Centro de Investigación Científica y de Educación Superior de Ensenada, Baja California

Maestría en Ciencias
en Ciencias de la Vida
con orientación en Microbiología Celular y Molecular

Identificación de secuencias genómicas virales en el piojo harinoso de la vid (*Planococcus ficus* Signoret, 1875) presente en viñedos de Ensenada, Baja California

Tesis
para cubrir parcialmente los requisitos necesarios para obtener el grado de Maestro en Ciencias

Presenta:

José Luis Duarte de Jesús

Ensenada, Baja California, México
2020

Tesis defendida por

José Luis Duarte de Jesús

y aprobada por el siguiente Comité

Dra. Jimena Carrillo Tripp
Directora

Miembros del comité

Dr. Miguel Ángel Martínez Mercado

Dra. Clara Elizabeth Galindo Sánchez

Dr. Jorge Alberto Cáceres Martínez

Dra. Patricia Juárez Camacho
Coordinadora del Posgrado en Ciencias de la Vida

Dra. Rufina Hernández Martínez
Directora de Estudios de Posgrado

Resumen de tesis que presenta **José Luis Duarte de Jesús** como requisito parcial para la obtención del grado de Maestro en Ciencias en Ciencias de la Vida con orientación en Microbiología Celular y Molecular.

Identificación de secuencias genómicas virales en el piojo harinoso de la vid (*Planococcus ficus* Signoret, 1875) presente en viñedos de Ensenada, Baja California

Resumen aprobado por:

_____
Dra. Jimena Carrillo Tripp
Directora de tesis

La vid, *Vitis vinifera* L., es la especie de vid más cultivada para la explotación de su fruto en las regiones de Europa, África y América. En México, Baja California (B.C.) es la entidad con mayor aporte a la producción nacional de uva con fines de uso industrial. Actualmente, la plaga de insecto más importante de los cultivos de *Vitis vinifera* L. en México y otros países es el piojo harinoso de la vid, *Planococcus ficus* (Signoret, 1875). En las últimas décadas, se ha explorado la diversidad de virus que infectan exclusivamente a insectos, con la finalidad de esclarecer sus interacciones ecológicas y encontrar, en última instancia, alternativas para el control de las poblaciones de insectos. Dicha prospección de virus específicos a *P. ficus* no había sido abordada. Por lo tanto, mediante técnicas de secuenciación masiva (RNA-seq) y análisis bioinformáticos, en el presente estudio se realizó la búsqueda e identificación de secuencias de virus con genoma de RNA o DNA en muestras de *P. ficus* colectadas en 14 viñedos de Ensenada, B.C. Dentro de los resultados de virus con genoma de RNA, se identificaron dos genomas casi completos de virus putativos relacionados con las familias *Tymoviridae* y *Dicistroviridae*. Adicionalmente, se encontraron fragmentos de genomas de virus putativos correspondientes a las familias *Tymoviridae*, *Reoviridae*, *Rhabdoviridae, Tombusviridae, Iflaviridae*, y al orden *Picornavirales*. Por otro lado, se logró la detección de secuencias relacionadas a virus de dsDNA de las familias *Baculoviridae*, *Poxviridae* y *Polydnaviridae*. Complementario al análisis bioinformático, se realizaron enriquecimientos de partículas tipo virus (VLPs) a partir de tejido de *P. ficus* y se analizaron por microscopía electrónica. Esto reveló la presencia de agregados de partículas que, en algunos casos, asemejan en tamaño y/o forma a los cuerpos de oclusión producidos por algunos virus de insecto durante la infección. La evidencia conjunta obtenida de los análisis realizados indica la presencia de uno o más virus de RNA putativos que pudieran asociarse específicamente con *P. ficus*. Este trabajo sienta las bases para futuros estudios de caracterización de virus específicos de *P. ficus* y virus de vid asociados a este insecto.

Palabras clave: *Planococcus ficus*, *Vitis vinifera* L., secuencias virales, virus específicos de insecto, control biológico, virómica, metagenómica.

Abstract of the thesis presented by **José Luis Duarte de Jesús** as a partial requirement to obtain the Master of Science degree in Life Sciencies with orientation in celular and molecular microbiology.

Identification of viral genome sequences in the vine mealybug (*Planococcus ficus*, Signoret 1875) present in vineyards of Ensenada, Baja California

Abstract approved by:

_____
Dr. Jimena Carrillo Tripp
Thesis director

The grapevine, *Vitis vinifera* L., is the most worldwide-spread species cultivated for grape production in Europe, Africa, and America. In Mexico, Baja California (B.C.) is the region with the highest grape production for industrial purposes. The vine mealybug, *Planococcus ficus* (Signoret, 1875), is the most important insect pest in grapevine growing areas in Mexico and other countries. Recently, the diversity of viruses infecting insects has been broadly explored to elucidate further ecological viral-hots interactions in many insect species, which in some cases has resulted in the application of virus-based biological control agents for insect pests. However, a prospection of viruses associated with *P. ficus* had not been addressed yet. In the present study, *P. ficus* individuals collected through 14 different vineyards in Ensenada, B.C., were analyzed by RNA metagenomics. Several sequences of putative RNA viruses were found, including two nearly complete genomes related to the *Dicistroviridae* and the *Tymoviridae* families. Moreover, several smaller sequences related to viral genomes of the *Tymoviridae*, *Reoviridae*, *Rhabdoviridae*, *Tombusviridae*, and the *Iflaviridae* families, as well as with *Picornavirales* order, were identified. A group of sequences related to dsDNA viral genomes of the *Baculoviridae*, *Poxviridae*, and the *Polydnaviridae* families was also found. As a complement to these bioinformatic results, virus-like particle (VLPs) enrichment and electron microscopy analyses were carried out. These experiments revealed the presence of some particles that resemble viral occlusion bodies in shape and size. The overall obtained evidence points to the presence of one or more putative RNA viruses that could specifically infect *P. ficus*. This study provides the first insight into the *P. ficus* virome. It sets the foundation to further studies aimed to characterize those viruses potentially infecting *P. ficus*, in addition to grapevine viruses associated with this mealybug.

Keywords: *Planococcus ficus*, *Vitis vinifera* L., viral sequences, insect specific viruses, biological control, viromics, metagenomics.

Dedicatoria

A Catalina, por supuesto.

El haber llegado a este lugar fue tu último regalo para mí, y lo hiciste sin que tú o yo lo supieramos; como dar un beso con amor mientras se está dormido, sólo una veladora de sueños como tú sería capaz.

"*Oh, be wiser you!*

Instructed that true knowledge leads to love"

-William Wordsworth

Lines left upon a Seat in a Yew-tree

Agradecimientos

Al Consejo Nacional de Ciencia y Tecnología por haberme beneficiado con la beca No. 911868 que me permitió sustentarme durante el desarrollo de mis estudios de maestría. Al CICESE y al personal académico (especialmente del Depto. de Microbiología) por su contribución tan importante a mi formación de posgrado, y al personal administrativo, por ser tan eficientes y profesionales en su labor. A cada uno de los profesores que me dio clase o laboratorio y me enseñaron lo que saben, especialmente a los "docs" Edgardo, Jimena, Carolina, Rafael, Leobardo, Miguel y Edgar, además de muchos otros doctores, posdocs y técnicos de CICESE, por sus clases tan instructivas, bien estructuradas y amenas.

Ahora, hay tantas personas a las que quiero agradecer (en tantas maneras) por haber llegado hasta aquí, que debo expresar mi inconformidad por tan breve espacio para hacerlo; pero ciertamente un texto como el presente no les haría justicia. Y si por descuido omito a alguien, sepan que cuando caiga en cuenta se los haré saber.

A la Dra. Jimena Carrillo Tripp, por su incansable esfuerzo de principio a fin en transmitirme su disciplina y su conocimiento, pero sobre todo por haber siempre exigido de mí lo mejor y siempre alentarme. Además, quiero agradecerle por haber puesto todo su empeño y estar siempre al pie del cañón buscando que este proyecto fuera una realidad (fue complicado muchas veces, pero lo sacamos ☺), y por haberme dado su confianza para comunicarle las buenas y las malas noticias. La admiro en lo personal y en lo académico.

Al Dr. Miguel Á. Martínez Mercado, por haber sido mi mentor en bioinformática, pero sobre todo, por tener siempre tanta disposición y paciencia para enseñar. Eres un gran profesional que me motivo y me oriento en cada etapa. Sin tu ayuda este trabajo simplemente no hubiera sido posible; más de una vez me ayudaste a brincar las bardas de callejones sin salida. Gracias además por todos los tips y consejos personales y académicos.

A la Dra. Clara E. Galindo Sánchez y al Dr. Jorge Cáceres Martínez, por su valioso tiempo, orientación, apoyo y comentarios, que me permitieron pulir y dar detalle a este trabajo. Además, le agradezco a la Dra. Clara por haberme permitido trabajar en el laboratorio de Genómica Funcional del Dpto. de Biotecnología Marina.

Quiero agradecer muy especialmente (pero de verdad mucho) a la M. en C. Anaid Saavedra Flores, por tu apoyo, enseñanza, orientación y mucha (pero de verdad mucha) disposición en la construcción de las

bibliotecas para la secuenciación. Dicen que el primer paso es el más importante de toda empresa, y gracias a tu pericia y destreza llevamos a cabo con éxito esa etapa.

A mis amigas y amigos de CICESE, Vero, Monse, Carlos, Mena, Andrea y Carmen, porque coincidimos en este barco y encontré a personas tan valiosas como ustedes. Gracias por estar ahí en las buenas, las malas y las peores; cuando mi mundo se hizo más pequeño. Por todas las salidas, risas, LA COMIDA (y la bebida), regaños y consejos, pero sobre todo por su tiempo (dentro y fuera de lo académico); no hay mejor regalo que ese. Sin ustedes no habría disfrutado tanto este tiempo de mi vida y seguro que más de una buena historia me habría perdido. Los quiero.

A mis “virusitas” Karendy (la jefa) y Ana Karen, porque fueron mis primeras amistades de CICESE y me enseñaron en mis primeros días de trabajo, ayudándome con muchos experimentos y no dejándome solo.

A mis amigos “biolokos” Rodrigo y Roberto, siempre estuvieron ahí para mi y me hicieron sentir como en casa cuando llegue a tierras extrañas. Por las reuniones, las cervezas, por las buenas y las malas experiencias en las que estuvieron conmigo. Me acompañaron, me enseñaron y confían en mí; yo apuesto por ustedes y sé que ustedes por mi; porque con ustedes ya perdí la cuenta de los años y espero que no se detenga.

A mis amigos de Toluca, Arlene, Mayeli, Daniel (Deni), Sonia, Richo (mi primo y amigo), porque aún estando lejos nunca dejaron que me sintiera solo y siempre estuvieron al pendiente de mi. Gracias por su amistad, consejos y disposición después de tantos años; aunque tengamos proyectos de vida tan diferentes no deseo más que verlos felices (como sé que ustedes a mi), sin importar que eso nos lleve lejos o no nos veamos tan seguido.

Finalmente, a mis padres, mis hermanos, mis sobrinos, mis tíos y mis primos. Ustedes son mi mundo y no imagino mi vida sin ustedes. Gracias por respaldarme en todos mis proyectos de vida y creer en mí incluso cuando yo dudo en hacerlo. Como siempre se los he dicho, soy quien soy por y para ustedes. Los amo.

Tabla de contenido

Lista de figuras

Lista de tablas

Capítulo 1. Introducción

1.1. Antecedentes

1.1.1. El viroma de insectos

Los virus se encuentran dentro de las entidades microbianas más abundantes, están asociados a células tanto eucariotas como procariotas y virtualmente se pueden encontrar en todo el planeta. A pesar de esta ubicuidad, se estima que se conoce apenas el 1% de su diversidad total (Zhang et al., 2019). Mucho del conocimiento actual de los virus se ha obtenido principalmente del enfoque en aquellos cuyos hospederos pueden ser cultivables o que son agentes patógenos tanto para el humano como para animales y plantas de importancia económica (Junglen y Drosten, 2013). Sin embargo, ahora se sabe que estas especies representan solo una pequeña fracción del total, puesto que algunos virus pueden causar infecciones subletales, asintomáticas o latentes, o encontrarse solo en pequeñas cantidades (Bolling et al., 2015; Nouri et al., 2018). Esto ha representado una limitante para los métodos de detección como la microscopía electrónica, el cultivo celular o la amplificación de DNA por la reacción en cadena de la polimerasa (PCR). Específicamente, al emprender el análisis de la diversidad viral en una muestra ambiental mediante técnicas moleculares surgen algunas complicaciones como el hecho de que no existe un solo gen que sea común a todos los genomas virales, y por tanto las técnicas dirigidas como la PCR o RT-PCR están limitadas por la necesidad de conocer previamente la secuencia blanco del virus (Liu et al., 2011).

La implementación reciente de las técnicas de secuenciación masiva de siguiente generación o NGS (por sus siglas en inglés) han permitido refinar la detección y caracterización de virus presentes en muestras provenientes de diferentes fuentes, generando así un incremento en la diversidad conocida de virus en la biosfera, referida actualmente como virosfera (Liu et al., 2011; Zhang et al., 2018). Los estudios enfocados al análisis de dicha diversidad basados en NGS son numerosos y se han abordado en pequeña y gran escala; desde los análisis macro-ecológicos a nivel de biomasa, hasta la búsqueda e identificación de todos los virus que se encuentran asociados con alguna especie hospedera en particular o viroma (Koonin y Dolja, 2013). El viroma engloba a todos los virus que presentan algún tipo de relación con una especie y que se pueden categorizar en cuatro grupos generales: i) los virus que se replican en el hospedero, ii) los que solo son transmitidos por el individuo sin replicarse en él, iii) los virus de endosimbiontes o parásitos del individuo y iv) los virus que se encuentran sin un rol específico (Conceição-Neto et al., 2015; Feng et al., 2017; Roossinck et al., 2015).

Durante la última década, la exploración de la diversidad viral en invertebrados ha suscitado un notable avance en el descubrimiento de los virus de artrópodos en general, y de insectos en particular (Mokili et al., 2012; Li et al., 2015a; Shi et al., 2016). El interés creciente por estudiar el viroma de insectos se debe a que tienen una alta capacidad de fungir como vectores u hospederos de virus y otros patógenos tanto de plantas como de animales (Junglen y Drosten, 2013; Gytis y Obbard 2015; Shi et al., 2018). Dentro del viroma de las distintas especies de insectos se ha podido reconocer que algunos virus son inocuos para el insecto, mientras que otros pueden afectar su adecuación de diversas formas (Roossnick, 2011), por ejemplo, reduciendo su tiempo de vida, su fertilidad (Laubscher y von Wechmar, 1992; Martin et al., 2012; Chen et al., 2014), o bien, siendo necesarios para su supervivencia o reproducción (Roossinck, 2011; Ryabov et al., 2009; Lefeuvre et al., 2019). De esta forma, el descubrir, identificar y caracterizar las diferentes especies de virus que componen el viroma de insectos es necesario para entender las complejas relaciones ecológicas que suceden entre los virus y los organismos celulares que les permiten completar su ciclo de reproducción (Feng et al., 2017; Nouri et al., 2018).

Un grupo de virus de interés particular del viroma de insectos es el de aquellos que se replican exclusivamente en dichos hospederos, denominándose así como virus específicos de insecto o ISVs (por sus siglas en inglés) (Bolling et al., 2015). Los ISVs ofrecen la oportunidad de desarrollar nuevas alternativas para el manejo o control de las poblaciones de insectos que infectan y a su vez de los patógenos que éstos transmiten (Roundy et al., 2017; Whitfield y Rotenberg, 2015), lo cual es de particular interés en los ámbitos clínico, agroalimentario y biotecnológico (Whifield y Rotenberg 2015). Desde este punto de vista, los ISVs también pueden verse como antagonistas indirectos para otros patógenos. En el sector agrícola, el control de plagas de insectos tiene una importancia central, tanto por los efectos negativos directos de la plaga sobre los cultivos, como por el riesgo secundario de transmisión de enfermedades que los insectos pueden ocasionar (Vincent et al., 2012). Actualmente, a través de los análisis de NGS, se han realizado esfuerzos para el descubrimiento y caracterización de los virus en insectos plaga, como un primer paso en el desarrollo de nuevos agentes de control biológico (Öhlund et al., 2019; Valles y Rivers, 2019).

1.1.2. El piojo harinoso de la vid, *Planococcus ficus*

El piojo o cochinilla harinosa de la vid, *P. ficus* (Signoret, 1875) es un insecto incluido en el orden Hemiptera, suborden Sternorryncha, superfamilia Coccoidea, y la familia Pseudococcidae (Ben Dov y Matile Ferrero, 1995). Una de las características morfológicas distintivas que presentan todos los pseudocóccidos (familia

Pseudococcidae) es la cutícula dividida en segmentos o placas sobre la cual depositan una secreción cerosa blanquecina (de lo que deriva el nombre común de “piojo harinoso” o “cochinilla harinosa”) (Walton y Pringle, 2004; Mani y Amala, 2016). *P. ficus* presenta un amplio rango geográfico ya que se ha reportado su presencia en más de 45 países (García-Morales et al., 2016).

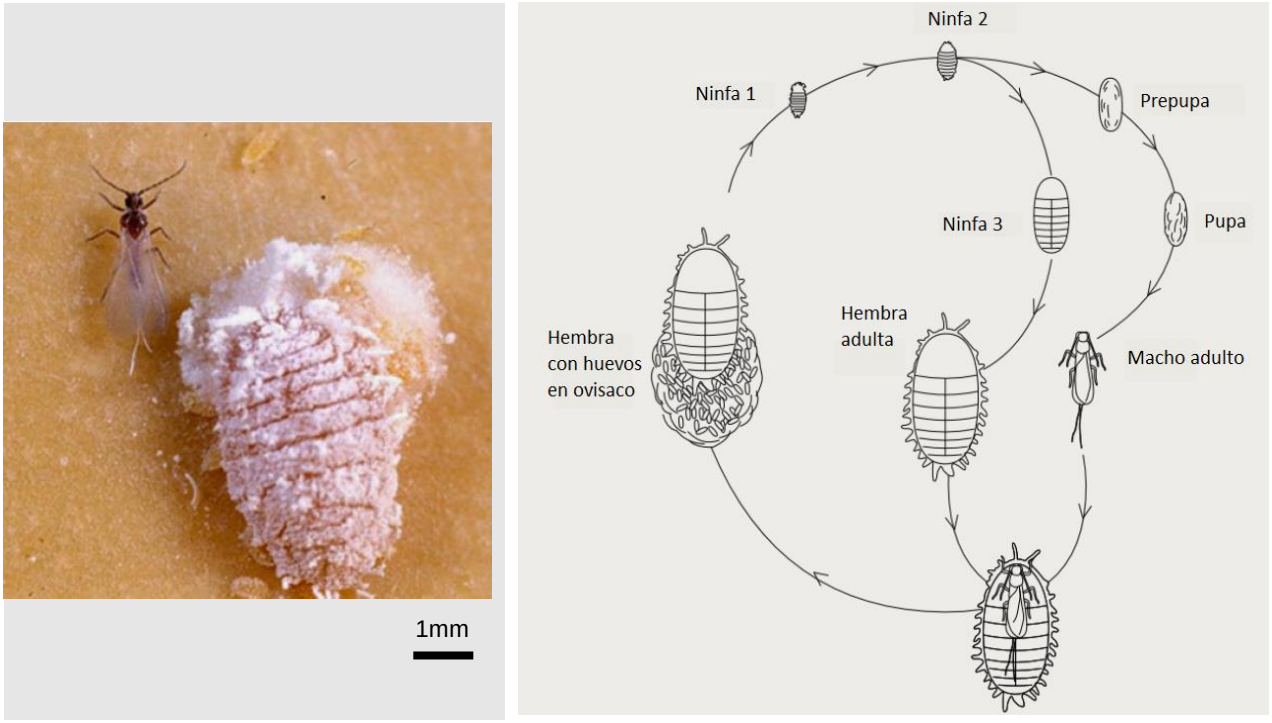

Figura 1. Dimorfismo sexual y ciclo de vida de *P. ficus*. Imágenes reproducidas de Estopá-Consuegra, 2015 (izquierda) Y Daane et al., 2004 (derecha).

Los individuos de *P. ficus* presentan dimorfismo sexual (**figura 1**). La hembra tiene un desarrollo heterometábolo, con 3 estadios ninfales antes de alcanzar el estadio adulto (Estopà Consuegra, 2015). Las hembras adultas son ápteras, con un aparato bucal completamente funcional, el cuerpo tiene forma ovalada de hasta 9 mm de longitud (en su eje más largo) y son mayormente sedentarias (Mani & Shivaraju, 2016c). Por su parte, el macho presenta un desarrollo holometábolo, con dos estadios ninfales y dos estadios de pupa (donde ocurre la metmorfosis) antes de alcanzar el estadio adulto (Estopà Consuegra, 2015). El macho adulto es alargado (de 1.5 mm de longitud) y surge de la pupa con alas y el aparato bucal atrofiado; vuela activamente y sólo vive de 1 a 3 días. La mayoría de los estadios de desarrollo se encuentran cubiertos por cera, excepto las ninfas de recién eclosión, llamadas “caminadores”, los cuales se mueven activamente sobre y entre plantas (Mani & Shivaraju, 2016c).

1.1.3. *P. ficus* y su estatus como plaga de la vid, *Vitis vinifera* L. en México y el mundo

P. ficus es una de las plagas de insecto de mayor importancia para los cultivos de la vid, *Vitis vinifera* L. en la región Mediterránea, África, el medio Oriente, Argentina y recientemente en California y México (Cox, 1989; Walton y Pringle, 2004; Daane et al., 2008). Específicamente en viñedos de Norte América (California y México), ésta es la especie de piojo harinoso más importante (Daane et al., 2008). Fue identificado por primera vez a principios de 1990 en el valle de Coachella, California y una década después fue identificado en Sonora, México (Danne et al., 2004; Castillo et al., 2008).

El primer registro que se tiene de *P. ficus* para México fue en la Costa de Hermosillo, Sonora en noviembre del año 2000. Para mayo de 2001 se registraron daños en la uva de mesa cv. Flame Seedless, con pérdidas de hasta 100% de la producción (Castillo et al., 2004). Posteriormente, la Secretaría de Agricultura y Desarrollo Rural (SAGARPA) reportó la presencia de *P. ficus* en Baja California en 2014 (CESVBC, 2018). Mediante un análisis de haplotipos, se determinó que el linaje de las poblaciones en Sonora es el mismo que el de las poblaciones de California, por lo que es posible que *P. ficus* haya sido introducido desde California a través del comercio (tanto legal como ilegal) de material vegetal infectado (Daane et al., 2018). Hasta el año 2018, Ensenada se reportó como el único de los 6 municipios de Baja California con presencia de *P. ficus*, en ~54 viñedos con diferentes grados de infestación (**figura 2**), que abarcan en conjunto 1,048.9 ha de las 3600 ha cultivadas (CESVBC, 2018).

Lo anterior es de capital relevancia ya que a nivel nacional Baja California ocupa el 4º lugar en producción anual (5,924 toneladas en 2017) de uva de mesa, mientras que ocupa el 1º lugar en la producción nacional anual de uva de uso industrial (17,924 toneladas en 2017), la cual contempla variedades de uva para la producción de jugos y vinos (SIAP 2018). Así mismo, Baja California ocupa el 1º lugar a nivel nacional en la producción de vino, aportando un 85% del total nacional (SIAP 2018). Dicha producción anual de vino en 2014 fue valorada por el Consejo Mexicano Vitivinícola A.C en un total de 2844.3 millones de pesos, representando así una de las fuentes de mayor ingreso para el estado (CESVBC 2018).

Figura 2. Las poblaciones de *P. ficus* en viñedos del Valle de Guadalupe, Ensenada, Baja California, México (mayo-junio de 2019). Crédito de las fotografías: Eduardo Hernández Navarro.

1.1.4. Daños de *P. ficus* a la vid

Los pseudocóccidos son insectos exclusivamente fitófagos que se alimentan succionando el contenido de los conductos floemáticos o del mesófilo (o ambos) de la planta mediante un estilete que forma parte de su aparato bucal (Gullan y Martin, 2009). Sus poblaciones son de hábitos crípticos, por lo que se hacinan debajo de la corteza del tronco o las ramificaciones y se alimentan de la savia. De esta forma, al igual que

otros fitófagos hemípteros, los pseudocóccidos reciben una dieta basada principalmente en carbohidratos, con un consumo limitado de aminoácidos y otros compuestos nitrogenados (Mani y Shivaraju, 2016b).

Muchos de los daños que genera *P. ficus* en los viñedos, directos e indirectos, se deben a su forma de alimentación y sus productos de desecho, aunque la intensidad de dichos daños varía dependiendo de la zona geográfica (Daane et al., 2008). El daño directo sobre la planta comienza con la sustracción de nutrientes de la misma y la consecuente reducción de su vigor, crecimiento y desarrollo. Las infestaciones severas de *P. ficus* pueden resultar en clorosis de las hojas y tras la alimentación continua del piojo durante varios ciclos anuales puede ocurrir la muerte de la planta (Daane et al., 2012). En cuanto al daño indirecto del piojo, el más inmediato se debe a la excreción continua de un desecho metabólico conocido como "mielecilla" que puede acumularse en las hojas o en los racimos de uvas y propiciar el crecimiento de hongos saprófitos sobre la planta. Estos hongos a su vez pueden reducir la calidad de los frutos (cuando se desarrollan sobre los mismos, causando su pudrición) o reducir la fotosíntesis de la planta (Mani y Shivaraju, 2016a). El daño indirecto afecta tanto a los productos como a los subproductos de la vid; Catania et al., (2007) reportaron cambios a las propiedades enológicas y organolépticas de vinos elaborados con frutos de plantas infestadas con *P. ficus* en regiones de cultivo de Europa.

Aunque los daños directos sobre la planta pueden no ser severos de forma inmediata, la presencia del piojo en viñedos con cualquier grado de infestación es un riesgo potencial de transmisión de virus fitopatógenos, por lo que muchas veces el rol del piojo como vector es más preocupante que la presencia del piojo en sí (Daane et al., 2008; Daane et al., 2012). *P. ficus* se ha identificado como uno de los vectores de la enfermedad del enrollamiento de las hojas de la vid o GLD (por sus siglas en inglés), una de las enfermedades más destructivas en *V. vinifera* L. que ha generado hasta un 30-40% de pérdidas en viñedos de California (García Morales et al., 2016). El agente etiológico de la GLD es el complejo grapevine leafroll-associated virus-n (GLRaV-n) que comprende 9 especies de virus, siendo el GLRaV-3 el más virulento y el cual puede ser transmitido por *P. ficus*, así como los virus GLRaV-1 y GLRaV-4 (Almeida et al., 2013; Herrbach et al., 2017).

En Baja California se han detectado plantas con síntomas de GLD y se ha comprobado la presencia de los virus GLRaV-1 y 3 (Monroy Corral, 2019), añadiendo así el factor de riesgo de dispersión de virus a la presencia *P. ficus* en viñedos de esta región.

1.1.5. *P. ficus* y su dinámica poblacional

Aunque se ha visto que *P. ficus* está presente en los viñedos a lo largo del año, la densidad de sus poblaciones varía de acuerdo a la región. Reportes en India indican que se presenta un pico anual de densidad poblacional de marzo a abril (Mani y Amala, 2016), mientras que en viñedos de California se han reportado dos picos anuales de alta densidad poblacional, uno de marzo a mayo y de nuevo en septiembre (Daane et al., 2004), con un total de dos hasta siete generaciones por año (Daane et al., 2008). Sin embargo, más estudios son necesarios para conocer la dinámica poblacional de esta especie en Baja California.

El desarrollo estacional del piojo depende de la fenología del cultivo; la fluctuación de las poblaciones del insecto está relacionada al desarrollo de la vid. Después de la poda en septiembre-octubre y hasta la mitad de diciembre, el piojo permanece en la parte basal del tronco, los cordones y el tallo o incluso en las raíces de la planta. En general, la abundancia de piojo comienza a aumentar a partir de este punto (Daane et al., 2004; Becerra et al., 2006; Mani y Amala, 2016). Posteriormente, comienza a migrar hacia las partes altas de la planta; desde las raíces y la parte baja del tronco, hacia las ramas, las panículas de las flores y las uvas (Vincent et al., 2012). Alcanza su máximo desarrollo antes de la cosecha de los racimos, aproximadamente de marzo a mayo. Después de la cosecha desciende de nuevo a las partes basales de la planta. Generalmente las poblaciones disminuyen de junio a septiembre, aunque se ha observado que puede presentarse un segundo incremento en sus poblaciones durante el otoño (Daane et al., 2004; Mani y Amala, 2016).

1.1.6. Control biológico de *P. ficus*

Debido a la importancia regional que presentan las variedades de vid en Baja California, a la potencialidad de *P. ficus* como vector de virus fitopatógenos, así como a los daños que se han reportado en viñedos de México y otros países, es necesario plantear estrategias para el monitoreo y control de esta plaga (Castillo et al., 2008).

Específicamente en México, el control químico de *P. ficus* es la práctica más común (Castillo et al., 2009). Sin embargo, los agentes químicos presentan inconvenientes importantes como el no ser específicos para el insecto plaga, el estar sometidos a regulación (o veto) continua por el grado de toxicidad que presentan,

el propiciar desarrollo de resistencia ante uso prolongado; y en el caso de los insecticidas de contacto, tener una eficiencia variable debido a que *P. ficus* se encuentra protegido por sus secreciones cerosas y por la corteza de la planta (Daane et al., 2012).

Así, el control biológico que consiste en el uso de las especies depredadoras o parasitoides de *P. ficus,* es una alternativa que supera algunas limitaciones del control químico como la toxicidad o la baja especificidad. En Baja California, se han planteado recientemente estrategias de control biológico con el coleóptero *Cryptolaemus montrouzieri* (depredador) o el encítrido *Anagyrus pseudococci* (parasitoide), que se han empleado como agentes de control biológico con cierto grado de efectividad en otros países (Castillo 2004; Gutierrez et al., 2008; Daane et al., 2012; CESVBC 2018). Por otro lado, el uso de estos y otros insectos antagonistas naturales ha demostrado tener una efectividad limitada o nula en algunas regiones del mundo, debido en parte a que las hormigas con que se asocia el piojo lo protegen y le proveen de refugios (Mani y Shivaraju, 2016a).

El uso de ISVs que puedan regular las poblaciones de insectos o modular su potencial como vectores de patógenos se ha propuesto como una estrategia de control biológico complementaria al uso de especies de insectos antagonistas o bien, como alternativa al control químico. Es por ello que se ha realizado la búsqueda, identificación y caracterización de virus asociados a diferentes especies plaga de insecto con potencial para el biocontrol (Valles y Hashimoto, 2009; Junglen y Drosten, 2013; Valles y Rivers, 2019). Sin embargo, actualmente no existe información acerca del viroma de *P. ficus* ni se ha reportado algún virus como específico para el mismo.

1.1.7. Estudios de la microbiota en hemípteros fitófagos y otros insectos

Debido a que la nutrición que obtienen los pseudocóccidos de la savia de las plantas que parasitan es subóptima (principalmente carbohidratos), requieren del aporte de nutrientes esenciales brindados por parte de procariotas intra/extracelulares y hongos, con los cuales mantienen relaciones simbióticas obligadas (Iasur-Kruh et al., 2015; Lin et al., 2019).

Además de su rol en la nutrición y el metabolismo (Prosser & Douglas, 1991; Douglas, 1998), se ha demostrado que los simbiontes bacterianos y fúngicos de los hemípteros y otros órdenes de insectos influyen en aspectos no metabólicos de su ecología y evolución como: i) incrementar o disminuir la

competencia del insecto como vector de virus patógenos (Caragata et al., 2016; Roundy et al., 2017) , ii) conferir fenotipos resistentes contra parásitos o virus patógenos para el insecto (Oliver et al., 2003; Teixeira et al., 2008), o incluso iii) inducir la feminización, la partenogénesis o la muerte de machos en las poblaciones de insectos (Werren et al., 2008).

Pese a que los procariotas y hongos son los endosimbiontes de insectos más estudiados, representan solo una fracción de su microbioma, existen estudios que demuestran cómo las especies del viroma también pueden influir en la fecundidad, reproducción o supervivencia de los insectos (D'Arcy et al., 1981; Laubscher y von Wechmar, 1992). Dichas observaciones han permitido plantear a lo largo del tiempo diversas aplicaciones de las especies del viroma de insectos para el manejo y control de plagas (Whitfield y Rotenberg, 2015; Nouri et al., 2018).

Actualmente, muchas especies de virus (la gran mayoría de ellos de la familia *Baculoviridae*) son usadas como biopesticidas en la agricultura, al igual que otros microorganismos como hongos o bacterias (Bonning y Nusawardani, 2007; Lacey et al., 2015). Más aún, se han generado y probado virus o proteínas virales recombinantes para reducir la supervivencia o la capacidad vectorial (de patógenos) en algunos insectos de interés en el ámbito agrícola (Whitfield y Rotenberg, 2015). Por otra parte, en el ámbito clínico se ha puesto énfasis en las especies del viroma de insectos que no matan a su hospedero pero que reducen su capacidad de actuar como vectores de transmisión de otros virus de importancia médica (Vasilakis y Tesh, 2015; Öhlund et al., 2019). Todo este cúmulo de aplicaciones pone de manifiesto la versatilidad de los virus como agentes de biocontrol de plagas de insectos. En este sentido, los hemípteros fitófagos han sido un foco de atención ya que prácticamente cada superfamilia (Aphodoidea, Aleyrodoidea, Psylloidea y Coccoidea) contiene especies representativas reconocidas como plagas de cultivos agrícolas a nivel mundial (Gullan y Martin, 2009).

En consecuencia, se han abordado estudios concernientes a la prospección del viroma de algunas especies de hemípteros fitófagos. Feng et al. (2017) realizaron una búsqueda de secuencias virales mediante NGS en *Aphis glycines*, que afecta el cultivo de la soya, donde reportan 6 segmentos genómicos virales, 3 de los cuales corresponden respectivamente a los virus Aphis glycines virus 2 (AGV-2), Rhopalosiphum padi virus (RhPV) y aphid lethal paralysis virus (ALPV), previamente reportados en esta especie (Moon et al., 1998; van Munster et al., 2002; Liu et al., 2016). De estos, los virus RhPV y ALPV ya se han caracterizado como virus que pueden provocar efectos negativos en las poblaciones de los áfidos que infectan (D'Arcy et al., 1981; Laubscher y von Wechmar, 1992). Por su parte, Nouri et al. (2016) realizaron una búsqueda de secuencias virales por NGS en la especie *Diaphorina citri,* que actúa como vector de la bacteria Candidatus

Liberibacter asiaticus, el agente etiológico de la enfermedad dorada de los cítricos. Los autores encontraron 6 secuencias con cierta **similitud**[1] a ISVs ya conocidos, pero suficientemente divergentes para inferir que pueden representar posibles nuevos ISVs. En contraste con lo anterior, al momento no se ha reportado algún análisis de secuencias virales en pseudocóccidos, ya sea por NGS o por alguna otra técnica molecular.

Aunque este tipo de investigaciones sobre el viroma de insectos basadas en metagenómica tiene la posibilidad de derivar en la identificación de virus con potencial biológico o biotecnológico (Whitfield & Rotenberg, 2015; Nouri et al., 2018), debe recalcarse que solo conciernen al reconocimiento e identificación iniciales de la diversidad presente en una especie hospedera, siendo necesarios análisis posteriores para realizar una caracterización biológica y física completa, evaluando las características específicas de los virus encontrados (Junglen y Drosten, 2013). Este proceso de caracterización *in silico*, *in vitro* e *in vivo* permite sentar las bases teóricas de la relación entre un insecto, su viroma y los patógenos que el insecto puede transmitir, lo cual puede después trasladarse a una aplicación específica (Bonning et al., 2014; Öhlund et al., 2019).

1.1.8. Los virus específicos de insectos o ISVs

El interés por la búsqueda de virus que se asocian con diferentes taxa de insectos se debe a la alta frecuencia en que estos últimos actúan como agentes de dispersión de dichos virus (Bolling et al., 2015; Vasilakis & Tesh, 2015). Ya sea como hospederos finales o como vectores, los insectos son reconocidos actualmente como una pieza clave en los ciclos de replicación de una amplia diversidad de virus, tanto de los que poseen un genoma de ácido ribonucleico (RNA, por sus síglas en inglés) como de aquellos con genoma de ácido desoxirribonucleico (DNA, por sus síglas en inglés). De las 7 categorías de la clasificación de Baltimore (Flint et al., 2015), a saber (aquí enunciadas por sus siglas en inglés), virus de (I) DNA de doble cadena (dsDNA), (II) de cadena simple positiva de DNA (+ssDNA), (III) de cadena doble de RNA (dsRNA), (IV) de cadena simple positiva de RNA (+ssRNA), (V) de cadena simple negativa de RNA (-ssRNA), (VI) de cadena simple de RNA que retrotranscribe (ssRNA-RT) y (VII) de doble cadena de DNA que retrotranscribe (dsDNA-RT); la mayoría de los virus asociados a insectos que se han descrito pertenecen a las categorías I, II, III, IV y V (Li et al., 2015; Shi et al., 2016).

[1] Nota: En adelante se remarca en **negrita** la primera aparición de conceptos incluidos en el glosario.

El estudio del viroma de insectos mediante técnicas de NGS ha derivado en una creciente identificación de virus que se replican únicamente en tejidos de un insecto hospedero, ISVs (Mokili et al., 2012; Bolling et al., 2015). Algunos de los ISVs de RNA mejor estudiados se agrupan dentro de las familias *Dicistroviridae*, *Iflaviridae*, *Flaviviridae, Bunyaviridae, Rhabdoviridae, Togaviridae, Reoviridae, Nodaviridae*, *Alphatetraviridae, Permutotetraviridae y Carmotetraviridae* (Vasilakis y Tesh, 2015; Ryabov, 2017; Nouri et al., 2018). Por otra parte, también se conoce una amplia diversidad de ISVs de DNA pertenecientes a las familias *Baculoviridae*, *Parvoviridae*, *Ascoviridae*, *Nudiviridae* y *Polidnaviridae*, así como en las subfamilias *Entomopoxvirinae* y *Betairidovirinae* (Williams et al., 2017).

A partir de la definición de ISV, se puede establecer una distinción dicotómica para dichos virus según los efectos que producen en sus respectivos hospederos, siguiendo la terminología de "virus abierto" y "virus encubierto", de uso común en virología de invertebrados. Se consideran como virus abiertos a todos aquellos que producen síntomas evidentes de infección bien definidos, causando en algunos casos, afecciones letales en las diferentes etapas del desarrollo del hospedero. Por su parte, los virus encubiertos son aquellos que causan infecciones asintomáticas, esto es sin daño evidente, pese a que pueden producir infecciones abiertas bajo ciertas condiciones (de Miranda y Genersch, 2010).

1.1.8.1. ISVs entomopoatógenos

Los ISVs abiertos han sido objeto de una trayectoria de investigacion más intensa y larga que los encubiertos, debido en parte a que históricamente fueron los primeros en ser reconocidos por los síntomas de infección que producen (Roossinck, 2011). Los ISVs de RNA de mayor interés por sus efectos como entomopatógenos pertenecen al orden *Picornavirales* (denomindos colectivamente como virus picorna-like) y a las familias *Dicistroviridae, Iflaviridae* y *Solinviviridae*, al género *Cypovirus* de la familia *Reoviridae* y al género *Alphanodavirus* de la familia *Nodaviridae* (Belloncik & Mori, 1998; Bonning, 2009; Chen et al., 2012; Ryabov, 2017; Brown et al., 2019). En cuanto a los ISVs de dsDNA, algunos de los entomopatógenos más importantes pertenecen a las familias *Baculoviridae* y *Ascoviridae*, y a las subfamilias *Betairidovirinae* (géneros *Iridovirus* y *Chloriridovirus*) y *Entomopoxvirinae* (géneros *Alpha-*, *Beta-* y *Gammaentomopoxvirus*) (Thiem, 1999). Es importante mencionar que si bien dichos taxa pueden impactar negativamente en las poblaciones, los efectos letales no siempre se observan, y de hecho se ha probado que en muchos casos las infecciones que producen son tales que el hospedero es asintomático (lo cual por definición es una infección encubierta). Sin embargo, se ha visto que en algunos casos dichas

infecciones asintomáticas tienen efectos detrimentales a largo plazo sobre el hospedero. Así pues, los efectos producidos por las infecciones de los ISVs mencionados pueden ser variables, desde cambios conductuales (McMenamin y Genersch, 2015; Valles y Rivers, 2019), anormalidades en el desarrollo (de Miranda y Genersch, 2010), reducción de la fecundidad (Valles et al., 2013), hasta la inviabilidad y muerte de los individuos (Cory y Myers, 2003; D'Arcy et al., 1981; Manousis y Moore, 1987; Williams et al., 2017).

La diversidad de hospederos de estos ISVs también es extensa, siendo algunos de principal importancia para el humano, ya sea como plagas de cultivos, como productores de recursos aprovechables (como en el caso de la apicultura o la sericultura), o por su rol ecológico como polinizadores de plantas silvestres y cultivadas (Caballero y Williams, 2008; McMenamin y Genersch, 2015). Considerando a las especies virales oficialmente reconocidas por el Comité Internacional para la Taxonomía de Virus (ICTV, por sus siglas en inglés), la mayoría de los hospederos de dicistrovirus e iflavirus son especies de hemípteros e himenópteros (Bonning, 2009; van Oers, 2010); los cypovirus infectan mayormente lepidópteros (Belloncik y Mori, 1998); los baculovirus infectan principalmente lepidópteros, himenópteros y dípteros (Harrison et al., 2018); los hospederos de iridovirus son mayormente dípteros, coleópteros y lepidópteros (Williams, 2008); las especies de hospederos de entomopoxvirus son lepidópteros, coleópteros, ortópteros y dípteros (L. A. King et al., 1998; Williams et al., 2017); por su parte la mayoría de los ascovirus se han encontrado sólo en especies de lepidópteros (Asgari et al., 2017). En la mayoría de los casos, los estadios más suceptibles a padecer enfermedades manifiestas son las larvas o estadios juveniles (Thiem, 1999; Williams et al., 2017).

Algunos de los ISVs picorna-like se han estudiado extensamente por sus efectos detrimentales, como malformidad, parálisis aguda o cambios de comportamiento. Por ejemplo, en el caso de *Apis mellifera,* se han descrito alrededor de 23 especies de virus (la mayoría de ellos de RNA de las familias *Dicistroviridae* e *Iflaviridae*) como agentes causales de diferentes afecciones (Chen y Siede, 2007; McMenamin y Genersch, 2015). De entre estas, las infecciones por el dicistrovirus Israeli acute paralyisis virus (IAPV) y el iflavirus deformed wing virus (DWV) en individuos de *A. mellifera* se han asociado a uno de los fenómenos más devastadores para dicha especie en diferentes regiones del mundo, el desorden del colapso de las colonias (Bonning, 2009; Vasilakis y Tesh, 2015). Por otro lado, en el caso de la hormiga roja *Solenopsis invicta* (una plaga de los bosques estadounidenses), se han identificado a la fecha 13 virus de RNA, de los cuales Solenopsis invicta virus 3 (SINV-3; del género *Invictavirus*) es el más virulento; reduce la fecundidad de las reinas y altera el comportamiento de las obreras infectadas, lo que a su vez provoca una muerte por inanición de las larvas y el colapso de las colonias (Valles & Hashimoto, 2009; Valles et al., 2014). Debido a sus características, se ha propuesto el uso de SINV-3 como un agente de biocontrol.

También se conocen ampliamente otros ISVs (de dsDNA), como los miembros de la familia *Baculoviridae,* por su impacto negativo en las poblaciones de insectos. De hecho, estos son los más ampliamente usados a escala industrial como agentes de biocontrol y se caracterizan por producir infecciones conocidas como "poliedrosis" nuclear (Haase et al., 2015; Ikeda et al., 2015). Este nombre se debe a que producen macroestructuras denominadas cuerpos de oclusión u OBs (por sus siglas en inglés), que ocluyen o agrupan a los viriones producidos y se forman a partir de un solo tipo de proteína, expresada abundantemente durante la infección, llamada poliedrina (Harrison et al., 2018). Otros virus de dsRNA del género *Cypovirus* (familia *Reoviridae*) también producen poliedrosis y OBs, aunque esto ocurre en el citoplasma de la célula (no en el núcleo). En el caso de las infecciones por cypovirus, se ha reportado que los efectos son crónicos en la mayoría de los casos y letales sólo esporádicamente; incluso las larvas con una fuerte infección pueden alcanzar el estado de pupa y adulto. No obstante, pueden causar alteraciones fisiológicas y metabólicas que provocan inanición de las larvas, lo cual las hace más sensibles a factores de estrés ambiental (Belloncik & Mori, 1998).

Independientemente de su aplicación como agentes de biocontrol, se ha demostrado que todos los virus entomopatógenos, sea que produzcan infecciones abiertas o no, tienen papeles reguladores clave de las poblaciones de insectos en los ecosistemas (Thiem, 1999).

1.1.8.2. ISVs mutualistas

Si bien los primeros ISVs se identificaron en función de los efectos citopáticos que producían, se ha probado que muchos, pese a que pueden producir efectos patogénicos en cierto grado, establecen relaciones mutualistas o comensalistas con el hospedero. A la luz de esto, se ha relegado a las infecciones que provocan la muerte del hospedero (i.e., el antagonismo entre un virus y el hospedero) como solo una de muchas posibles interacciones ecológicas (Roossinck, 2011; Öhlund et al., 2019).

En general, se han caracterizado virus de distintas familias que sostienen relaciones mutualistas con plantas, hongos, procariotas e insectos y otros animales (Roossinck, 2011; Lefeuvre et al., 2019). Respecto a insectos, se ha observado que algunos virus pueden incrementar la fecundidad y supervivencia de sus hospederos. Por ejemplo, el desarrollo de morfotipos alados en el áfido *Dysaphis plantaginea* está relacionado con la presencia del virus Dysaphis plantaginea densovirus (DplDNV), siendo aquellos insectos infectados por el virus los que pueden llegar a colonizar otras plantas, contribuyendo así el virus a la

dispersión del insecto (Ryabov et al., 2009). En el caso del áfido *Acyrthosiphon pisum*, la infección con bacteriófagos de su endosimbionte heredable *Hamiltonella defensa* dota a esta última de un fenotipo que le confiere a su vez resistencia al áfido contra el ataque del parasitoide *Aphidius ervi*, contribuyendo así los bacteriófagos de forma indirecta a la supervivencia del áfido (Oliver et al., 2009).

Uno de los ejemplos más notables de virus mutualistas es la relación entre las avispas braconidas (Himenoptera: Braconidae) e ichneumonidas (Himenoptera: Ichneumonidae) parasitoides y los virus de dsDNA de los géneros *Bracovirus* e *Ichovirus*, de la familia *Polydnaviridae* (Roossinck, 2011; Renault, 2012). La relación entre estos virus y las avispas se ha descrito como simbiogénica, donde el genoma del virus está integrado en el genoma de la avispa, y las formas "episomales" del virus (cuando se forman viriones) llevan tanto el genoma del virus como algunos genes de la avispa. Las avispas parasitoides inyectan sus huevos en larvas de himenópteros (orugas), que proporcionan un nicho para el desarrollo de las larvas y pupas de la avispa. Durante la puesta de los huevos, la avispa también inyecta viriones que expresan (exclusivamente) los genes de la avispa que permiten suprimir la respuesta inmune de la oruga. En ausencia de esta supresión, los huevos son degradados por la oruga (Roossinck, 2011).

Otros virus de las familias *Ascoviridae* y *Reoviridae* se han encontrado en una asociación mutualista más compleja con avispas parasitoides. Por ejemplo, Diadromus pulchellus ascovirus 4 (DpAV4) se caracteriza por causar infecciones letales en el insecto hospedero de la avispa. Se ha demostrado experimentalmente que una infección de DpAV4 en *Acrolepiopsis assectella* inhibe la formación de cápsulas de melanina (las estructuras que degradan el huevo de la avispa), con lo cual a su vez se promueve la supervivencia de los huevos (Renault et al., 2002; Renault, 2012). Sin embargo, las infecciones de DpAV4 (un virus no simbiogénico con la avispa) por si solas pueden matar al hospedero antes de que el huevo se desarrolle (Roossinck, 2011). Por otra parte, se ha observado que en los casos de parasitismo de la avispa donde se encuentra el reovirus Diadromus pulchellus idnoreovirus 1 (DpRV1), las infecciones de DpAV4 se retardan lo suficiente para que el huevo logre su desarrollo. Se ha argumentado que una de las posiblies causas de los numerosos ejemplos de simbiosis entre diferentes virus que infectan insectos y las avispas braconidas e ichneumonidas se debe al carácter parasitoide de estas úlitmas, donde la asociación supone una mejora en la adecuación de estas últimas (Roossinck, 2011).

1.1.9. Virus que infectan a la vid, *Vitis vinifera* L.

La vid, *V. vinifera* L. pertenece al género *Vitis* de la familia Vitaceae, la cual comprende 11 géneros y alrededor de 700 especies distribuidas en Asia, Norte América y Europa, en regiones tropicales, subtropicales o mediterráneas (Rzedowski y Calderón 2005; Terral et al., 2010). Estas especies son plantas arbustivas trepadoras (raramente herbáceas) con frutos en forma de bayas (Rzedowski y Calderón 2005). El género *Vitis* comprende 11 especies, de las cuales sólo *V. vinifera* ha adquirido una importancia económica y agrícola a nivel mundial, si bien otras especies como *V. rupestris*, *V. riparia* y *V. berlandieri* y sus híbridos son usadas por su capacidad de resistencia a distintos patógenos. En la actualidad, se utilizan alrededor de 5000 cultivares diferentes de *V. vinifera* alrededor del mundo para la producción de uva de mesa (fruto fresco), pasa (fruto seco) y vino (Maliogka et al., 2015).

Al igual que ocurre con otras especies propagadas vegetativamente, la vid es especialmente susceptible a las infecciones de distintos patógenos como virus, viroides, fitoplasmas y bacterias. Martelli (2017) presenta una lista de 70 especies de virus de RNA y DNA que se han identificado en vid, de las cuales alrededor de 30 se han asociado a enfermedades. Las infecciones virales que producen efectos abiertos a menudo son mixtas (i.e., varios virus replicándose activamente en la planta), por lo que estas pueden varíar en los síntomas y grado de severidad. Las enfermedades producidas por virus de mayor importancia en los viñedos a nivel mundial son el enrollamiento de la hoja o GLD, enfermedad de la madera rugosa, el decaimiento/degeneración de las hojas y el complejo fleck, las cuales son consecuencia de infecciones por virus de ssRNA de las familias *Closteroviridae*, *Betaflexiviridae, Secoviridae*, y *Tymoviridae,* respectivamente. Además de estos, se conocen otros virus de ssRNA (de las familias *Alphaflexiviridae, Potyviridae, Virgaviridae, Bromoviridae* y *Tombusviridae*), dsRNA (familias *Reoviridae, Endornaviridae* y *Partitiviridae*), ssDNA (familia *Geminiviridae*) y dsDNA (familia *Caulimoviridae*) cuyos efectos aún no se conocen del todo.

Los insectos vectores de virus de vid incluyen a insectos cocoideos como piojos harinosos y otras cochinillas (Hemiptera: Coccoidea), así como alfareros (Hemiptera: Membracidae) y chicharritas (Auchenorrhyncha: Cicadellidae) (Martelli, 2017). Estos insectos inoculan los viriones al alimentarse de los conductos del floema de la planta (Deiz et al., 2008). Más específicamente, algunos géneros de piojo harinoso como *Heliococcus, Phenacoccu*s y *Planococcus*, y de cochinillas blandas como *Pulvinaria, Neopulvinaria Parthenolecanium, Coccus, Saissetia, Parasaissetia* y *Ceroplastes* son vectores de virus de los géneros *Ampelovirus* (familia *Closteroviridae*) y *Vitivirus* (familia *Betaflexiviridae*) (Maliogka et al., 2015; Martelli, 2017). Por su parte, algunos virus de los géneros *Marafivirus* y *Maculavirus* (familia *Tymoviridae*) son

transmitidos por los géneros de membrácidos y cicadélidos *Dalbulus*, *Macrosteles* y *Aconurella* (King et al., 2011).

1.1.10. Estudios de virómica para el descubrimiento e identificación de secuencias virales

La virómica (como una extensión de la metagenómica) es el estudio de la diversidad de virus presentes en una muestra mediante el análisis de las secuecias virales. Al implementar el uso de NGS en virómica, se han podido identificar virus en muestras de agua, suelo, heces, muestras intestinales de humanos y de tejidos de otros animales y plantas (Wooley et al., 2010; Rosario y Breitbart, 2011; Roossinck et al., 2015). Potencialmente, cualquier virus en las muestras, cultivable o no cultivable, nuevo o conocido, puede ser detectado a partir de un análisis de virómica con NGS, que incluye tres etapas principales: i) la extracción de ácidos nucleicos virales de la muestra de origen, ii) la secuenciación en alguna plataforma de NGS y iii) el análisis bioinformático (Anderson y Schrijver, 2010 ; Radford et al., 2012; Mokili et al., 2012; Nooij et al., 2018). Para cada una de estas etapas hay diferentes técnicas que pueden aplicarse de forma complementaria o excluyente dependiendo de los objetivos de la investigación (Hall et al., 2014; Roossinck et al., 2015).

1.1.10.1. Obtención de ácidos nucleicos virales

En virómica se deben tomar algunas consideraciones importantes: i) que la abundancia relativa de los ácidos nucleicos virales es relativamente baja respecto a la abundancia de los ácidos nucleicos procedentes del hospedero y de organismos como bacterias u hongos en la mayoría de las muestras, ii) que el tamaño promedio del genoma de las bacterias u hongos es mucho más grande con relación al tamaño de los genomas virales, y iii) en el caso de muestras procedentes de tejidos donde ocurren infecciones virales mixtas, es frecuente obtener diferentes títulos virales para cada uno de los virus contenidos en la muestra (Mokili et al., 2012). Por lo tanto, en un estudio de virómica se busca la eliminación o reducciónde los ácidos nucleicos no virales de la muestra y/o el enriquecimiento de los ácidos nucleicos virales, con los cuales se prepararán posteriormente las bibliotecas de secuencias que son el punto partida para la secuenciación masiva (Thurber et al., 2009; Conceição-Neto et al., 2015).

Una técnica usada en virómica para aumentar el título de ácidos nucleicos virales en una muestra es el enriquecimiento físico de partículas tipo virus o VLPs (por sus siglas en inglés) (Conceição-Neto et al., 2015; Roossinck et al., 2015). Generalmente comienza con un tratamiento de la muestra con cloroformo seguido de una digestión con DNasas y/o RNasas. El cloroformo rompe las membranas lipídicas de las células, exponiendo así su DNA y RNA, los cuales serán eliminados con la digestión subsecuente de nucleasas. Las partículas virales que permanecen en la muestra pueden concentrarse realizando uno o varios ciclos de homogenización, filtración y ultracentrifugación (Mokili et al., 2012; Hall et al., 2014). Luego de dicho procesamiento, es posible contener a las VLPs en un buffer o solución libre de restos de materia celular, y extraer los ácidos nucleicos que contienen. Se ha demostrado que una de las principales ventajas del enriquecimiento de VLPs es que permite la recuperación de viriones que se encuentran en alta y baja proporción (Roossinck et al., 2015). Por otro lado, también se ha demostrado que esta técnica puede generar sesgos en la diversidad de virus muestreados (Kleiner et al., 2015; Wu et al., 2015). Por ejemplo, el tratamiento con cloroformo puede eliminar la membrana lipídica de algunos virus con envoltura, lo que puede ocasionar la desestabilización de su cápside. El filtrado puede retener algunos virus de tamaño mayor al poro usado, mientras que en el ultracentrifugado se pueden perder algunas cápsides virales que no resistan las gravedades a las que se somete la muestra (Hall et al., 2014). Por otra parte, tampoco se pueden recuperar con enriquecimientos de VLPs a los virus que se encuentran en una fase del ciclo infectivo donde no se producen viriones (por ejemplo, como ocurre en el caso de algunos virus de plantas) (Flint et al., 2015).

Una alternativa al enriquecimiento de VLPs usada en virómica para reducir la proporción de ácidos nucleicos no virales (y aumentar la proporción de material genético viral) es el uso de fracciones del transcriptoma (el total de RNA presente en la muestra) (Pecman et al., 2017). El transcriptoma contiene tanto al RNA del hospedero como al RNA viral, por lo que se debe realizar uno o más pasos de selección de los RNAs presentes para remover la mayor cantidad de RNA no viral. Se ha demostrado que el RNA ribosomal (rRNA) es el componente más abundante del transcriptoma, representando alrededor del 50 a 85% del RNA total. Por lo cual, un paso de remoción del rRNA es esencial en casi cualquier análisis de secuenciación del transcriptoma. Las fracciones del transcriptoma que se usan más frecuentemente en la preparación de bibliotecas para virómica son el RNA total sin rRNA y los RNAs pequeños (sRNAs) (Pecman et al., 2017; Zhang et al., 2018). En el primer caso, la remoción del rRNA permite detectar aquellos virus con genoma de RNA o DNA (mediante sus transcritos), y si se omite el uso de oligo dT para la selección exclusiva de los mRNAs con extremos 3’ poliadenilados, se pueden recuperar también los genomas virales con extremos 3’ no poliadenilados (Wu et al., 2015). Por su parte, algunas especies de sRNAs (p. ej. los RNA de interferencia) en una muestra son producto del sistema de respuesta antiviral de las células y se

producen en grandes cantidades solo durante la infección viral en plantas, insectos y vertebrados, por lo que sus secuencias puede usarse para reconstruir los genomas virales a partir de los cuales se originaron (Roossinck et al., 2015; Aguiar et al., 2016).

Si bien ambas aproximaciones permiten una detección amplia de genomas virales, se ha demostrado que presentan algunas diferencias. El RNA total sin rRNA permite generar secuencias más largas (llamadas "contigs") durante el análisis bioinformático, dando una mayor **cobertura de secuenciación** al genoma viral, con lo cual posteriormente pueden detectarse regiones con similitud a las secuencias virales reportadas en las bases de datos. Esto la vuelve una aproximación más adecuada para el descubrimiento de nuevos virus (Pecman et al., 2017; Shi et al., 2018). Sin embargo, presenta la desventaja de que la mayor parte de las secuencias generadas después de la secuenciación siguen siendo de origen no viral y deben ser descartadas (Martin et al., 2012; Sadeghi et al., 2018), mientras que las secuencias virales que se encuentren en bajo título pueden no ser detectadas (Roossinck et al., 2015). Por su parte, los sRNAs permiten trabajar con menos secuencias no virales (reduce el ruido de la información de fondo), lo que permite un procesamiento bioinformático más rápido. Además, se ha demostrado que el uso de sRNAs es una aproximación más sensible para detectar aquellos virus que solo presentan intermediarios de RNA de doble cadena y también virus de DNA de cadena simple (Roossnick et al., 2015; Wu et al., 2015; Pecman et al., 2017). No obstante, debido a que se generan lecturas muy cortas, el ensamble *de novo* de genomas virales completos se dificulta más que cuando se trabaja con lecturas generadas a partir de RNA total sin rRNA (Pecman et al., 2017).

1.1.10.2. Secuenciación en plataformas de NGS

Las plataformas de NGS actuales son muy variadas en sus principios de funcionamiento y se clasifican típicamente en tecnologías de 2.ª y 3.ª generación (Anderson y Schrijver, 2010). Estas tecnologías han sustituido a los secuenciadores de 1.ª generación (que no se consideran NGS) en los estudios de metagenómica, debido a que tienen más alto rendimiento, menor coste de secuenciación y más versatillidad en el procesamiento de muestras (si bien ninguna ha mejorado la precisión de las tecnologías de 1.ª generación) (Liu et al., 2011; Heather y Chain, 2016). Actualmente, las tecnologías de NGS existen bajo diferentes marcas comerciales, de las cuales Illumina (de 2.ª generación) es la más utilizada en estudios de virómica (Kumar et al., 2017).

Hasta el momento, todas las plataformas Illumina funcionan mediante un principio llamado “secuenciación por síntesis” (Stark et al., 2019). El proceso de secuenciación en conjunto consta de 3 etapas principales, a saber, la generación de bibliotecas (llevada a cabo por el experimentador), seguida por la generación de grupos de secuncias (también llamadas “islas de secuencias” o “clusters”) y la secuenciación paralela de las islas por ciclos de síntesis de terminación reversible (estas últimas dos etapas son llevadas a cabo en la plataforma de NGS) (**figura 3**) (Metzker, 2010).

Una vez realizada la reducción de ácidos nucleicos de origen celular en la muestra (idependientemente de la técnica usada), se debe comenzar la preparación de bibliotecas de secuenciación. Este paso es crucial para todas las plataformas de NGS actuales, ya que los ácidos nucleicos iniciales (ya sea RNA o DNA) extraídos de la muestra de interés no pueden procesarse directamente. Una biblioteca de secuenciación es una colección de secuencias de dsDNA o RNA (sintetizados a partir de los ácidos nucleicos iniciales y por ende representativos de los mismos) de tamaño medio estandarizado, que son ligadas a secuencias artificiales llamadas adaptadores, las cuales a su vez permiten a la plataforma de NGS reconocer y “leer” las secuencias resultantes. Así pues, la preparación de las bibliotecas se considera una etapa más del proceso de secuenciación ya que consiste en la generación de ácidos nucleicos sintéticos aptos para un secuenciador, a partir de aquellos obtenidos del material biológico. En el primer paso de la preparación de bibliotecas a partir de RNA para plataformas Illumina, los ácidos nucleicos de origen (i.e., el RNA del material biológico) debe ser fragmentado en moléculas más pequeñas de tamaño medio, típicamente de 300-500 pb (**figura 3**, A). A continuación, se sintetizan secuencias de DNA complementario a los fragmentos (cDNA), y luego moléculas de dsDNA. Finalmente este dsDNA (llamado inserto) es ligado a secuencias de adaptadores (que contienen motivos de unión al fragmento, un índice para identificar la biblioteca en la que se encuentra y un motivo complementario a la celda de flujo del secuenciador) (Illumina 2015).

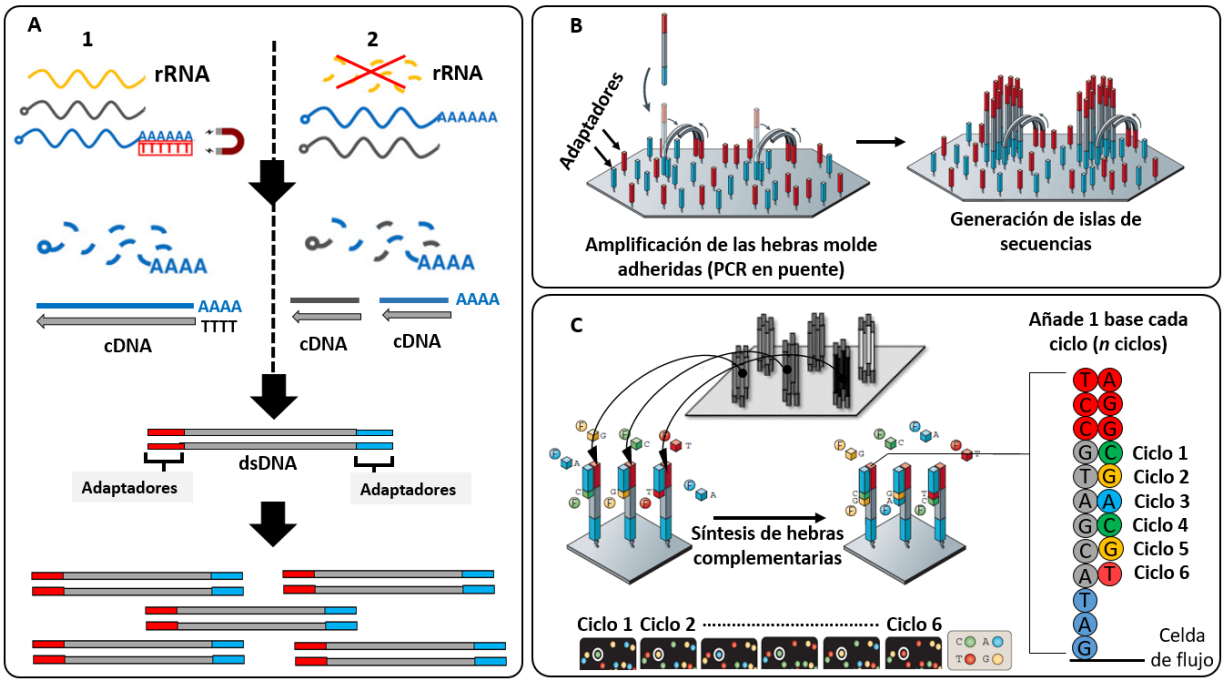

Figura 3. Secuenciación en plataformas Illumina. En (A) se muestra la preparación de bibliotecas a partir de la selección de RNAs con extremo 3' poliadenilado (1), o bien, a partir de RNA total, con la previa eliminación del rRNA (2). Una vez seleccionada la fracción de RNA a usar, se continúa con la fragmentación del RNA, seguido de la síntesis de cDNA, dsDNA, ligación de adaptadores y la amplificación de fragmentos de dsDNA por PCR. En el secuenciador (B y C), los fragmentos de dsDNA se unen a la celda de flujo. Las hebras unidas se amplifican mediante PCR en puente para formar "islas de secuencias" o "clusters". Finalmente (C), las secuencias de las islas sirven de molde para sintetizar hebras complementarias con nucleótidos marcados con fluoroforos. En las cajas negras al fondo de (C), indicadas con los ciclos 1 a 6 (el número real de ciclos es variable), cada punto de color representa una isla de secuencias que emite una señal colectiva de acuerdo al nucleótido añadido en ese ciclo. En cada ciclo, se muestra el color correspondiente al nucleótido añadido (encerrando en círculo el punto de la secuencia que se toma como ejemplo) en las secuencias nacientes de cada isla; dicha señal es leída por el equipo e interpretada como la base correspondiente. Se muestra un ejemplo de la secuencia molde adherida a la celda y la hebra naciente en el extremo derecho de (C). Figuras tomadas y modificadas de Metzker (2010) y Depledge et al. (2019).

Así, los fragmentos resultantes de dsDNA de la biblioteca son compatibles con la celda de flujo del secuenciador, que está cubierta por oligos complementarios a la secuencia de los adaptadores de los fragmentos (**figura 3**, B). La segunda etapa del proceso de secuenciación (la formación de islas de secuencias) comienza con la unión de las hebras (separadas) de los fragmentos a los oligos de la celda. Las hebras unidas son entonces amplificadas en un proceso iterativo de "PCR en puente" para formar millones de copias de cada secuencia (Metzker 2010; Radford et al., 2012; Stark et al., 2019). Las hebras sentido y antisentido del inserto deben procesarse por separado, por lo cual, luego de un primer PCR en puente usando ambas, se eliminan las hebras sentido o antisentido de la celda (y en una ronda posterior se procesan las que fueron eliminadas inicialmente). El proceso de amplificación anterior es estandarizado en tiempo, de ahí que el tamaño de los fragmentos deba también seguir un estándar.

Una vez generadas las islas (cada una con millones de copias de una misma secuencia), se comienza con la síntesis de cadenas complementarias usando nucleótidos marcados con fluoróforos (i.e., comienza la etapa final del proceso de secuenciación) (**figura 3**, C); por cada ciclo de síntesis se añade el mismo nucleótido en millones de cadenas nacientes, con lo cual se genera una señal luminosa aditiva, que puede leer el equipo y traducirse a una base determinada. Así pues, el equipo recibe simultaneamente las señales luminosas específicas de cada una de las islas por cada ciclo de síntesis, visualizadas en conjunto como puntos o zonas de color. Luego de varios ciclos, se forman millones de secuencias digitales llamadas "reads" de 100-300 bases de longitud (Metzker 2010; Stark et al., 2019).

1.1.10.3. Procesamiento bioinformático

Una vez que se obtienen los reads, se debe proceder al análisis de las secuencias. Los flujos de trabajo bioinformático para el procesamiento de reads generados por una plataforma Illumina son muy variados y dependerán del objetivo particular del estudio (Nooij et al., 2018). En los estudios dirigidos a la búsqueda e identificación de secuencias virales nuevas y conocidas, se incluyen típicamente: el análisis de calidad de los reads "crudos", el filtrado de secuencias no virales, el ensamble de reads para generación de secuencias más largas llamadas contigs, los análisis de similitud de secuencias y finalmente, el análisis de abundancia o cobertura de las secuencias virales encontradas (Wu et al., 2015; Viljakainen y Juravansuu 2020).

De manera general, el análisis de calidad permite descartar aquellos reads cuya secuencia no cumple con un valor de calidad establecido, usualmente con la escala de "Phred score", cuyo valor mide la probabilidad de error en la asignación de las bases en la secuencia del read. Los reads que pasan el filtro de calidad pueden entonces ser analizados de dos formas: mediante el mapeo o el ensamble (Bleidorn 2017). En el primer caso, los reads se someten a un alineamiento (mapeo) a secuencias conocidas, ya sean no virales (genes o genomas de organismos celulares como el hospedero o microorganismos asociados) o virales de interés que se crea que pueden estar representadas en la muestra, con lo cual se pueden clasificar los reads (Viljakainen y Jurvansuu, 2020). Alternativamente, los reads pueden someterse a un ensamble para reconstruir fragmentos más largos y representativos de las secuencias de origen llamadas contigs; esto puede llevarse a cabo usando una secuencia previamente conocida para guiar el ensamble (i.e., ensamble por referencia), o bien sin secuencia guía (i.e., ensamble *de novo*) (Bleidorn 2017). Los análisis de mapeo y ensamble pueden utilizarse de forma complementaria; un flujo de análisis común en virómica consiste en primero mapear los reads a secuencias no virales de referencia, recuperar los reads no mapeados y

después usar estos últimos para un ensamble *de novo* (Viljakainen y Jurvansuu, 2020). Esta última variante de ensamble es de especial interés en los análisis de virómica enfocados a la recuperación de contigs que representan secuencias “nuevas” o que no se han reportado previamente en bases de datos.

Para determinar si una secuencia es novedosa o no, determinar el organismo o virus del cual procede esa secuencia, la siguiente etapa del trabajo bioinformático es el análisis de similitud. El principio teórico de estos análisis es la **homología** inferida por similitud de la(s) **secuencia(s) problema** (en lo sucesivo denominadas como *query* o *queries,* para el plural) con las secuencias contenidas en las bases de datos (en lo sucesivo denominadas como **secuencias de referencia**, *subject* o *subjects,* para el plural) (Pearson, 2013). Dada la importancia central de este análisis en la virómica, se hará incapié en algunos aspectos básicos.

La herramienta más usada en los análisis de similitud es BLAST (Basic Local Alignment Search Tool) (Altschul et al., 1990) con sus diferentes programas (BLASTN: nucleótido-nucleótido, BLASTX: nucleótido-proteína y tBLASTX: nucleótido traducido-nucleótido traducido) (Camacho et al., 2008). El **valor de expectación *e*** reportado durante las búsquedas BLAST es una medida estadística utilizada para evaluar las coincidencias de una secuencia problema con las secuencias contenidas en una base de datos; indica el número de alineamientos que se espera ocurran por azar utilizando una **base de datos** de tamaño determinado, por lo que entre menor sea su valor para una base de datos determinada, mayor es la significancia del alineamiento (Pevsner 2015).

Las bases de datos más frecuentemente usadas para el filtrado y búsquedas de secuencias virales y no virales son las proporcionadas por el Centro Nacional de Información Biotecnológica (NCBI, por sus siglas en inglés), de proteínas, nucleótidos, no redundante, de referencia, de taxonomía, de dominios conservados, etc. (Sayers et al., 2011). Así mismo, el Instituto Europeo de Bioinformática del Laboratorio Europeo de Biología Molecular (EMBL-EBI, por sus siglas en inglés) (McWilliam et al., 2013) proporciona numerosas herramientas y recursos análogos o complementarios al NCBI para el análisis de secuencias. Además, existen bases de datos más específicas diseñadas para virómica como vFam (Skewes-Cox et al., 2014) o VOGDB (http://vogdb.org/).

1.1.11. El papel del análisis de secuencias en la taxonomía y descubrimiento de virus

La clasificación taxonómica de la diversidad viral es una tarea que se ha vuelto progresivamente más compleja durante las últimas dos décadas, no solo porque el ritmo con que se descubren nuevos virus ha incrementado drásticamente, sino porque la variabilidad genética que se observa en los grupos virales que se describen es cada vez mayor (Zhang et al., 2019). Para abordar la problemática que lo anterior plantea, se ha desarrollado toda una gama de técnicas, herramientas y estrategias moleculares, físicas y bioinformáticas para analizar y caracterizar las distintas propiedades de los virus (Wu et al., 2015).

La clasificación taxonómica de virus (al igual que la de organismos celulares) busca apegarse a un ordenamiento jerárquico que represente las relaciones evolutivas de las entidades a clasificar, es decir, que la agrupación de los taxa se aproxime o represente su historia natural (ICTV, 2020a). Por supuesto, los sistemas de clasificación taxonómica siempre son asintóticos a la verdadera historia evolutiva, puesto que están limitados a las propiedades que se conocen en un momento dado. Así pues, una taxonomía viral confiable debe tomar en cuenta dos aspectos fundamentales: i) recoger la colección de todas las particularidades de los virus, que los distinga de otras entidades biológicas y los defina como un universo de estudio concreto, y ii) considerar la mayor cantidad de propiedades generales y particulares de la diversidad de virus conocida para poder agruparla jerárquicamente y darle robustez al ordenamiento (ICTV, 2020a).

Entre los componentes que definen a los virus como entidades biológicas, el tipo y secuencia de su genoma son los elementos más fundamentales. Históricamente, la clasificación taxonómica en rangos supraespecíficos (como familias u órdenes) dependía en gran medida del análisis de secuencias, pero la clasificación más fina (géneros o especies) y el reconocimiento de especies recaían principalmente (aunque no exclusivamente) en otros aspectos biológicos como la morfología, estructura y composición del virión, serología, patogenicidad o rango de hospederos (Simmonds et al., 2017). Debido a la creciente necesidad de clasificar y ordenar toda la información proveniente de estudios de virómica, así como a la mayor disponibilidad de herramientas para el análisis de secuencias y para la inferencia de propiedades de los virus a partir de su secuencia, la clasificación taxonómica de virus actual da un mayor peso a los criterios genéticos, siempre y cuando se cumplan los requerimientos necesarios (Simmonds et al., 2017; Roux et al., 2019; ICTV, 2020a).

Así, en la clasificación taxonómica de virus basada en secuencias se consideran algunos aspectos esenciales como la ausencia de un gen universal para todos los genomas virales conocidos y su origen polifilético

(Koonin et al., 2006). Lo anterior implica que, a diferencia de los organismos celulares, la totalidad de virus no puede integrarse en un mismo árbol filogenético (Koonin et al., 2015). Por lo tanto, para las reconstrucciones filogenéticas se hacen agrupamientos de genomas que presentan un gen o grupos de genes compartidos homólogos (con origen monofilético) (Koonin y Dolja, 2013; Koonin et al., 2015). Los genes virales homólogos considerados como los más relevantes para dichos agrupamientos son aquellos considerados exclusivos de virus, i.e., que presentan únicamente homólogos distantes con genes de organismos celulares, y se denominan "genes marcadores" (Koonin et al., 2006; Koonin y Dolja, 2013).

En el caso de los genomas virales de RNA (a excepción de los que se replican por retrotranscripción), el gen que codifica para la polimerasa de RNA dependiente de RNA (RdRP) es uno de los casos más notables de conservación (de Farias et al., 2017; Venkataraman et al., 2018). La secuencia del gen codificante varía ampliamente entre genomas pero existe una alta conservación del **dominio** funcional de la secuencia proteíca y de la estructura tridimensional del centro catalítico (que contiene una región o subdominio de "palma de la mano αβ"), con los motivos llamados A, B y C como los más prominentes; los motivos A y C están conservados en la mayoría de las polimerasas, y el motivo B tiene una secuencia exclusiva de las RdRPs (Gorbalenya et al., 2002). Por otro lado, si bien las RdRPs de todos los genomas virales son homólogas, se ha comprobado que las RdRPs de los virus de +ssRNA y la de virus de dsRNA están más estrechamente relacionadas entre sí que con las encontradas en virus de -ssRNA (Koonin y Dolja, 2015).

Además de la conservación de la RdRP, se han reconocido grupos de genes marcadores (**genes ortólogos**) en módulos con un orden canónico (altamente conservados) en virus de RNA, que a su vez han permitido definir grandes conjuntos de virus (denominados informalmente como "superfamilias") como los de picornavirus-like (también llamados picorna-like), alphavirus-like y flavivirus-like (Koonin y Dolja, 2015; Koonin et al., 2020). Por ejemplo, el módulo conservado de genes que codifican proteínas no estructurales en los picornavirus-like (el grupo más grande y diverso) es S3H-VPg-PC3-RdRP (S3H: superfamilia 3 de helicasa; VPg: proteína ligada al genoma; PC3: proteasa quimotripsina), además de uno o más genes codificantes de proteína de cápside JRC (o "Jelly-roll capsid protein") (Gall et al., 2008). Sin embargo, el arreglo anterior puede presentar variaciones entre los miembros de la superfamilia, que comprende al orden *Picornavirales* y algunas familias de +ssRNA (p. ej. familias *Caliciviridae, Secoviridae* y *Comoviridae*) y dsRNA (p.ej. familia *Totiviridae* y otras familias no pertenecientes al orden *Reovirales*) (Koonin y Dolja, 2015). Por su parte, el grupo alphavirus-like incluye al orden *Tymovirales* así como otras familias que infectan plantas (como *Bromoviridae, Closteroviridae* y *Virgaviridae*) y animales (como *Hepeviridae* y *Tetraviridae*). Los miembros de este grupo conservan el módulo de genes asociado a la replicación como MTR-(PRO)-S1H-RdRP (MTR: metil-transferasa; PRO: papaína-proteasa, puede no presentarse; S1H:

superfamilia 1 de helicasa), así como genes que codifican para proteínas de movimiento (MP) o cápside (CP), aunque también se presentan variaciones entre los miembros del grupo (Martelli et al., 2002; King et al., 2011; Koonin et al., 2015).

En el caso de los virus de dsDNA, que representan la gran mayoría de los virus que infectan procariotas (aunque también se encuentran grupos que infectan eucariotas), se han encontrado grupos de genes ortólogos que pueden definir nuevos taxones como ocurre con el orden propuesto "*Megavirale*s" (Colson et al., 2013). Este grupo, contiene a los virus nucleocitoplasmáticos de dsDNA (NCLDV), comprende 7 familias (*Iridoviridae, Poxviridae, Mimiviridae, Phycodnaviridae, Ascoviridae, Asfaviridae* y *Marseilleviridae*) y fue definido en función de un grupo de 50 genes homólogos, que se relacionan solo distantemente con otros virus de dsDNA como los de las familias *Baculoviridae, Nudiviridae* y *Polydnaviridae* (Koonin y Yutin, 2010; Colson et al., 2013). Considerando lo anterior, se ha utilizado la homología de genes, tanto en genomas de RNA como de DNA, así como otras propiedades de las secuencias, como guía del agrupamiento de virus en jerarquías taxonómicas progresivamente más restrictivas.

Tomando lo anterior como fundamento, en la práctica la asignación taxonómica con base en el análisis de genomas virales consiste en revelar la homología de una secuencia problema con un grupo particular de virus, a pesar de que la similitud de secuencia pueda ser muy limitada (Li et al., 2015; Shi et al., 2016). En principio, si las búsquedas de homología basadas en similitud indican un alto porcentaje de **identidad**, la asignación se facilita, pero en caso contrario (que ocurre frecuentemente en estudios de virómica) se siguen estrategias alternativas (Pearson 2013). Entre estas, se pueden determinar las características principales de la secuencia como la organización genómica mediante la predicción de marcos de lectura abierta u ORFs (regiones que abarcan genes putativos), así como el orden, número y posición de los dominios (o módulos) conservados en las secuencias de proteínas (y sitios de corte) correspondientes a dichos ORFs del genoma (Simmonds et al., 2019). Así, pueden identificarse módulos característicos de ciertos grupos virales y en función de su disposición en el genoma, se puede proponer una clasificación tentativa. Por ejemplo, una secuencia que presente una similitud limitada (≤30%) con miembros del orden *Picornavirales*, pero que contenga una organización genómica con el módulo canónico no estructural de este grupo (S3H-VPg-3CPro-RdRP) y/o el módulo canónico estructural (3 dominios de cápside JRC), puede en principio considerarse como una secuencia relacionada (Yasmin et al., 2020). Posteriormente, puede vincularse con alguna de las familias si se sabe que la secuencia representa un genoma monopartita (como en las familias *Picornaviridae* o *Dicistroviridae*) o multipartita (como en la familia *Secoviridae*), si contiene un ORF (familias *Picornaviridae* e *Iflaviridae*) o más (como la familias *Dicistroviridae* y *Marnaviridae*), si el

módulo canónico se encuentra antes o después del módulo estructural (3 dominios de cápside JRC) o si presenta sitios de corte en la proteína predicha (si es o no una poliproteína con diferentes dominios funcionales o se escinde en unidades separadas) (Gall et al., 2008). Además, se pueden identificar secuencias no codificantes distintivas como sitios internos de entrada al ribosoma o "IRES", secuencias repetitivas (terminales) conservadas, motivos de pliegues estructurales, entre otras (Jan, 2006; Simmonds et al., 2017). Cabe mencionar que, a diferencia de los virus de DNA, los genomas virales de RNA contienen en su mayoría genes marcadores, por lo que reconocer la homología de una secuencia viral putativa de RNA, y por ende recuperarla, durante el análisis de secuencias es relativamente más sencillo que en el caso de una secuencia putativa de DNA (Shi et al., 2016; Koonin et al., 2015). En los casos donde la identificación de regiones homólogas con base en la comparación de secuencias falla, es posible complementar el análisis con otras aproximaciones como el modelado *in silico* de estructura tridimensional de proteínas y su posterior comparación con bases de datos estructurales para deducir la homología (Hily et al., 2018).

Es importante mencionar que los análisis de la organización genómica se facilitan con las bases de datos actuales de genes ortólogos o de proteínas y genomas de referencia, ya que proveen una anotación funcional de las secuencias y los dominios que se encuentran en ellas. De esta forma, en principio es posible asignar una función putativa a las secuencias *queries* en función de sus hits con las bases de datos, y además obtener una primera aproximación a los grupos con los que presenta mayor afinidad. Así, algunas bases de datos como la Pfam (El-Gebali et al., 2019) agrupan a las familias de RdRPs con base en análisis de posición específica, generando grupos de secuencias (conocidos como perfiles ocultos de Markov o pHMMs) y luego etiquetando a los pHMMs con la afiliación taxonómica de los virus cuyas secuencias se incluyeron. De esta manera, se encuentran las polimerasas de algunos virus de +ssRNA (como la Pfam00978, RdRP_2), las RdRPs de virus de dsRNA (como los de la familia *Reoviridae,* RdRP_5), o el grupo de RdRPs de virus de -ssRNA, especialmente aquellos del orden *Mononegavirales* (familia Pfam00946, mononeg_RdRP). Por otro lado, algunas bases de datos, como eggNOGs (Huerta-Cepas et al., 2016) o VOGDB (http://vogdb.org/), proporcionan grupos de genes ortólogos y su respectiva anotación de acuerdo a su exclusividad en genomas virales (i.e., si son genes marcadores o no).

Finalmente, algunos autores como Roux et al. (2019) han propuesto estándares internacionales para el análisis, reporte y organización de las secuencias virales en las bases de datos de acceso público, denominados colectivamente como Información Mínima de Genomas de Virus No Cultivados o MIUViGs (por sus siglas en inglés). Entre los tópicos más importantes de dichos estándares destacan i) la predicción *in silico* del hospedero, ii) la asignación taxonómica y iii) la diversidad y distribución potencial. Todos estos

puntos son importantes y complementarios para inferir algunas de las propiedades biológicas de los virus. En cuanto al primer punto, proponen diferentes aproximaciones bioinformáticas para la predicción del rango posible de hospederos, de entre las cuales se menciona que la asignación taxonómica previa es el método más fiable (con el razonamiento de que el rango posible de hospederos va en función del grupo taxonómico al que pertenece), indicando a su vez que esta última puede realizarse con base en la detección de genes marcadores. Respecto al punto dos, se ha propuesto el uso de vOTUs (unidades taxonómicas operacionales virales) para asignar a las secuencias virales putativas a un rango de especie, cuando estas cumplen con un 95% de identidad nucleotídica promedio (%ANI, respecto a la secuencia más corta), a lo largo de una fracción ≥85% del alineamiento pareado con la secuencia de referencia. En cuanto al tercer punto, se propone que los análisis de abundancia de las secuencias en diferentes localidades geográficas o regiones anatómicas del hospedero provea una aproximación a los aspectos biológicos del virus como su distribución geográfica, o tropismo celular a los tejidos del hospedero respectivamente.

En síntesis, el análisis de secuencias virales permite no solo obtener información y criterios para establecer un sistema de clasificación taxonómica, sino también inferir algunas propiedades relativas a la biología de los virus como la estructura y organización del virión, el rango de hospederos o el tropismo celular. Esto ha permitido que se dé mayor peso a los criterios bioinformáticos en la taxonomía actual de los virus, bajo la idea de que la clasificación con base en las secuencias no está limitada por la ausencia de otras propiedades biológicas, sino por la capacidad de predicción de las herramientas bioinformáticas disponibles (Simmonds et al., 2017). Pese a esto, es evidente que existe información que escapa a los alcances del análisis de secuencias (como la patogenia, los ciclos de replicación o la dinámica de la infección en los hospederos), por lo que es importante recalcar que los criterios biológicos y bioinformáticos son siempre complementarios para la definición de especies virales, que es la tarea última de la clasificación taxonómica (Simmonds et al., 2017; ICTV, 2020a).

1.2. Justificación

Los virus pueden impactar de forma importante en la ecología de las distintas especies hospederas. Específicamente hablando de ISVs, se ha reportado ampliamente su impacto sobre la fecundidad, reproducción y supervivencia de sus insectos hospederos. Así mismo, existe evidencia sobre la influencia de los ISVs en la competencia de sus hospederos como vectores de patógenos. En conjunto, estas características pueden trasladarse en aplicaciones para el control de plagas. Los estudios de búsqueda y

caracterización de especies virales en hemípteros fitófagos se han limitado a áfidos o psílidos, siendo nula la información existente acerca del viroma de pseudocóccidos. Determinar cuáles (y cuántos) son los genomas virales que pueden encontrarse en *P. ficus* representa un primer paso en la búsqueda e identificación de potenciales nuevos ISVs para esta especie.

1.3. Objetivos

1.3.1. Objetivo general

Identificar fragmentos genómicos de virus de RNA y DNA a partir de la secuenciación de RNA aislado de *Planococcus ficus.*

1.3.2. Objetivos específicos

1) Generar bibliotecas para secuenciación masiva usando RNA total sin rRNA a partir de muestras de *P. ficus*.

2) Reconstruir fragmentos de genomas virales *in silico* mediante análisis bioinformáticos de datos de secuenciación masiva.

3) Determinar las relaciones filogenéticas de los fragmentos genómicos virales obtenidos.

Capítulo 2. Metodología

2.1. Colecta y procesamiento de muestras

2.1.1. Selección de sitios de muestreo

Los muestreos de *P. ficus* se realizaron durante el 2018 y 2019 en trece ranchos/viñedos del valle de Guadalupe y un viñedo del valle de San Vicente, ubicados al norte y sur de la ciudad de Ensenada, B.C. respectivamente. Todos los muestreos se realizaron en colaboración con el Comité Estatal de Sanidad Vegetal de Baja California (CESVBC), con base en los registros históricos de dos años (2016-2018) de presencia y grado de infestación en cada viñedo. Para la selección de los sitios de muestreo se consideraron los siguientes puntos: i) que el viñedo contara con un alto grado de infestación de *P. ficus*, ii) que se contara con el permiso de los productores para muestrear, y iii) que el sitio se encontrara lo más distante posible de otro sitio de muestreo. A petición del CESVBC y por razones de confidencialidad de la información, la identidad y ubicación exacta de los viñedos que se muestrearon no se detalla.

2.1.2. Muestreo

Los muestreos de *P. ficus* dentro de cada sitio se realizaron de manera dirigida, no sistemática. La presencia de signos de fumagina y/o de cera o mielecilla en el tronco o ramas de las vides se usó como indicador visual para elegir las plantas a muestrear en cada surco. Una vez ubicada una planta con sospecha o presencia de piojos, se realizó un descortezado del tronco o ramas para inspeccionar y encontrar las agregaciones de piojos. La toma de los piojos en la superficie expuesta se realizó mediante un aspirador entomológico que cuenta con un tubo colector, o bien utilizando palillo y pincel. Las muestras de colecta de cada viñedo se guardaron en bolsas plásticas herméticas y cada una se distinguió por sitio y fecha de colecta (se denominaron como P1, P2, P3, y así sucesivamente). Una vez en el laboratorio, las muestras se refrigeraron a 4 °C por un periodo no mayor a una semana para continuar con la limpieza (ver sección siguiente) y finalmente se congelaron a -20 °C. Debido a que algunas muestras de colecta (P1, P2, etc.) provenían del mismo sitio de muestreo pero se colectaron en diferente fecha, y ya que se consideró solo una muestra por cada uno de los 14 sitios diferentes de muestreo, en lo siguiente la numeración de las muestras utilizadas no es continua (las muestras consideradas son las nombradas como: P1, P2, P3, P4, P5, P7 P8, P9, P10, P11, P12, P16, P17 Y P18).

2.1.3. Limpieza de muestras

Las muestras se limpiaron para remover remanentes de corteza, otros insectos, o micelio de hongo que se hubiesen colectado en campo. Esto permitió aumentar la concentración de piojos en la muestra, además de facilitar las extracciones posteriores de DNA y RNA.

2.2. Identificación molecular de *P. ficus*

Se realizó la extracción de DNA total para una fracción de individuos de cada sitio de colecta y después se realizó la identificación molecular de la(s) especie(s) respectivas en cada submuestra por PCR y secuenciación del gen de la citocromo oxidasa subunidad I (*COI*).

2.2.1. Extracción de DNA total

En la extracción de DNA total se siguió un protocolo modificado de Doyle y Doyle (1990), utilizando ejemplares de hembras en los últimos estadios de desarrollo. Se consideraron 13 de los 14 sitios de muestreo (ya que no se contaba con muestra suficiente del sitio restante) y dos réplicas de un sitio extra (sitio CC, un viñedo ubicado en el Valle de Guadalupe), donde se ha demostrado la presencia de otras especies de piojo harinoso además de *P. ficus* (Carrillo Tripp, 2019; datos sin publicar). Se colocaron de 5 a 10 individuos por sitio de colecta en tubos de microcentrífuga de 1.5 mL y se les añadieron 120 µL de buffer de extracción (CTAB 2% w/v, NaCl 1.4 M, EDTA 20 mM, Tris-HCl 10 mM pH=8, β-mercaptoetanol 0.2%). Los tubos se incubaron a temperatura ambiente por 5 min y se homogenizó la muestra con ayuda de un pistilo de plástico, luego de esto se incubaron por 30 min a 65 °C. Posteriormente se añadieron a cada tubo 16 µL de acetato de amonio 10 M y se incubaron los tubos en hielo por 10 min. A continuación, se agregó 1 volumen de cloroformo:alcohol isoamílico (24:1) frío y se mezcló en vórtex. Los tubos se centrifugaron por 15 min a 13000 x g y se transfirió el sobrenadante a tubos nuevos. Se añadieron 0.6 volúmenes de isopropanol frío, se mezcló por inversión y se dejó precipitar la muestra por 12 hrs. a -20 °C. Posteriormente, los tubos se centrifugaron a 13000 x g por 15 min y se decantó el sobrenadante. La pastilla formada se lavó dos veces con etanol frío al 70%, centrifugando a 13000 x g por 5 min. El exceso de alcohol del último lavado se eliminó calentando el tubo a 65°C por ~10 min y la pastilla se resuspendió en 30 µL

de agua libre de nucleasas (previamente tratada con dietil-pirocarbonato o DEPC, en lo posterior referida como agua DEPC a menos que se especifique de otra forma). Finalmente, se verificó la concentración y pureza del material obtenido en un equipo Nanodrop® 2000 (ThermoScientific, Waltham, Massachusetts, EUA) y se visualizó la integridad del DNA extraído mediante electroforesis en gel de agarosa al 0.8% teñida con bromuro de etidio, utilizando un fotodocumentador UV-Vis BioRad (Bio-Rad Laboratories, Inc., Hercules, CA, EUA).

2.2.2. PCR multiplex y uniplex

Se siguió un protocolo de PCR multiplex de Daane et al., 2011 utilizando los primers diseñados por dichos autores, y están dirigidos al gen *COI* y son específicos para 7 especies de piojo harinoso: *Pseudococcus calceolariae*, *Ps. longispinus*, *Planococcus ficus*, *Ps. maritimus*, *P. citri*, *Ps. viburni* y *Ferrisia gilli*. El tamaño de la banda amplificada en la PCR se toma como referente para la identificación, ya que a cada especie le corresponde un tamaño de banda particular: 650 pb (*Ps. calceolariae*), 600 pb (*Ps. longispinus*), 450 pb (*P. ficus*), 400 pb (*Ps. maritimus*), 350 pb (*P. citri*), 250 pb (*Ps. viburni*) y 150 pb (*F. gilli*).

Para cada una de las 13 muestras de DNA total, la reacción de PCR se realizó en un volumen final de 10 µL con las concentraciones finales de los siguientes reactivos: 0.2 mM de dNTPs, 0.025 U/µL de DreamTaq DNA polimerasa (ThermoScientific, Waltham, Massachusetts, EUA), 1X de DreamTaq buffer (ThermoScientific, Waltham, Massachusetts, EUA), 0.2 µM de cada primer (7 primers *Forward* (Fw) específicos y un *Reverse* (Rv) universal) y 1 µL de DNA a una concentración de 90 ng/µL. El programa de PCR usado fue: 95°C, 30seg. – 30 ciclos (94°C, 30 seg. – 53°C, 90 seg. – 72°C, 90 seg.) – 72°C, 10 min. Con los productos de PCR obtenidos se realizó una electroforesis en gel de agarosa al 2% teñida con bromuro de etido. Posteriormente, se analizó el gel en un fotodocumentador UV-Vis BioRad (Bio-Rad Laboratories, Inc., Hercules, CA, EUA).

Para descartar la posible presencia de otras especies de piojo harinoso distintas a *P. ficus*, se consolido el DNA de todas las muestras en un solo volumen (llamado “DNA final”) y se realizaron PCRs uniplex, utilizando los primers de las especies *P. ficus, P. maritimus* y *P. citri*. Los PCR uniplex se realizaron utilizando el mismo programa para la PCR multiplex (modificando el tiempo de alineación a 30 seg) y un volumen de reacción de 50 µL con los reactivos a las mismas concentraciones descritas anterioremente, ajustando

únicamente la cantidad de DNA final a 30 ng totales. Los productos de los PCR uniplex se analizaron por electroforesis de manera similar a los productos de la PCR multiplex.

2.2.3. Secuenciacion de los productos de PCR uniplex

Se realizó una secuenciación capilar (secuenciación Sanger) de los productos de cada PCR uniplex realizado, así como un análisis de BLASTN en la plataforma del NCBI (https://blast.ncbi.nlm.nih.gov/Blast.cgi) y un alineamiento de los fragmentos obtenidos con el programa ClustalW de la plataforma EMBL-EBI (https://www.ebi.ac.uk/Tools/msa/clustalo/) para confirmar la identidad de *P. ficus*.

2.3. Construcción de bibliotecas a partir de RNA total sin rRNA y secuenciación

2.3.1. Extracción de RNA total

El RNA total de las 14 muestras de piojo consideradas se extrajo con el reactivo TRIzol LS® (Thermo Fisher Scientific® Inc., Waltham, MA, EUA) siguiendo las especificaciones del fabricante. El protocolo seguido constó de dos rondas de extracción, utilizando como material inicial 100-130 mg de masa de piojos (principalmente de hembras adultas) para cada sitio. La primera ronda de extracción, se realizó con TRIzol LS® diluido 3:1 con agua DEPC y siguiendo el protocolo del fabricante hasta la separación de fases acuosa y orgánica. A continuación, la fase acuosa ontenida se tomó como material de inicio de la segunda ronda de extracción, donde se utilizó TRIzol LS® concentrado y se siguió el protocolo del fabricante hasta el final, haciendo la resuspensión de la pastilla de RNA en 100 µL de agua DEPC. Al finalizar la extracción, se midió la concentración y la relación de absorbancia 260/280 de cada muestra en un equipo Nanodrop® 2000. El RNA total extraído se almacenó a -80°C.

2.3.2. Muestras de secuenciación y limpieza del RNA

El RNA total extraído de cada sitio de colecta se consolidó en grupos, los cuales se usaron a su vez como material de inicio en la generación de bibliotecas de secuenciación. A partir de este punto, se refiere a cada una de estas agrupaciones como “muestras de secuenciación final o MF”. Se generaron 3 muestras de secuenciación finales (MF1, MF2 y MF3) que incluyen a los 14 sitios de muestreo. Por razones de la estandarización y temporalidad en que se desarrolló el protocolo de limpieza con DNasa y otros protocolos posteriores, la MF2 incluyó solamente 3 sitios, mientras que las MF1 y MF3 incluyeron 5 y 6 sitios, respectivamente.

Se realizó la limpieza del RNA total con el kit Quick-RNA™ MiniPrep Plus (Zymo Research® Corp., Irvine, CA, EUA) utilizando el tratamiento con DNasa incluida en dicho kit. Se comenzó tomando 2500 ng totales de RNA de cada sitio correspondiente a la MF2 y 3500 ng totales de RNA de cada sitio correspondiente a las MF1 y MF3, depositándolos de acuerdo a su respectivo agrupamiento, en tubos de microcentrifuga con 300 µL de agua DEPC. A continuación, se añadieron 300 µL de buffer de lisis (*RNA lysis buffer*) del kit al tubo de cada muestra y se mezcló el volumen total por pipeteo. El volumen resultante se transfirió a una columna “amarilla” del kit y se centrifugó a 12000 x g por 30 seg. El volumen recuperado en el tubo de elución se mezcló con 0.5 volúmenes de etanol al 95% y se transfirió a una columna “verde” del kit. Se centrifugó a 12,000 x g por 30 seg. A partir de este punto se continuó con el módulo de purificación del RNA como lo especifica el protocolo del fabricante, incluyendo el tratamiento con DNasa. El volumen final de elución fue de 30 µL. La integridad de los RNAs resultantes se verificó en un gel de agarosa al 2% revelado con un fotodocumentador UV-Vis BioRad. Se cuantificaron las concentraciones finales y se registraron las relaciones de absorbancia A260/A280 y/o A260/A30 con un equipo Nanodrop® Lite (MF1) y Nanodrop® 2000 (MF2 y MF3). La concentración de RNA final se corroboró con un fluorómetro Qubit® (ThermoFisher Scientific®, Waltham, MA, USA) usando el kit Qubit® RNA BR Assay con rango de 20-1000 ng (ThermoFisher Scientific®, Waltham, MA, EUA).

2.3.3. Generación de bibliotecas y secuenciación

A partir de las tres MFs de RNA anteriores, se generaron 3 bibliotecas de secuenciación. Para la depleción de rRNA del RNA total se utilizó el kit Zymo-Seq RiboFree® Universal cDNA kit (Zymo Research® Corp., Irvine, CA, USA). La cantidad de RNA total inicial usada de las MF1, MF2 y MF3 fue de 1000 ng, 700 ng y

1000 ng totales de RNA, respectivamente, los cuales se colocaron separadamente en 8 µL de agua libre de DNasas/RNasas incluida en el kit. A continuación, se realizaron las etapas de alineación de primers, transcripción reversa, incubación de pre-depleción, reacción de depleción y terminación de la depleción, especificadas en el protocolo del kit. El DNA complementario (cDNA) obtenido al final de dichas etapas para cada MF se limpió con perlas magnéticas integradas en el kit Zymo-Seq RiboFree®, eluyendo el material final en un volumen de 25 µL de buffer de elución de DNA. Antes de proseguir, se realizó una cuantificación con fluorómetro Qubit® del cDNA recuperado con el kit Qubit® ssDNA Assay (Thermo Fisher Scientific Inc., Waltham, MA, EUA) cuyo rango de detección es de 1-200 ng.

A continuación, se inició con la generación de las bibliotecas usando el kit TruSeq® RNA Sample Preparation V2 de Illumina® (Illumina® Inc., San Diego, CA, EUA) siguiendo el protocolo *low sample*. La biblioteca de MF2 se generó primero y luego se generaron las bibliotecas de MF1 y MF3. Al cabo de este proceso las bibliotecas generadas fueron nombradas como PV1, PV2 y PV3, con la numeración correspondiente a la MF de RNA de origen. Se inició el protocolo del kit TruSeq® a partir de la etapa de síntesis de DNA de doble cadena (dsDNA) usando el cDNA de cada muestra obtenido anteriormente y se continuó con las etapas de reparación de extremos, adenilación de extremos, ligación de adaptadores y enriquecimiento de fragmentos de dsDNA por PCR como lo especifica el protocolo del fabricante. Se realizaron dos limpiezas mediante perlas magnéticas para la biblioteca 2, una al 1X y otra al 0.75X, mientras que para las bibliotecas 1 y 3 se realizó únicamente la limpieza con perlas al 0.75X. Luego de las limpiezas, todos los productos de PCR fueron visualizados en un gel de agarosa al 4% teñido con GelRed® (Biotium®, San Francisco, CA, EUA) que se reveló con un fotodocumentador UV-Vis BioRad, para verificar la obtención del barrido esperado de fragmentos con rango de 300 a 500 pb. La concentración final de cada biblioteca se cuantificó con un equipo Nanodrop® 2000 y con fluorómetro Qubit®, empleando con este último el kit Qubit® dsDNA High Sensitivity Assay Assay (Thermo Fisher Scientific Inc., Waltham, MA, EUA), que cuenta con un rango de detección de 0-100 ng. Adicionalmente, la distribución de tamaño de la población de fragmentos de dsDNA de las bibliotecas obtenidas se analizó con un chip de DNA preparado con el kit Agilent DNA 1000 (Agilent Technologies Inc., Santa Clara, CA, EUA) y fue leído en el equipo Agilent 2000 Bioanalyzer.

La concentración en nanomoles (nm) de cada biblioteca generada se calculó mediante la siguiente ecuación:

$$nM = \left[\frac{\frac{ng}{\mu L}}{(660)(tamaño\ medio\ de\ fragmentos)}\right] * 10^6 \quad (1)$$

Las 3 bibliotecas fueron normalizadas a la concentración de 8 nM, diluyendo la muestra en el volumen apropiado de buffer Tris-HCl pH= 8.5 con Tween 20 al 0.1 %. Una vez a 8 nM, se colocaron 10 µL de cada biblioteca en un nuevo tubo para obtener un grupo de bibliotecas de 30 µL a 8 nM. Las bibliotecas generadas en este trabajo se secuenciaron en una línea de celda de flujo de la plataforma HiSeq® X de Illumina® con una química de 2x150 bases y con un rendimiento estándar de 355 millones de reads/línea en la compañía Novogene (Novogene corporation, Inc., Sacramento, California, USA). Dado que en este trabajo se estimaron ~30 millones de reads para cada biblioteca, se utilizó el ~26% de la línea (~90 millones de reads en total).

2.4. Análisis bioinformático para las secuencias de bibliotecas de RNA total sin rRNA

Para el procesamiento bioinformático de los datos obtenidos de la secuenciación (reads) de las 3 bibliotecas de RNA total sin rRNA, se adaptó un flujo de trabajo basado en los análisis realizados por Feng et al. (2017), Shi et al. (2016) y otros estudios de virómica en insectos u otros organismos eucariontes (Sadeghi et al., 2018; Gilbert et al., 2019; Roux et al., 2019; Batson et al., 2020; Viljakainen & Jurvansuu, 2020). Las secuencias de cada biblioteca (reads o contigs) se analizaron de forma independiente en cada etapa, esto es, no se combinaron en un solo archivo para ser procesadas, a menos que así se indique. Un resumen de los análisis, programas y bases de datos, se presenta en la **tabla 1**, con una breve descripción de la finalidad de cada etapa.

2.4.1. Control de calidad para reads crudos

Los seis archivos fastq (dos por cada biblioteca, con los reads sentido o 1P y antisentido o 2P) se sometieron a un control de calidad. Estos archivos fueron procesados con el software Trimmomatic v. 0.36 (Bolger et al., 2014) para la remoción de adaptadores e índices del inserto (ligados a los fragmentos de dsDNA durante la preparación de bibliotecas), así como para el análisis y recorte de la secuencia de los reads usando el parámetro de “sliding window” de 4:15 (análisis del read en ventanas de 4 bases, realizando un corte si el valor promedio de Phred score en la ventana es < 15).

Los 6 archivos fastq fueron analizados en conjunto con el sotware multiQC (Ewels et al., 2016), antes y después del corte de calidad para verificar que los reads cumplieran con los parámetros de calidad especificados.

Tabla 1. Resumen del flujo de trabajo bioinformático. Ver texto para referencias y versiones de programas y bases de datos.

Análisis incluidos en el flujo de trabajo	**Programa(s)**	**Breve descripción**
Control de calidad para reads crudos	Trimmomatic	Se retiran adaptadores de los reads crudos y se descartan aquellos con un valor de Phred score menor al valor especificado.
	MultiQC	
Filtrado de reads no virales	Bowtie2 con secuencias de referencia	Se mapean los reads a una lista de secuencias de organismos celulares; los reads mapeados son eliminados.
Ensamble *de novo*	Megahit	Se generan contigs a partir de los reads del paso anterior.
Filtrado de contigs preliminares no virales	BLASTN con NT	Se realizan análisis de similitud entre los contigs generados y las secuencias de las bases de datos no redundantes del NCBI; se descartan aquellos con mejores hits a secuencias de organismos celulares.
	BLASTX con NR	
Reensamble de contigs preliminares	Cap3	Se reensamblan los contigs recuperados del análisis previo
Búsqueda de ORFs y predicción de secuencias proteicas	Prodigal	Se hace una búsqueda de ORFs en los contigs del reensamble y la posterior predicción del producto proteico sobre los ORFs.
Búsqueda de secuencias virales a partir de proteínas predichas	Hmmsearch con VOGs/Pfam	Se realizan análisis de similitud entre las secuencias de proteínas predichas para los ORFs anteriores y las bases de datos de i) dominios conservados, ii) grupos de genes ortólogos y iii) secuencias virales de referencia.
	RPS-BLAST con CDD	
	BLASTP con RefSeq	
Revisión manual de contigs y recuperación de secuencias virales putativas	-	Se inspeccionan y recuperan aquellos contigs cuyas proteínas mostraron similitud a secuencias de grupos selectos de virus de RNA o DNA. Además, se inspeccionan (cuando es relevante) las regiones no codificantes de los contigs.
Análisis de cobertura de los contigs seleccionados	Bowtie2	Se mapean los reads filtrados por calidad a los contigs seleccionados del paso anterior.
Análisis filogenéticos	MAFFT	Se construyen alineamientos que posteriormente se editan y se usan para reconstrucciones filogenéticas.
	TrimmAl	
	RAxML	
	iTol	

2.4.2. Filtrado de reads no virales

Los reads correspondientes al hospedero (*P. ficus*), a bacterias, a la microbiota asociada a este hospedero y a posibles fuentes de contaminación de laboratorio, fueron sustraídos mediante un mapeo a genomas/genes de referencia descargados del NCBI (**Anexo A**) con el programa Bowtie2 v. 2.2.3 (Langmead y Salzberg 2012), en la modalidad –very sensitive. Para cada biblioteca, se retuvieron solo los pares de reads (1P y 2P correspondientes) que no mapearon a dicha lista de genomas y se combinaron en un solo archivo fasta mediante la paquetería Samtools v. 1.7 (Li et al., 2009).

2.4.3. Ensamble *de novo*

La generación de contigs a partir de los reads se realizó utilizando el ensamblador Megahit v. 1.1.3 (Li et al., 2015a). Las métricas de los ensambles se obtuvieron con el programa stats de la paquetería BBTools v. 38.70 del JGI (Sourceforge.net/projects/bbmap/). Los contigs obtenidos de esta etapa se denominan "preliminares".

2.4.4. Filtrado de contigs preliminares no virales

Los contigs preliminares con tamaños ≥ 500 nt se compararon contra las bases de datos NT (https://www.ncbi.nlm.nih.gov/nuccore/) y NR (https://www.ncbi.nlm.nih.gov/protein/) del NCBI, utilizando los software BLASTN y BLASTX de la paquetería ncbi-blast v. 2.4.0. En los análisis anteriores, se especificó únicamente el reporte de aquellas secuencias cuyos hits tuvieran un valor de e ≤ $1x10^{-10}$ para BLASTN y de e ≤ $1x10^{-3}$ para BLASTX. Los contigs con hits a taxa no virales (procariota y eucariota) se descartaron. Los contigs con hits a secuencias virales y los contigs sin hits a las secuencias de las bases de datos anteriores se conservaron para análisis posteriores.

2.4.5. Reensamble de contigs preliminares

Los contigs recuperados del análisis anterior se sometieron a un reensamble usando el programa Cap3 (Huang y Madan, 1999), especificando un mínimo de empalme de 65 sitios y una identidad de ≥ 98% en la región del empalme entre secuencias. Los contigs obtenidos se denominaron “contigs reensamblados” y a partir de este punto las secuencias de las tres bibliotecas se combinaron en un solo archivo, etiquetando los contigs previamente para conocer su biblioteca de origen.

2.4.6. Detección de ORFs y predicción de secuencias proteicas

Se realizó una búsqueda de marcos de lectura abierta u ORFs (por sus siglas en inglés) en las secuencias de los contigs reensamblados de longitud ≥500 nt, utilizando el programa Prodigal v. 2.6.3 (http://compbio.ornl.gov/prodigal/). En la búsqueda se utilizó la opción meta, que permite detectar ORFs completos e incompletos (sin codón de paro) de cada contig. Así mismo, con Prodigal se recuperó un archivo fasta con la secuencia de aminoácidos predicha para cada ORF. Solo las secuencias ≥ 100 aa de longitud fueron consideradas en análisis posteriores.

2.4.7 Búsqueda de secuencias virales a partir de proteínas predichas

Las secuencias de proteínas predichas recuperadas del paso anterior se sometieron a un análisis de similitud con el programa BLASTP (especificando el reporte de aquellas secuencias cuyos hits tuvieran un valor de e ≤ $2x10^{-3}$) para detectar homología a las secuencias virales de la base de datos de proteínas de referencia RefSeq viral v. 99 (en lo posterior referida únicamente como RefSeq) del NCBI (http://www.ncbi.nlm.nih.gov/refseq/). De forma paralela, las proteínas predichas se sometieron a una búsqueda de dominios virales conservados con el software RPS-BLAST (especificando el reporte de secuencias cuyos hits presentaran un valor de e ≤ $2x10^{-2}$) usando la base de datos CDD v. 3.18 del NCBI (https://www.ncbi.nlm.nih.gov/cdd/).

La anotación de los ORFs que obtuvieron hit con alguna de las bases de datos anteriores se complementó con análisis de posición específica, utilizando el programa Hmmsearch de la paquetería HMMER 3 v. 3.2.1

(Eddy 2011; http://hmmer.org/), la base de datos de grupos de genes virales ortólogos (VOGs) VOGDB v. 83 (http://vogdb.org/) y la base de datos de GyDB v 2.0 de retroelementos y elementos móviles (Llorens et al., 2011). Para todos los análisis de HMMER se especificó el reporte de secuencias cuyos hits presentaran un valor de e ≤ 0.01, tanto para el valor de la secuencia completa como para el mejor dominio.

2.5. Inspección manual de contigs y recuperación de secuencias virales putativas

Los contigs finales obtenidos de los análisis anteriores se inspeccionaron manualmente para recuperar aquellas secuencias cuyos hits con las diferentes bases de datos utilizadas indicaran una secuencia viral putativa. A partir de este punto, las secuencias *queries* con hit a las secuencias virales de RNA y los *queries* con hit a secuencias de virus de dsDNA se analizaron separadamente (a menos que se especifique lo contrario).

Las secuencias fueron inspeccionadas y seleccionadas con los siguientes criterios:

- Tamaño de las secuencias *queries* (y sus respectivos contigs), con especial énfasis en aquellos cuyo tamaño se asemeje al de la secuencia de referencia con la que obtuvieron hit.

- Valores del alineamiento para los hits resultantes de los análisis de BLASTP, RPS-BLAST y HMMER, considerando los valores de significancia mencionados anteriormente. En los resultados de similitud de BLAST se hizo especial énfasis en los hits con **cobertura del *query*** (qcov) ≥90%, sin importar el porcentaje de identidad (id). Respecto a los resultados de HMMER, se hizo especial énfasis en aquellos hits con valores de score ≥ 30, y se inspeccionó la relación de este último valor con su respectivo *bias* (de acuerdo a las recomendaciones de Eddy, 2011), además de la amplitud del alineamiento (buscando que se cubriera la mayor parte del *query* y/o del *subject*) y su respectiva probabilidad posterior.

- Linaje taxonómico resuelto para las secuencias con hit a la base de datos RefSeq. Se enfatizó la búsqueda de las secuencias con hit a los miembros del grupo *Riboviria,* así como a diferentes familias de virus dsDNA, con especial énfasis en las familias *Poxviridae*, *Baculoviridae*, *Iridoviridae*, *Ascoviridae* y el orden *Herpesvirales*. Se inspeccionaron los diferentes *queries* que obtuvieron hit con secuencias asignadas a un mismo txid. En el caso de los *queries* con hit a las secuencias de la

base de datos VOGDB, se analizó la afiliación taxonómica de VOGs *subject* correspondientes a "ssRNA", "dsRNA", o bien "dsDNA; no RNA stage" (de las mismas familias de virus de dsDNA consideradas para los hits con RefSeq), filtrando y seleccionando de acuerdo a los valores del alineamiento y el tamaño de las secuencias *queries* que presentaron hit.

- Secuencias con hits a dominios conservados de proteínas de virus de RNA o DNA.

- Las secuencias que tuvieron hit con alguno de los pHMMs de la base de datos GyDB no fueron considerados.

Para facilitar la selección de secuencias en función de la asignación taxonómica tentativa, se generaron gráficos krona con la paquetería Krona Tools (Ondov et al., 2011) para todos aquellas secuencias de proteína con hit a las secuencias de la base de datos RefSeq (tanto de virus de RNA como de dsDNA). Además, para los contigs correspondientes a las proteínas predichas con hit a las secuencias de los grupos taxonómicos de interés, se construyeron histogramas con la paquetería ggplot (Gregory et al., 2016) en la plataforma R v. 3.0.1 (R Core Team, 2018) para representar el número de contigs de acuerdo a rangos de tamaño.

Con la finalidad de descartar posibles falsos positivos, las proteínas *queries* seleccionadas manualmente con base en sus hits a secuencias de virus de RNA de las bases de datos de RefSeq y VOGDB se sometieron posteriormente a un análisis de BLASTP con la base de datos NR. Si los valores de id, qcov y *e* de cada hit con las bases de datos NR eran mejores para un hit con secuencia no viral que un hit con secuencia viral, el *query* se descartaba.

En la búsqueda de secuencias relacionadas a los taxa virales de dsDNA de interés, se hizo especial énfasis en inspeccionar los hits con la lista de VOGs "exclusivos" o de genes marcadores de virus de dsDNA propuesta por Roux et al. (2019), a partir de la base de datos VOGDB v. 83. Por otra parte, las secuencias *queries* que inicialmente obtuvieron hit con secuencias de virus de dsDNA de las bases de datos especializadas (i.e., RefSeq y VOGDB), se sometieron a un análisis de BLASTP con la base de datos NR antes de la selección manual. Se construyeron gráficos de krona para las secuencias con hit a la base de datos NR y VOGDB/RefSeq para facilitar la clasificación de sus respectivos contigs. Se inspeccionaron considerando los valores de id, qcov y *e* de cada hit con las bases de datos. Si los valores de alineamiento eran mejores (i.e., mayor qcov, id o valor de e) para un hit con secuencia no viral que un hit con secuencia viral, el *query* se descartaba.

2.6 Análisis de cobertura de los contigs seleccionados

Los reads filtrados por calidad se mapearon a los contigs seleccionados para determinar el nivel de cobertura de las secuencias en la muestra. En este análisis se consideraron únicamente las secuencias de virus de RNA putativos seleccionadas. El mapeo de reads a los contigs de referencia se realizó con el programa BWA-MEM v. 0.7.17 (Li y Durbin, 2009). El filtrado por calidad del mapeo, así como el indexado y conteo de los reads mapeados a los contigs seleccionados se realizó con la paquetería Samtools v. 1.7. Adicionalmente, se calculó la cobertura absoluta y la cobertura normalizada con el programa pileup de la paquetería BBTools y se representaron, respectivamente, en histogramas y heatmaps generados con el programa Tidyverse v. 1.3.0 (Wickham et al., 2019) utilizando la plataforma de R v. 3.6.

Con los resultados del mapeo se determinó si un contig se encontraba representado en cada una de las tres bibliotecas. Un contig se considera representado si hay una cobertura mínima de 1X a lo largo de una porción continua de su secuencia ≥70%. En relación con esto, la biblioteca a partir de la cual se ensamblaron los contigs seleccionados se denomina “biblioteca de origen”, mientras que las otras bibliotecas, donde puedan encontrarse representados, se denominan “secundarias”.

2.7. Análisis filogenéticos de las secuencias virales putativas seleccionadas

Todas las reconstrucciones filogenéticas se realizaron a nivel de aminoácidos, con las secuencias completas de proteínas hipotéticas de los ORFs de contigs seleccionados donde se detectaron dominios de RdRP o cápside de virus de RNA. Las secuencias más cercanamente relacionadas a dichas proteínas hipotéticas fueron obtenidas mediante una búsqueda de BLASTP en la base de datos NR del NCBI, seleccionando aquellas que presentaron una qcov ≥ 80%. En todos los casos, se consideró solamente la porción de la secuencia de los *subjects* que cubrió el alineamiento local con el *query*. El alineamiento múltiple de secuencias (MSA) se llevó a cabo con el programa MAFFT v 7.429 (Katoh et al., 2002) del servicio web del EMBL-EBI (https://www.ebi.ac.uk/Tools/msa/mafft/) sin modificar los parámetros por defecto. Los alineamientos se editaron con el programa TrimAl v. 1.3. (Capella-Gutiérrez et al., 2009) en la plataforma Phylemon 2 (Sánchez et al., 2011; http://phylemon2.bioinfo.cipf.es/), especificando un porcentaje de gaps por sitio de hasta 10%. Los alineamientos editados resultantes fueron usados para la generación de árboles filogenéticos basados en máxima verosimilitud (ML) con el software RAxML v. 8.2 (Stamatakis, 2014), disponible en la plataforma CIPRES Science gateway v. 3.3 (Miller et al., 2012; http://www.phylo.org/). En

las reconstrucciones filogenéticas con RAxML se especificó un análisis completo (análisis rápido de bootstrap seguido de una búsqueda del árbol con el mejor puntaje de ML), con una selección automática del modelo de sustitución y el criterio de convergencia de bootstrap (BS). En la reconstrucción de árboles se estableció un grupo externo en cada caso. Los árboles filogenéticos fueron visualizados y editados con el programa iTOL v. 5.6.1 (Letunic & Bork, 2007). Además, para aquellos ORFs que representaron la secuencia completa o casi completa de una poliproteína estructural o no estructural de virus de RNA, se construyeron matrices pareadas de porcentaje de identidad, representándolas mediante heatmaps con un script de R modificado de Gilbert et al. 2019, ejecutado en una plataforma de R v. 3.5.

2.8. Diseño de primers para confirmación experimental de secuencias virales

Con base en los análisis de similitud y de organización genómica, se diseñaron nueve pares de primers con los programas Primer3 (Untergasser et al., 2012) y primer-BLAST (Ye et al., 2012) para una posterior confirmación experimental (no realizada en el presente trabajo) de ocho de las secuencias virales putativas seleccionadas.

2.9. Análisis de partículas tipo virus (VLPs)

2.9.1. Enriquecimiento de VLPs

El enriquecimiento de partículas virales se realizó siguiendo un protocolo modificado de Feng et al. (2017). Se consideraron los sitios P1, P5 y P16 (ya que fueron las primeras muestras en ser colectadas), empleando un total de 3 g de piojos de cada uno. Las tres muestras se procesaron individualmente. Cada muestra fue macerada con 10 mL de buffer PBS 1X (NaCl 137 mM; KCl 2.7 mM; Na_2HPO_4 10 mM; KH_2PO_4 1.8 mM). Posteriormente se añadieron 5 mL de cloroformo y se agitó vigorosamente. El volumen total se centrifugó a 7600 X g por 20 min. a 4 °C. La fase acuosa del sobrenadante fue recuperada y se le aplicó un segundo lavado con 0.5 volúmenes de cloroformo, centrifugando con las mismas condiciones anteriores. El sobrenadante recuperado se pasó por un filtro Millex®-HV con membrana PVDF y abertura de 0.45 µm (Cork Co., Tullagreen, Ireland). El volumen total de filtrado se sometió a ultracentrifugación a 105000 X g por 3 hrs. a 4 °C en tubos OptiSeal® de 4.7 ml (*part number* 361621; Beckman Coulter Inc., Brea, CA, EUA)

usando un rotor TLA-110 de ángulo fijo (Beckman Coulter Inc., Brea, CA, EUA). La pastilla obtenida se cubrió con ~1.5 mL de buffer TES 1X (Tris-HCl 10 mM, pH=7.5; EDTA 2 mM; NaCl 15 mM) y se almacenó el tubo en hielo por 48 horas. Posteriormente, se resuspendió la pastilla pasando el material a través de una aguja de 21G (DL MÉDICA, Edo. de Méx., México) conectada a una jeringa, evitando la formación de burbujas. Después, se transfirió el resuspendido a un tubo nuevo donde se llevó a 7 mL con buffer TES 1X. El volumen total se homogenizó usando la jeringa con aguja de forma similar al proceso anterior. Se centrifugó a 7000 X g por 5 min a 4 °C y el sobrenadante obtenido se pasó por un filtro Millex®GP de membrana PES con abertura de 0.2 µm (Cork Co., Tullagreen, Irlanda). A continuación, las muestras se sometieron a ultracentrifugación a 128000 X g / 4 hrs. a 4 °C utilizando un colchón de sacarosa al 40%, en los mismos tubos OptiSeal correspondientes a cada muestra. El sobrenadante resultante se descartó y la pastilla se resuspendió con buffer TES 1X (~500 µL). Las suspensiones se mantuvieron en refrigeración a 4 °C por un periodo no mayor a 1 semana, hasta su preparación para la observación al microscopio electrónico de transmisión (TEM). Todas las ultracentrifugaciones se realizaron en una ultracentrífuga OPTIMA®MAX-XP (Beckman Coulter Inc., Brea, CA, EUA).

2.9.2. Fijación de muestras y observación al TEM

Se tomó una alícuota de 100 µL de cada uno de los volúmenes de concentrados de VLPs obtenidos (previamente homogenizados por pipeteo) y se procesaron separadamente para la fijación. El volumen de cada alícuota se mezcló con 40 µL de glutaraldehido al 50% y 860 µL de buffer TES 1X (la concentración final de glutaraldehido fue 2%). La mezcla se dejó fijando por 1 hr. a 4°C. Las muestras se concentraron nuevamente por ultracentrifugación en tubo OptiSeal a 100000 X g por 1 hr., se retiró el fijador y se resuspendieron en TES 1X. Las muestras fijadas fueron teñidas con Acetato de Uranilo y observadas posteriormente en el microscopoio electrónico de transmisión (TEM, por sus siglas en inglés) (Hitachi High-Tech Co., Tokio, Japón) del Laboratorio Nacional de Microscopía Avanzada de CICESE.

Capítulo 3. Resultados

3.1. Identificación molecular de *P. ficus*

3.1.1. Extracción de DNA total

El DNA total extraído correspondiente a cada sitio de colecta se muestra en la **tabla 2**. En la mayoría de los casos la concentración fue mayor a 1000 ng/µL, con una relación A260/280 ≥ 1.8. La presencia de DNA degradado fue conspicua al observarse un barrido en gel de agarosa al 0.8% (no mostrado) a lo largo del carril para la mayoría de las muestras. Sin embargo, también se pudo observar una banda definida tenue en la parte superior del carril para la mayoría de las muestras que evidenció la presencia de DNA íntegro. Por otra parte, la relación de absorbancia A260/280 fue ≥ 1.9 en todos los casos, por lo que el material se consideró apto para realizar la PCR multiplex y continuar con la identificación molecular.

Tabla 2. Datos de concentración y relación de absorbancia 260/280 de los DNAs obtenidos a partir de piojos colectados en 13 de 14 viñedos muestreados en el municipio de Ensenada.
*El sitio CC se empleó únicamente como referencia para la identificación molecular de la(s) especie(s) de piojo harinoso y no fue considerado para la construcción de bibliotecas de secuenciación.

Sitio de muestreo	Concentración (ng/µL)	A260/280
P1	1599.8	1.9
P2	1211.2	2.04
P3	1143	1.96
P4	984.7	1.97
P5	1851.4	1.98
P7	1191.5	2.02
P8	232.6	2
P9	803	1.95
P10	1038.6	2.02
P12	1600.1	1.96
P16	1478.2	1.98
P17	944.6	1.94
P18	1226	2.03
CC réplica 1 (CC1)	219.9	2.05
CC réplica 2 (CC2)	1302.5	2.01

3.1.2. PCR multiplex

La mayoría de las muestras presentó un producto de PCR del tamaño esperado (450 bp) para *P. ficus* (**figura 4**). Solo una de las réplicas del sitio CC (CC2) presentó un producto de PCR evidente de ~600 pb, que corresponde con el tamaño de producto esperado para *Ps. longispinus*. Es importante mencionar que algunas muestras (P9, P10 y P17) presentaron bandas tenues de ~400pb y que, de acuerdo con Daane et al. (2011), coinciden con el tamaño esperado para la especie *Ps. maritimus*.

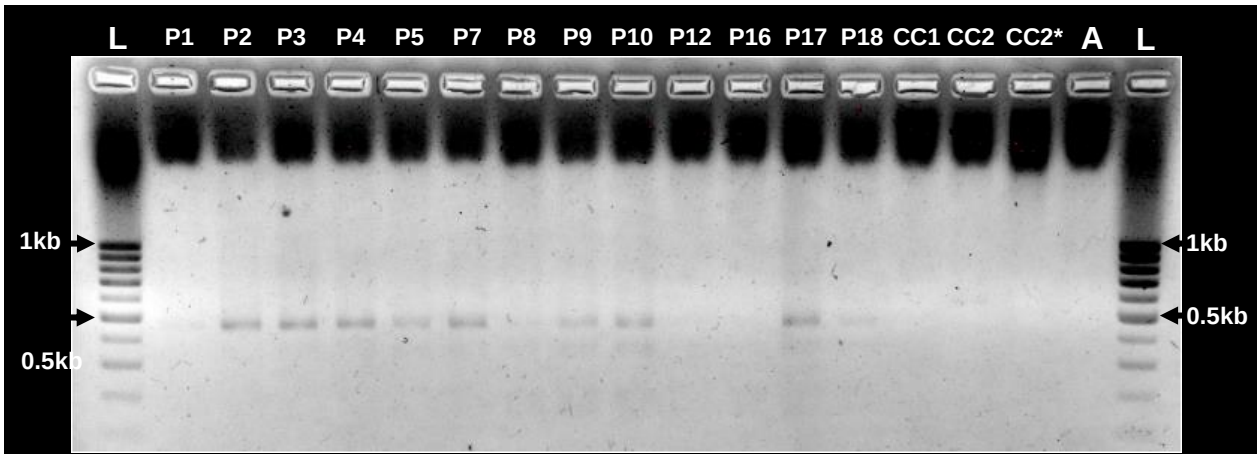

Figura 4. Productos de la PCR multiplex para especies de piojo harinoso. L: Marcador de peso molecular de 100 pb; A: control de agua y reactivos sin DNA; CC1, CC2: réplicas del sitio CC. CC2*: CC2 no diluído. Gel de garosa al 2% (45min a 90V).

Los PCR uniplex confirmaron la presencia de dos amplicones de tamaño distinto, siendo el fragmento obtenido con los primers de *P. ficus* de ~450 pb, mientras que el producto de PCR tanto con los primers de *Ps. maritimus* como con los de *P. citri* fue de ~300 pb.

3.1.3. Secuenciación de productos de PCR uniplex

Los análisis de similitud (megablast y blastn) y alineamiento de los amplicones secuenciados mostraron que en todos los casos los fragmentos obtenidos en los PCRs uniplex corresponden a secuencias parciales del gen *co1* de *P. ficus*, independientemente del tamaño de dichos fragmentos y a pesar de haber usado primers supuestamente específicos para las especies *Ps. maritimus* y *P. citri*. Esto indica que los primers de las tres especies amplificaron fragmentos de distintos tamaños del DNA de *P. ficus* en las muestras analizadas. Por lo tanto, este resultado confirmó la identidad y unicidad de *P. ficus* en las muestras usadas

(sin embargo, es necesario resaltar que en la identificación molecular se utilizó sólo una porción de los piojos colectados de cada sitio).

3.2. Construcción de bibliotecas a partir de RNA total sin rRNA y secuenciación

3.2.1. Extracción de RNA total

El RNA total de muestras de piojo de los 14 sitios de muestreo considerados en la preparación ulterior de bibliotecas se muestra en la **tabla 3**, junto con los valores de relación de absorbancia 260/280. Casi todas las muestras presentaron concentraciones de RNA > 1000 ng/µL, con valores de A260/280 ≥ 1.8.

Tabla 3. Datos de concentración y relación 260/280 de RNA extraído para cada uno de los 14 sitios considerados en la construcción de bibliotecas.

Sitio de muestreo	Concentración (ng/µL)	A260/280
P1	2100	1.9
P2	469.5	1.79
P3	2144.3	2.13
P4	1608.8	1.95
P5	1264.9	1.87
P7	1662.7	1.90
P8	1683.8	1.93
P9	1702.6	1.91
P10	2179	1.75
P12	1681.1	1.9
P13	599.2	1.82
P16	2013.1	1.61
P17	1935.1	1.9
P18	1909.5	1.81

3.2.2. Muestras de secuenciación y limpieza de RNA

Se construyeron tres diferentes agrupaciones de RNA llamadas muestras finales (MF), que contienen el RNA extraído del material biológico colectado en los 14 sitios de muestreo (**tabla 4**). Después de la consolidación del RNA en MFs y del proceso de estandarización del protocolo de limpieza de RNA con columnas miniprep y tratamiento con Dnasa, se logró obtener concentración suficiente de RNA final de cada muestra (>100 ng/µL) para comenzar la generación de bibliotecas.

Tabla 4. Concentraciones [ng/µL] y relaciones de absorbancia A260/A280 y A260/A230 de muestras finales de secuenciación de RNA después de la limpieza y tratamiento con DNasa. ND: no determinado.
[a]: determinado con Nanodrop; [b]: determinado con Qubit

Muestras de RNA	[ng/µL][a]	[ng/µL][b]	A260/280[a]	A260/230[a]
MF1	136	136	1.9	ND
MF2	165.9	97	2.12	1.29
MF3	180.2	146	2.06	2.08

Adicionalmente, se tomaron las relaciones de absorbancia A260/A280 y A260/A230, cuyos valores se usan como indicadores de pureza del RNA, i.e., para verificar ausencia de cantidades sustanciales de reactivos orgánicos utilizados durante la extracción/limpieza del RNA (Thermo Scientific T042-TECHNICAL BULLETIN). Puede notarse que las relaciones de absorbancia A260/280 de cada una de las muestras de RNA presenta valores ≥1.9, lo que indica un pico de absorbancia a 260 nm, con poca o nula absorbancia a 280 nm; estos valores se observan típicamente en muestras de RNA libres de proteínas o fenoles. Por otra parte, se obtuvo un valor de A260/A230 > 2 para la muestra 3, lo cual no ocurre con el valor correspondiente a la muestra 2. Una posible explicación para un valor de A260/A230 < 2 (i.e., debajo de lo aceptable) es que durante la extracción de RNA pudieron arrastrarse contaminantes residuales orgánicos como fenol, cloroformo o TRIzol, especialmente durante la separación de fases orgánica y acuosa. O bien, pudieron ser residuos provenientes de los buffers usados en la limpieza por columna, como el *RNA prep buffer* que contiene cloruro de guanidina, el cual tiene una absorbancia elevada a 230 nm, o el TRIzol LS, que presenta picos de absorbancia a 270 nm y a 230 nm (*RNA prep buffer* Safety data sheet; Thermo Scientific T042-TECHNICAL BULLETIN).

Los RNAs de todas las muestras finales de secuenciación (MF) presentaron bandas íntegras de rRNA (28S y 18S) en un gel de agarosa al 2%, por lo que se consideraron aptos para comenzar con la remoción de los rRNA (**figura 5**).

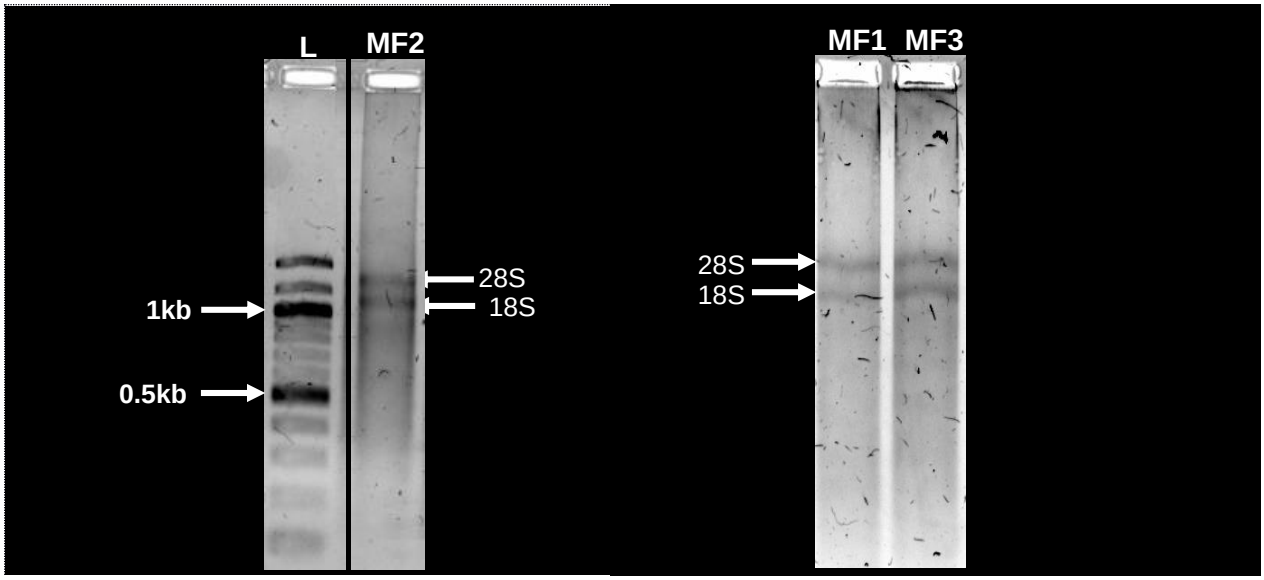

Figura 5. RNA total de las muestras finales para la construcción de bibliotecas. L: marcador de peso molecular de 100 pb de dsDNA. Geles de agarosa al 2% (80V por 60 min). En el gel de MF2, los carriles no fueron contiguos.

3.2.3. Generación de bibliotecas de secuenciación

Se generaron 3 bibliotecas de secuenciación a partir de los 3 grupos de RNA especificados anteriormente (MF1-3). La cantidad total de RNA inicial para la biblioteca PV2 fue de 770 ng, mientras las bibliotecas PV1 y PV3 se iniciaron con una cantidad de 1000 ng totales cada una. Las concentraciones de RNA inicial, cDNA, dsDNA y la concentración nanomolar correspondientes a las etapas de preparación de cada biblioteca se muestran en la **tabla 5**. Luego del PCR de enriquecimiento y la limpieza con perlas magénticas, se visualizó la población de fragmentos de dsDNA de cada biblioteca en un gel de agarosa al 4% (**figura 6**). El tamaño de la población de fragmentos de dsDNA observada en dicha electroferesis coincidió con las mediciones de tamaño realizadas en BioAnalyzer, siendo el tamaño promedio obtenido de la población de fragmentos de dsDNA de las bibliotecas PV1, PV2 y PV3 de 376 pb, 397 pb y 390 pb, respectivamente (**figura 7**).

Tabla 5. Cuantificación por espectrofotómetro Nanodrop® 2000, fluorómetro Qubit® y/o Bioanalyzer® de los productos obtenidos durante la construcción de las bibliotecas.
*La cuantificación del ds cDNA de la biblioteca 2 se realizó antes/después de la segunda limpieza con perlas AMPURE XP.

Análisis	PV1	PV2	PV3
Cuantificación por Nanodrop de RNA total [ng/µL]	136	165.9	180
Cuantificación por Qubit® de RNA total [ng/µL]	136	97	146
Cuantificación por Qubit® de cDNA [ng/µL]	5.2	4.74	4.58
Cuantificación por Nanodrop® de dsDNA [ng/µL]	19.5	*20/13	15.2
Cuantificación por Qubit® de dsDNA [ng/µL]	11.3	11.7	11.7
Concentración nanomolar estimada [nmol/L]	43.12	47.14	45.45
Tamaño promedio de fragmentos de dsDNA	397	376	390

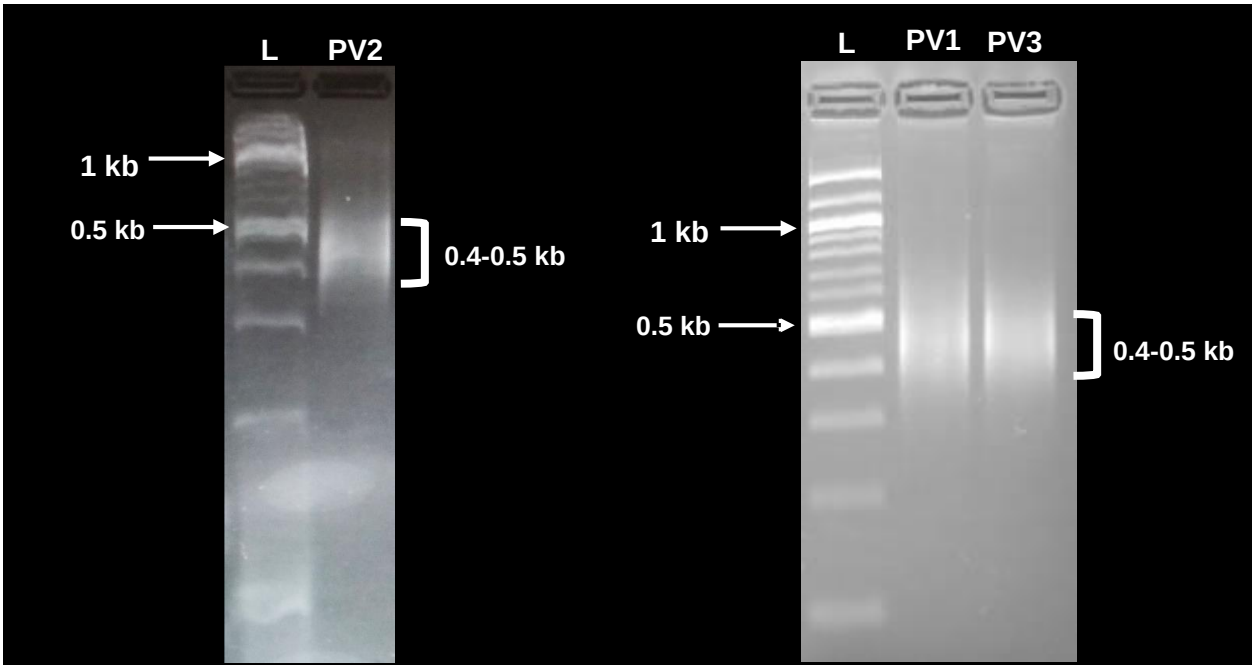

Figura 6. Fragmentos de dsDNA final obtenido para cada biblioteca. En corchetes se señala la región con mayor densidad de fragmentos observados en el barrido. Para todos los geles, L: marcador de peso molecular de 100 pb. Se enfatizan las bandas de 1 kb y 0.5 kb. Gel de agarosa al 4% (75-90 min a 90V).

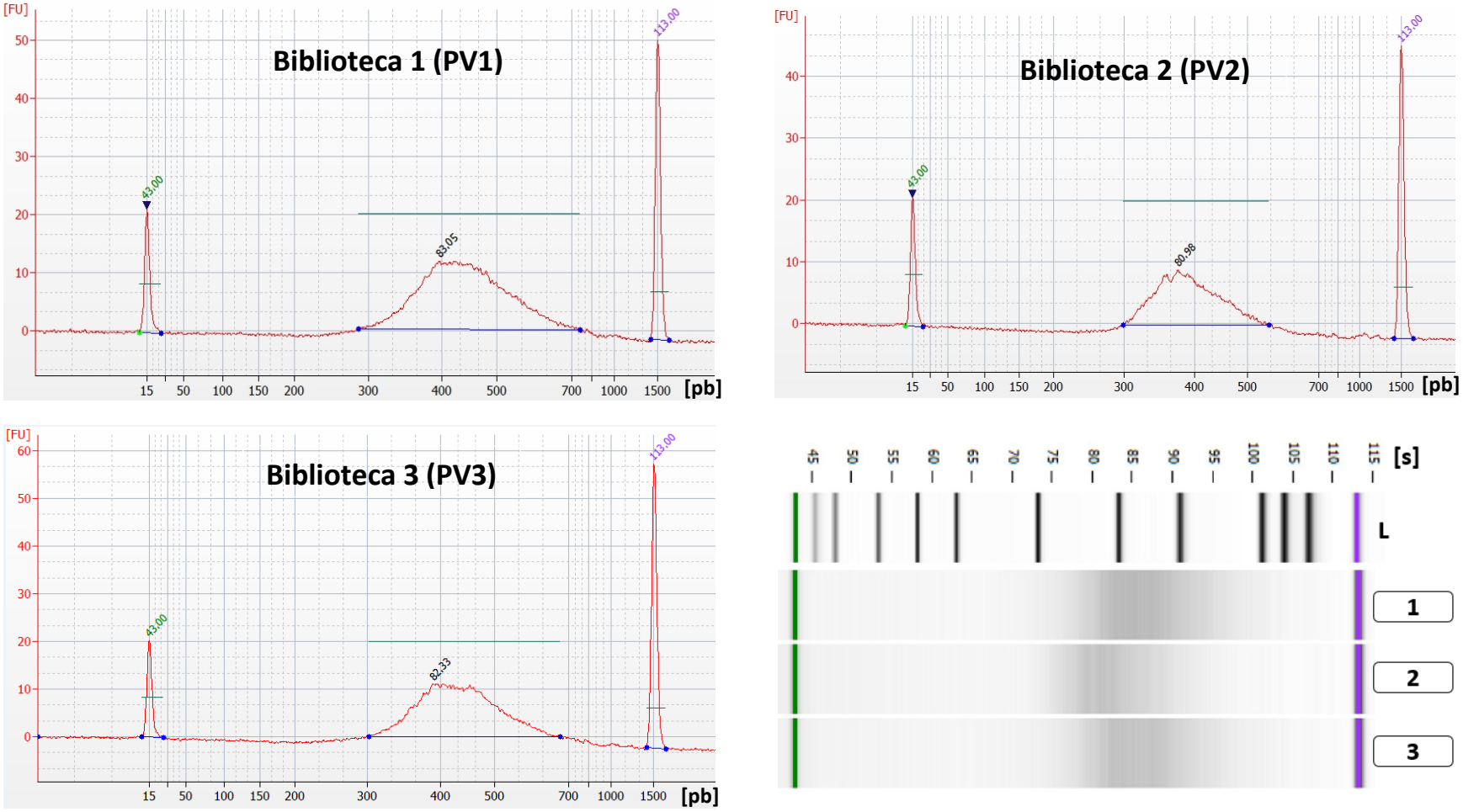

Figura 7. Análisis de la población de fragmentos de dsDNA de cada biblioteca en unidades de fluorescencia (FU) vs tamaño en pares de bases. Para cada biblioteca se muestran los marcadores de peso molecular de referencia de 15 pb y 1500 pb. La amplitud de la curva es indicativa del rango de tamaños para cada población. La intensidad de los picos es proporcional a la concentración del tamaño indicado. Abajo a la derecha: la comparación de los fragmentos en un gráfico de FU vs tiempo (segundos) indica el tamaño promedio de los fragmentos de dsDNA en un momento de tiempo dado (la migración de las bandas es de derecha a izquierda).

3.3. Análisis bioinformático

3.3.1. Control de calidad para reads crudos

El análisis de calidad realizado previo al corte de calidad indicó que los reads de las tres bibliotecas presentaban valores promedio de Phred score por arriba de 30, sin embargo, la PV2 presentó el valor medio de Phred score más bajo de todos los reads crudos (**Anexo B**). Antes del corte de calidad, se observó que la biblioteca con el mayor número de reads (1P y 2P) fue la PV1, mientras que la PV2 fue la que presentó el menor número (**tabla 6**). Así mismo, se observó que la PV1 y la PV3 mostraron el menor contenido de secuencias sobre-representadas antes del corte de calidad, siendo la PV2 la que mostró el mayor porcentaje. Posterior al corte de calidad, la biblioteca PV1 fue la que conservo el mayor número de reads, seguida por la biblioteca PV3 y finalmente la PV2 (lo cual era esperado dados los valores de Phred score iniciales). Por otra parte, la longitud de los reads filtrados por calidad de la PV2 también es menor comparada con la longitud de los reads filtrados de las otras dos bibliotecas. Además, la biblioteca PV2 continuó manteniendo un porcentaje relativamente alto de reads sobre-representados respecto a las otras bibliotecas.

Tabla 6. Valores obtenidos en el análisis de calidad, previos (PRE) y posteriores (POST), al corte de calidad de reads crudos.

Archivos de reads crudos	PRE			POST		
	Longitud (pre)	Millones de secuencias (pre)	% de secuencias sobre-representadas	Longitud (post)	Millones de secuencias (post)	% de secuencias sobre-representadas
PV1_1P	150 bp	24.8	2.13	145	20.3	0.11
PV1_2P	150 bp	24.8	1.97	145	20.3	0.11
PV2_1P	150 bp	16.9	11.89	136	10.5	3.07
PV2_2P	150 bp	16.9	10.78	141	10.5	2.94
PV3_1P	150 bp	22.9	1.54	144	17.9	0.11
PV3_2P	150 bp	22.9	1.43	144	17.9	0.13

3.3.2. Filtrado de reads no virales

La menor proporción de reads sustraídos por el mapeo a secuencias de referencia se observó en la biblioteca PV2, donde se removieron aprox. el 25% de los reads, mientras que para la biblioteca PV1 y PV3

se filtraron aprox. el 32% y el 45% de los reads, respectivamente (**tabla 7**). Así, la biblioteca PV2 conservó una mayor proporción de reads que no corresponden a alguna de las secuencias de referencia utilizadas.

Por otra parte, debido a que el mapeo se realizó utilizando un solo índice para todos los genomas/secuencias de referencia, no fue posible determinar la porción de reads que se mapearon a cada una de dichas referencias.

Tabla 7. Reads 1P y 2P de alta calidad de cada biblioteca, antes (pre-) y después (post-) del mapeo a secuencias no virales de referencia. Los números de cada casilla indican millones de reads.

Archivos de reads	PV1_1P	PV1_2P	PV2_1P	PV2_2P	PV3_1P	PV3_2P
Procesados (pre-filtrado)	20.3	20.3	10.5	10.5	17.9	17.9
Usados para ensamble (post-filtrado)	13.7	13.7	7.8	7.8	9.7	9.7

3.3.3. Ensamble *de novo*

Las métricas del ensamblado de reads filtrados por calidad indicaron que la biblioteca PV1 fue para la que se obtuvo el mayor número de contigs preliminares, seguida por PV3 y PV2 (**tabla 8**).

Tabla 8. Métricas del ensamble realizado para los reads de las tres bibliotecas. Se muestra además el número de contigs ordenados por longitud.

Métricas	PV1	PV2	PV3
N de secuencias	181454	85876	103727
Más larga	13259	11230	11626
N/L50	50789/610	25797/553	28803/600
N/L90	147392/342	70674/331	84673/338
Número de secuencias ordenadas según su longitud			
Rangos de tamaño	**PV1**	**PV2**	**PV3**
Total	181454	85876	103727
100-249	181454	85876	103727
250-499	180506	85021	103149
500-999	77214	31856	41670
1000-2499	16993	6463	9808
2500-4999	2032	441	1104
5000	208	19	98
10000	7	1	5

En todas las bibliotecas el valor de L50 fue de 550-610 nt. Este último valor es central, ya que indica que la mitad de las bases consideradas en el ensamble están representadas en contigs de igual o mayor longitud a este valor. En la distribución de tamaños de los contigs ordenados por longitud se observa que en cada rango, la biblioteca PV1 es aquella donde se rescata la mayor cantidad de contigs. Al igual que con la tendencia observada en los reads filtrados por calidad, la biblioteca PV2 es la que tiene el menor número de contigs en cada corte de longitud.

3.3.4. Filtrado de contigs preliminares no virales

Los resultados de los análisis de similitud de BLASTN y BLASTX con los contigs preliminares indicaron que una porción del 60-90% no obtuvo hit con alguna de las secuencias de las bases de datos utlizadas (por lo que se denominan contigs indeterminados), mientras que sólo una pequeña porción (~1%) obtuvo hit con secuencias virales y el resto presentó hit con secuencias de organismos celulares. Estos últimos contigs fueron descartados para análisis subsecuentes (**tabla 9**).

Es importante mencionar que el análisis de BLASTX no pudo terminarse para ninguna de las tres bibliotecas debido a limitaciones del tiempo de ejecución, por lo que cabía esperar que en el set de contigs preliminares sometidos a reensamble (sección siguiente) se encontraran todavía secuencias de origen celular.

Tabla 9. Resultados de los análisis de similitud con contigs preliminares. Se muestra el número de secuencias sometidas y el número de secuencias con hit a las secuencias de las bases de datos no redundantes. El total de secuencias restantes corresponde a las que se emplearon en análisis posteriores.

	Contigs preliminares			
	Contigs sometidos (≥ 500 nt)	**Número de hits no virales**		**Restantes (%)**
		BLASTN	**BLASTX**	
PV1	77214	3759	4710	69528 (90%)
PV2	31856	3671	10350	20059 (63%)
PV3	41670	3808	6989	32290 (77%)

3.3.5. Reensamble de contigs preliminares

El número total de contigs observado después del reensamble fue menor que el total inicial, mientras que el conteo de los contigs por rango de tamaño indicó que después del reensamble hubo una reducción de los contigs < 1000 nt y un aumento de los contigs ≥ 1000 nt (**tabla 10**).

Tabla 10. Comparación de contigs antes (pre-) y después (post-) del reensamble con Cap3. Se ordenan los contigs por rangos de tamaño.

Rango de tamaño	PV1		PV2		PV3	
	Pre	Post	Pre	Post	Pre	Post
500-999	69528	68119	20059	19808	32290	31727
1000-2499	13236	13429	2911	2976	6057	6151
2500-4999	1200	1391	150	15	492	569
5000	93	132	6	8	40	56
10000	2	3	0	1	1	1

3.3.6. Búsqueda de secuencias virales a partir de proteínas predichas.

Los análisis de similitud para detectar similitud entre las secuencias recuperadas hasta este punto y las secuencias virales reportadas en bases de datos se realizaron a nivel de aminoácidos. De la predicción de ORFs en los contigs reensamblados y la obtención de sus respectivos productos proteícos putativos se obtuvieron un total de 41837 secuencias de proteínas de tamaño ≥ 100 aa, que corresponden a 40411 contigs, lo que indica que la mayoría de los contigs presentan un solo ORF de interés. Estas secuencias de proteínas fueron sometidas a los análisis con BLASTP, HMMER y RPS-BLAST, de los cuales se obtuvieron hits con las secuencias de las diferentes bases de datos, de los diferentes grupos taxonómicos de interés, como se muestra en la **tabla 11**.

Como puede observarse en el krona de la **figura 8**, el análisis de BLASTP usando la base de datos RefSeq indica que de las 4408 secuencias de proteínas que obtuvieron hit, una porción >80% de los hits corresponde a grupos taxonómicos de virus de dsDNA. Aquellos contigs que corresponden a proteínas *queries* con hit a grupos como *Caudovirales*, *Phycodnaviridae*, *Ortervirales*, *Mimiviridae,* o virus no clasificados (“unclassified viruses”), fueron descartados puesto que dichos taxa se han caracterizado como virus que infectan predominantemente a bacterias, algas o protistas, o bien, son retrovirus (orden

Ortervirales). Tomando los contigs correspondientes a las proteínas que tuvieron hit con los grupos taxonómicos de interés (ver **tabla 11**), la ditribución de su número en función del tamaño de secuencia indica que, en todos los casos, hay aproximadamente 50 contigs de tamaño ≥ 2000 nt (**figura 8**, B). Este ordenamiento es relevante ya que si bien el tamaño es siempre un criterio de completitud relativo al grupo taxonómico del que se trate, es muy probable que dichos contigs representen fragmentos (pequeños) de genomas virales, ya que los genomas virales de RNA más pequeños reportados son de ~2000 nt y los genomas de dsDNA de los grupos taxonómicos de interés son de una longitud dos ordenes de magnitud mayor a los genomas de RNA. A partir de este punto, las secuencias que presentaron hit con el grupo *Riboviria* se analizaron individualmente.

Tabla 11. Resultados de los análisis de similitud con las proteínas predichas a partir de contigs reensamblados. Se muestra el número de secuencias sometidas y el número de secuencias con hit a las secuencias de las bases de datos. Se muestra además el total de secuencias *queries* con hit a las secuencias de cada grupo taxonómico de interés.

Contigs rensamblados			
Contigs	**Secuencias de proteínas sometidas (≥ 100 aa)**	**Número de hits BLASTP (RefSeq Viral)**	**Número de hits HMMER (VOGDB_83)**
40411	41837	4408	1490

Número de hits por grupo taxonómico de interés		
Taxón	**Número de hits BLASTP (RefSeq Viral)**	**Número de hits HMMER (VOGDB_83)**
Riboviria	335	209
Baculoviridae	406	425
Poxviridae	565	320
Iridoviridae	247	238
Ascoviridae	82	94
Herpesvirales	87	325

Por su parte, las secuencias de proteína *queries* con hit a secuencias de dsDNA de las bases de datos de RefSeq viral y/o HMMER se sometieron a un análisis de BLASTP usando la base de datos NR para descartar posibles secuencias de orgamismos celulares con similitud a secuencias virales de dsDNA.

Se observó que de los 1297 *queries* que presentaron hit a secuencias de virus de dsDNA de la base de datos VOGDB, 878 *queries* (~68%) también presentaron hit a las secuencias de virus de dsDNA de la base de

datos RefSeq viral (**figura 9**, A). Sin embargo, el análisis de BLASTP usando la base de datos NR indicó que 1171 secuencias (krona de la **figura 9**, B) de las 1297 secuencias *queries* iniciales (aprox. un 90%) también presentaron hit, prácticamente todas con secuencias de organismos celulares, incuyendo a 856 de las 878 secuencias *queries* que presentaron hit con secuencias virales de dsDNA de RefSeq y VOGDB; siendo solo 22 secuencias las que presentaron hit con secuencias de estas dos últimas bases de datos pero no con NR (**figura 9**, A). Por otra parte, 315 secuencias (~24.3%) presentaron hit con las secuencias de VOGDB y secuencias de organismos celulares de NR, pero no con la base de datos RefSeq viral y solo 104 secuencias (~8%) presentaron hit únicamente con la base de datos VOGDB.

Por otro lado, respecto a los análisis de similitud de BLASTP con la base de datos de RefSeq, se tuvo que 1243 secuencias *queries* presentaron hit con secuencias de virus de dsDNA, pero no con las secuencias de la base de datos VOGDB (**figura 9**, A). De estas, 1213 presentaron tuvieron hit con la base de datos NR, lo que representa un %97.2 del total de 1243 *queries* iniciales (krona de la **figura 9**, C). A su vez, de esta fracción, solo 4% presentó hit con secuencias virales de la base de datos NR. Así, considerando solo el mejor hit, un total de 74 secuencias (5.9% del total inicial) de las 1243 iniciales obtuvieron como el mejor hit a secuencias de virus de dsDNA de alguna de las dos bases de datos.

En conjunto, los análisis de similitud con BLASTP y HMMER indicaron que de un total de 2540 secuencias *queries* iniciales (con hit a secuencias de virus de dsDNA deRefSeq y/o VOGDB), sólo una pequeña fracción (< 10%) presentó hit exclusivamente con secuencias de virus de dsDNA y no con organismos celulares al compararse con la base de datos NR.

A partir de este resultado, se inspeccionaron los valores de alineamiento para las secuencias que tuvieron hit exclusivamente con secuencias de dsDNA, o hit tanto a secuencias virales (RefSeq/VOGDB) y secuencias de organismos celulares (NR). En este análisis se observó que para más del 90% de las secuencias, los hits con secuencias de organismos celulares presentaron valores de qcov o id mayores a los que presentaban con secuencias virales, por lo que fueron descartadas. Así, sólo un total de ~200 secuencias virales putativas de dsDNA se consideraron en la selección manual.

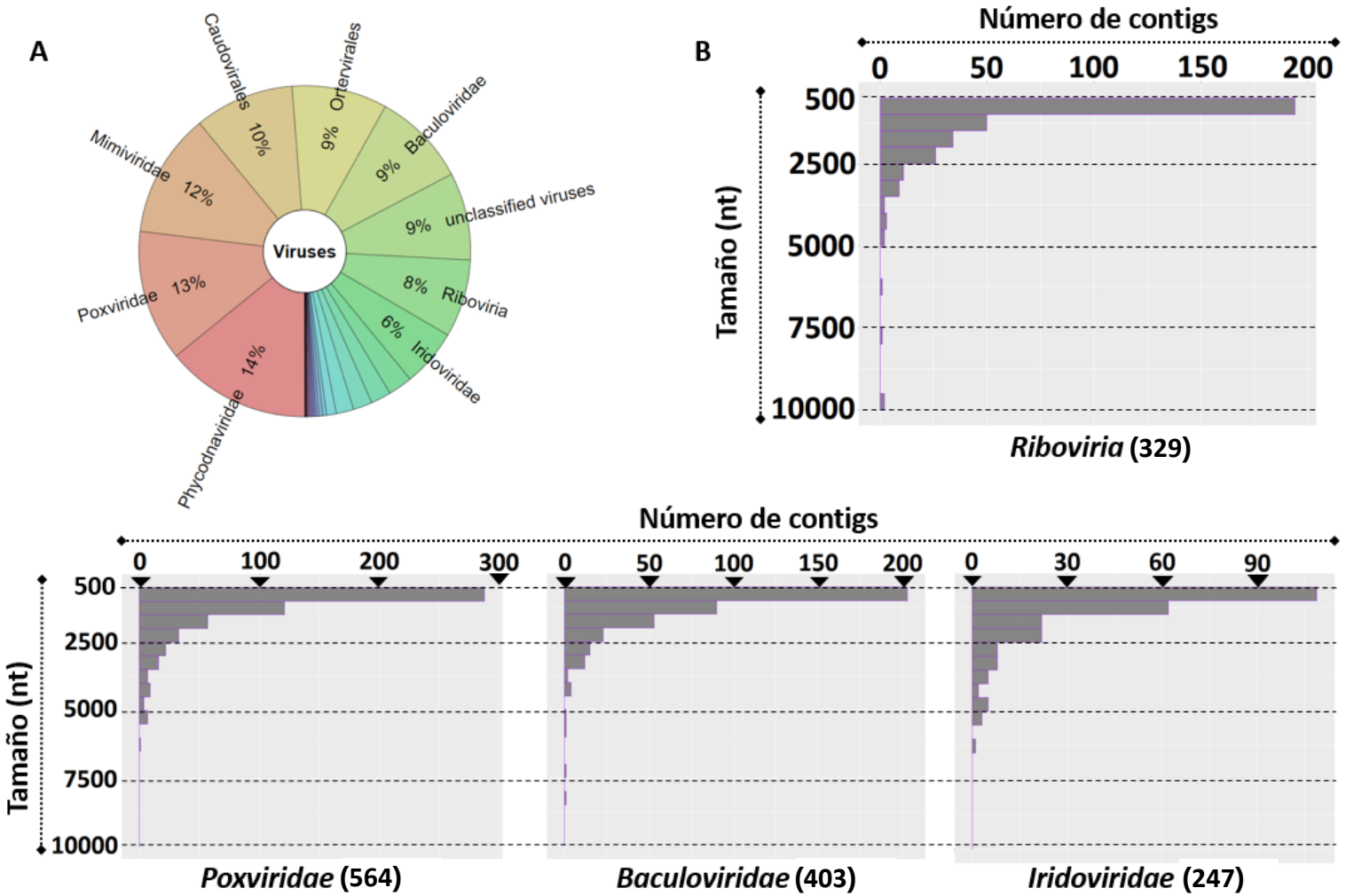

Figura 8. Gráficos krona de asignación taxonómica e histogramas de frecuencia para secuencias con hit a la base de datos RefSeq. (A) El gráfico krona considera las 4408 secuencias de proteínas *queries* con hit a la base de datos de RefSeq, donde se observa que la mayoría de los hits obtenidos correspondieron a virus de dsDNA. (B) Histogramas del número de contigs correspondientes a las proteínas *queries* con hit a secuencias de virus de los grupos taxonómicos de mayor interés, ordenados según su tamaño (se muestran solo los grupos taxonómicos donde se concentra el mayor número de hits). Al pie de cada histograma se indica el grupo taxonómico al que corresponde y en paréntesis el número total de contigs asignados a ese grupo (nótese que el número de contigs es menor o igual al número de proteínas *queries* mostradas en la tabla 4 ya que a cada contig le corresponde una o más proteínas). Cada barra del histograma representa un rango de 500 nt.

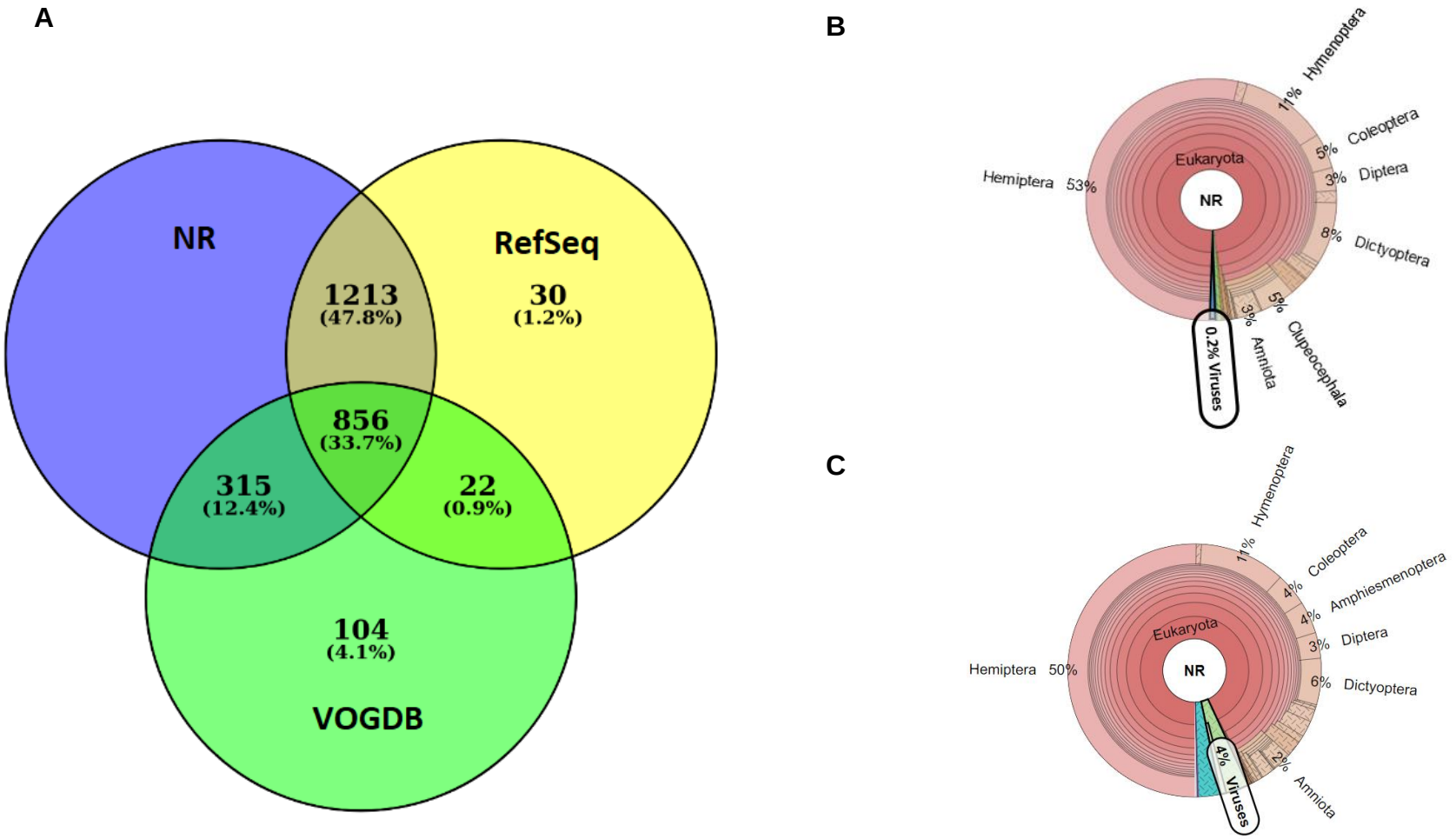

Figura 9. Diagrama de Venn y gráficos krona para las proteínas *queries* con hit a secuencias de dsDNA. (A) El diagrama de Venn muestra el número y porcentaje de proteínas *queries* que tuvieron hits con las secuencias de virus de dsDNA de las bases datos RefSeq y/o VOGDB, y de estas, el número y porcentaje de *queries* que posteriormente tuvieron hit con las secuencias de la base de datos NR (las intersecciones de círculos indican el número y porcentaje de *queries* que tuvieron hit con las secuencias de dos o más bases de datos). (B) Gráfico krona para las 1171 secuencias de proteínas *queries* con hit a las bases de datos VOGDB y NR, que muestra el resultado de la asignación taxonómica de las secuencias tomando como referencia su mejor hit con la base de datos NR. (C) Gráfico krona para las 1213 secuencias de proteínas *queries* con hit a las bases de datos RefSeq y NR, que muestra el resultado de la asignación taxonómica de las secuencias tomando como referencia su mejor hit con la base de datos NR.

3.4. Inspección manual de contigs, recuperación de secuencias virales putativas

Con base en la inspección manual de los resultados obtenidos para las secuencias de proteínas predichas, se seleccionaron 27 contigs que representan secuencias virales putativas de RNA. Las proteínas predichas correspondientes a dichos contigs mostraron homología tentativa con las secuencias de virus de RNA de las bases de datos empleadas, así como la presencia de dominios estructurales y/o no estructurales de virus de RNA (en la mayoría de los casos). Considerando la asignación taxonómica de las secuencias de acuerdo a los mejor(es) hit(s) en los análisis de BLASTP y/o la organización genómica de los contigs correspondientes, se recuperaron dos contigs relacionados al género *Marafivirus* (familia *Tymoviridae*), un contig relacionado con la familia *Dicistroviridae* (orden *Picornavirales*), seis contigs relacionados con la familia *Iflaviridae* (orden *Picornavirales*), cinco contigs relacionados con el orden *Picornavirales,* aunque sin ninguna afiliación clara con alguna de sus familias, ocho contigs relacionados con la familia *Reoviridae*, cuatro contigs relacionados con la familia *Rhabdoviridae* y un contig relacionado con la familia *Tombusviridae*. De este grupo de secuencias, se seleccionaron ocho contigs que presentaban algunas de las características más prominentes o conservadas de sus respectivos grupos y/o una completitud mayor del genoma según su asignación taxonómica, con los cuales se realizó el diseño de primers para una posterior confirmación experimental (no abordada en el presente trabajo; ver **Anexo D**).

La **tabla 12** presenta la información de los 27 contigs seleccionados de acuerdo a su afiliación taxonómica tentativa con los diferentes taxa de virus de RNA, y se muestran los resultados más importantes obtenidos en los análisis de similitud y dominios con las secuencias de proteínas predichas. Se enlista el ID del mejor hit (*subject*) obtenido en los análisis de BLASTP usando las bases de datos de RefSeq y NR, la cobertura del *query* en el alineamiento, la identidad entre secuencias, y el valor *e* de significancia. Se muestra además un listado de los dominios virales (y su valor *e* de significancia) detectados en las secuencias de proteínas predichas de los ORFs de cada contig. La **figura 10** muestra los resultados del mapeo en reads/Kb para las 27 secuencias. Una descripción más detallada de los resultados relevantes de las secuencias mencionadas se presenta en las siguientes secciones. Se integran además los resultados de la cobertura absoluta y relativa, así como los análisis filogenéticos correspondientes.

Por otro lado, cabe destacar que las bases de datos más comprehensivas y con las cuales se pudo tener más certeza del posible origen viral de las secuencias seleccionadas fueron las bases de datos de RefSeq, NR y CDD, del NCBI. Durante la selección manual de secuencias virales putativas de RNA, se lograron rescatar inicialmente secuencias que obtuvieron hit únicamente con las diferentes secuencias de la base de datos VOGDB, sin embargo, no se incluyeron porque i) su longitud era muy corta respecto a lo esperado

tomando en cuenta la afiliación taxonómica de los VOGs (o el tamaño de la secuencia *subject*), ii) la secuencia no presentaba resultados adicionales con otras bases de datos que soportaran este resultado, o iii) al someter dichas secuencias al BLASTP con la base de datos NR (realizado luego de la selección inicial para evitar falsos positivos o sesgos de las búsquedas contra la base de datos RefSeq), las secuencias mostraban hits con secuencias de origen no viral (celular) o con taxa virales que no eran de interés (p. ej. fagos, virus de protistas, etc.). Aunado a lo anterior, también se observó que tomando en cuenta únicamente los resultados de los análisis con VOGDB para una secuencia determinada, en muchos casos los VOGs con los que una secuencia *query* tuvo hit presentaban una afiliación taxonómica que entraba en conflicto (p. ej. VOGs de familias diferentes) con valores de significancia similares (lo cual puede deberse, en algunos casos, al hecho de que son pHMMs construidos con secuencias conservadas incluso entre varias familias de virus). Así, considerando lo anterior, la descripción de los resultados obtenidos para las secuencias recuperadas de virus de RNA se hizo enfocándose principalmente en los resultados con las bases de datos mencionadas del NCBI, que serán descritos a continuación, aunque debe resaltarse que los resultados de los análisis con VOGDB ayudaron a dar mayor solidez a los resultados obtenidos para dichas secuencias.

La selección de secuencias virales putativas de DNA se realizó considerando los mejores hits de las secuencias de proteínas *queries* a las bases de datos RefSeq, VOGDB y NR. Así, se recuperaron cinco contigs relacionados a la familia *Poxviridae*, siendo cuatro de ellos asignados a la subfamilia *Entomopoxvirinae*. También se recuperaron dos contigs relacionados a la familia *Baculoviridae* (una de ellas al género *Alphabaculovirus* y la otra al género *Betabaculovirus*) y uno más relacionado a la familia *Polydnaviridae* y el género *Bracovirus* (**tabla 13**). Dado el pequeño tamaño de los contigs respecto al genoma de los taxa de virus de dsDNA de referencia, y debido a que no se encontró un taxa viral de referencia común para los contigs, ni proteínas útiles para la reconstrucción filogenética de virus de dsDNA (p.ej. DNA polimerasa), no se presenta reconstrucciones de organización genómica en las secciones siguientes. Cabe mencionar que del conjunto de ~200 secuencias consideradas para la selección manual de estos grupos, casi todas presentaron hit con los mismos *subjects* de las secuencias finales seleccionadas, siendo estas últimas las que presentaron los mejores valores de alineamiento de todos los *queries*. Además, en la selección manual se observó que incluso las secuencias que presentaron hit a VOGs o dominios conservados de virus de dsDNA (por ejemplo los *subjects* VOG0732, de UDP-glucosiltransferasa ecdisteroide; el dominio pfam01728, deFtsJ-like metiltransferasa; el dominio pfam01607, de Peritrofina-A de unión a quitina; el dominio pfam13402 de peptidasa M60 y enhancina; cd00022, de dominio de repetición BIR, de inhibición de apoptosis de baculovirus), también presentaron hits con secuencias de organismos celulares de la base de datos NR.

Tabla 12. Resumen de las secuencias de virus de RNA putativos recuperadas en la selección manual y resultados de los análisis de similitud con BLASTP/RPS-BLAST/HMMER. Las secuencias de *subjects* que se presentan corresponden al mejor hit de las bases de datos NR y RefSeq. Los espacios con "-" indican que la secuencia no obtuvo hit con las bases de datos (los valores de alineamiento no aplican).

Taxonomía tentativa		Contig (Longitud en nt)	No. de ORF / long. en aa de proteína predicha	BLASTP vs NR				BLASTP vs RefSeq				RPS-BLAST/HMMER	
				Descripción del hit	qcov	id	e	Descripción del hit	qcov	id	e	Dominio; ID	e
Tymovirales (+ssRNA)	*Tymoviridae, Marafivirus*	PF_mfv_1 (6423)	ORF1/2093	AKZ17760.1: polyprotein [Grapevine Syrah virus 1]	98	93.6	0	YP_002756536.1: putative 230 kDa polyprotein [Grapevine Syrah virus 1]	98	93.8	0	MTR; PF01660	4.08E-73
												PRO; PF05381	1.03E-17
												HEL1; PF01443	9.75E-48
												RdRP_2; PF00978	4E-6
												Tymo_coat; PF00983	8.81E-09
		PF_mfv_2 (4305)	ORF1/1379	YP_009351862.1: polyprotein [Grapevine rupestris vein feathering virus]	98	97	0	YP_009351862.1: Grapevine rupestris vein feathering virus	98	97	0	PRO; PF05381	2.76E-25
												HEL1; PF01443	7.85E-50
												RdRP_2; PF00978	9.24E-16
												Tymo_coat; PF00983	8.72E-12
Picornavirales (+ssRNA)	*Dicistroviridae, Cripavirus*	PF_dv_1 (9902)	ORF1/1925	QKF95572.1: hypothetical protein 1 [Leibnitzia anandria dicistrovirus]	99	82	0	YP_009345032.1: hypothetical protein 1 [Wuhan insect virus 33]	92	32	0	HEL; PF00910	4.84E-25
												PC3; PF00548	0.0058
												RdRP_1; cd01699	8.05E-45
			ORF2/844	QKF95573.1: hypothetical protein 2 [Leibnitzia anandria dicistrovirus]	100	92	0	NP_733846.1: capsid protein precursor, partial [Aphid lethal paralysis virus]	91	35	0	Rhv_capsid; cd00205	3.36E-22
												Rhv_capsid; cd00205	4.08E-26
												CRPV_capsid; PF08762	8.66E-28
	Iflaviridae	PF_ifv_1 (2209)	ORF1/529	YP_009342053.1: hypothetical protein [Wuhan coneheads virus 1]	96	35	2E-59	YP_009342053.1: hypothetical protein [Wuhan coneheads virus 1]	96	35	1E-62	Rhv_like; cd00205	1.07E-13

Picornavirales (+ssRNA)	*Iflaviridae* (cont.)	PF_ifv_2 (1236)	ORF1/411	YP_009342053.1: hypothetical protein [Wuhan coneheads virus 1]	90	36	7E-68	YP_009342053.1: hypothetical protein [Wuhan coneheads virus 1]	91	36	5E-71	Calici_coat; PF00915	1.79E-07
		PF_ifv_3 (2158)	ORF1/684	YP_009342053.1: hypothetical protein [Wuhan coneheads virus 1]	97	35	3E-121	YP_009342053.1: hypothetical protein [Wuhan coneheads virus 1]	97	35	2E-124	CRPV_capsid; PF08762	3.45E-12
		PF_ifv_4 (751)	ORF1/192	YP_009342053.1: hypothetical protein [Wuhan coneheads virus 1]	92	50.5	2E-47	YP_009342053.1: hypothetical protein [Wuhan coneheads virus 1]	92	50.5	2E-50	HEL; PF00910	5.56E-11
		PF_ifv_5 (1253)	ORF1/417	ACN94442.1: protease/RNA-dependent RNA polymerase [Nasonia vitripennis virus]	100	60	1E-108	YP_009111311.1: polyprotein [Dinocampus coccinellae paralysis virus]	95	50	2E-128	RdRP_1; cd01699	3.8E-09
		PF_ifv_6 (895)	ORF1/264	ACN94442.1: protease/RNA-dependent RNA polymerase [Nasonia vitripennis virus]	100	60	1E-108	YP_009342053.1: hypothetical protein [Wuhan coneheads virus 1]	100	60	6E-100	RdRP_1; cd01699	2.01E-33
	No clasificado (Picorna-like)	**Contig (Long. en nt)**	**No. de ORF / long. en aa de proteína predicha**	**BLASTP vs NR**				**BLASTP vs RefSeq**				**RPS-BLAST/HMMER**	
				Descripción del hit	**qcov**	**id**	**e**	**Descripción del hit**	**qcov**	**id**	**e**	**Dominio; ID**	**e**
		PF_pv_1 (1057)	ORF1/213	QKK82984.1: hypothetical protein 1 [Teucrium fruticans picorna-like virus]	100	58	4E-77	YP_009337723.1: hypothetical protein 1 [Hubei picorna-like virus 51]	100	58	1E-75	Rhv_like; Cd00205	1.93E-14

		PF_pv_2 (3401)	ORF1/580	AUH27291.1: coat protein [Maize-associated picornavirus]	96	44.5	4E-143	YP_009551962.1: coat protein [Cherry virus Trakiya]	98	40	1E-124	Rhv_like; Cd00205	0
												CRPV_capsid; PF08762	6.93E-07
			ORF2/292	-	-	-	-	-	-	-	-	-	-
		PF_pv_3 (2004)	ORF1/667	QIN54759.1: polyprotein [Tetranychus urticae-associated picorna-like virus 1]	71	28	7E-46	YP_009551963.1: replicase [Cherry virus Trakiya]	57	30.5	6E-51	HEL; PF00910	6.45E-22
		PF_pv_4 (1345)	ORF1/395	AUH27292.1: polyprotein [Maize-associated picornavirus]	85	29	6E-25	YP_009551963.1: replicase [Cherry virus Trakiya]	81	29	3E-23	-	-
		PF_pv_5 (1631)	ORF1/540	AUH27292.1: polyprotein [Maize-associated picornavirus]	93	44.5	5E-136	YP_009337724.1: hypothetical protein 2 [Hubei picorna-like virus 51]	92	46	1E-137	RdRP_1; cd01699	3.51E-56
Reovirales (dsRNA)	*Reoviridae*	**Contig (Long. en nt)**	**No. de ORF / long. en aa de proteína predicha**	**BLASTP vs NR**				**BLASTP vs RefSeq**				**RPS-BLAST/HMMER**	
				Descripción del hit	**qcov**	**id**	**e**	**Descripción del hit**	**qcov**	**id**	**e**	**Dominio; ID**	**e**
		PF_rv_1 (3620)	ORF1/1149	AYP67577.1: RNA-dependent RNA polymerase [Shelly beach virus]	97	36	0	YP_392501.1: RNA-dependent RNA polymerase (S1) [Operophtera brumata reovirus]	99	34	7E-172	RdRP_5; PF07925	0
		PF_rv_2 (1779)	ORF1/524	QBA09478.1: hypothetical protein [Reoviridae sp. BF02/7/10]	95	31	4E-65	YP_392502.1: Operophtera brumata reovirus (S2)	91	25	1E-40	-	-

		PF_rv_3 (1821)	ORF1/562	APG79179.1: hypothetical protein 3 [Hubei diptera virus 21]	95	31	1E-55	YP_392503.1: hypothetical protein (S3) [Operophtera brumata reovirus]	86	31	1E-51	-	-
		PF_rv_4 (3413)	ORF1/1086	AVO64752.1: S4 [High Island virus]	48	42	6E-133	YP_392504.1: polyhedrin (S4) [Operophtera brumata reovirus]	41	37	8.7E-83	TBCC_N; PF16752	0.014
		PF_rv_5 (1980)	ORF1/607	YP_392505.1: hypothetical protein [Operophtera brumata reovirus]	61	23	4E-08	YP_392505.1: hypothetical protein (S5) [Operophtera brumata reovirus]	61	23	3E-11	-	-
		PF_rv_6 (1668)	ORF1/524	QBA09481.1: hypothetical protein [Reoviridae sp. BF02/7/10]	97	26	2E-45	YP_392506.1: hypothetical protein (S6) [Operophtera brumata reovirus]	94	24	3E-27	-	-
		PF_rv_7 (1713)	ORF1/517	RVD90501.1: hypothetical protein TUBRATIS_31610 [Tubulinosema ratisbonensis]	55	25	4E-17	YP_392507.1: hypothetical protein (S7) [Operophtera brumata reovirus]	35	20	4E-04	-	-
		PF_rv_8 (735)	ORF1/205	AWA82242.1: hypothetical protein [Eccles virus]	99	27	2E-17	YP_392508.1: hypothetical protein (S8) [Operophtera brumata reovirus]	76	24	4E-12	-	-
Mononegavirales (-ssRNA)	*Rhabdoviridae*	**Contig (Long. en nt)**	**No. de ORF / long. en aa de proteína predicha**	**BLASTP vs NR**				**BLASTP vs RefSeq**				**RPS-BLAST/HMMER**	
				Descripción del hit	**qcov**	**id**	**e**	**Descripción del hit**	**qcov**	**id**	**e**	**Dominio; ID**	**e**
		PF_rbv_1 (1018)	ORF1/235	YP_009337067.1: putative nucleoprotein [Hubei dimarhabdovirus virus 2]	99	51	1E-74	YP_009337067.1: putative nucleoprotein [Hubei dimarhabdovirus virus 2]	99	51	1E-77	Rhabdo ncap; PF00945	1.08E-45

		PF_rbv_2 (1199)	ORF1/121	QBP46938.1: matrix protein [Tetrastichus brontispae RNA virus 1]	56	37	8E-6	YP_009301741.1: matrix protein [Wuhan Insect virus 7]	72	33	3E-08	MATRX_BEFB; VOG1466	8E-13
		PF_rbv_3 (2405)	ORF1/402	YP_009301742.1: glycoprotein [Wuhan Insect virus 7]	85	38	E-83	YP_009301742.1: glycoprotein [Wuhan Insect virus 7]	86	38	5E-86	Rhabdo_glycop; PF00974	3.07E-17
		PF_rbv_4 (2529)	ORF1/841	YP_009337071.1: RNA-dependent RNA polymerase [Hubei dimarhabdovirus virus 2]	99	60	0	YP_009301743.1: RNA-dependent RNA polymerase [Wuhan Insect virus 7]	99	0	0	Mononeg_RdRP; PF00946	0
		Contig (Long. nt)	**No. de ORF / long. en aa de proteína predicha**	**BLASTP vs NR**				**BLASTP vs RefSeq**				**RPS-BLAST/HMMER**	
				Descripción del hit	**qcov**	**id**	**e**	**Descripción del hit**	**qcov**	**id**	**e**	**Dominio; ID**	**e**
Tolivirales (+ssRNA)	*Tombusviridae*	PF_tbv_1 (2720)	ORF1/441	APG76440.1: hypothetical protein 3 [Sanxia water strider virus 14]	95	35	2E-62	YP_009337188.1: hypothetical protein 2 [Sanxia tombus-like virus 5]	94	38	1E-77	RdRP_3	9.39E-61
			ORF2/164	YP_009337379.1: putative capsid protein [Hubei permutotetra-like virus 4]	98	30	2E-09	YP_009337379.1: putative capsid protein [Hubei permutotetra-like virus 4]	98	30	1E-12	Viral_Coat	4.54E-07

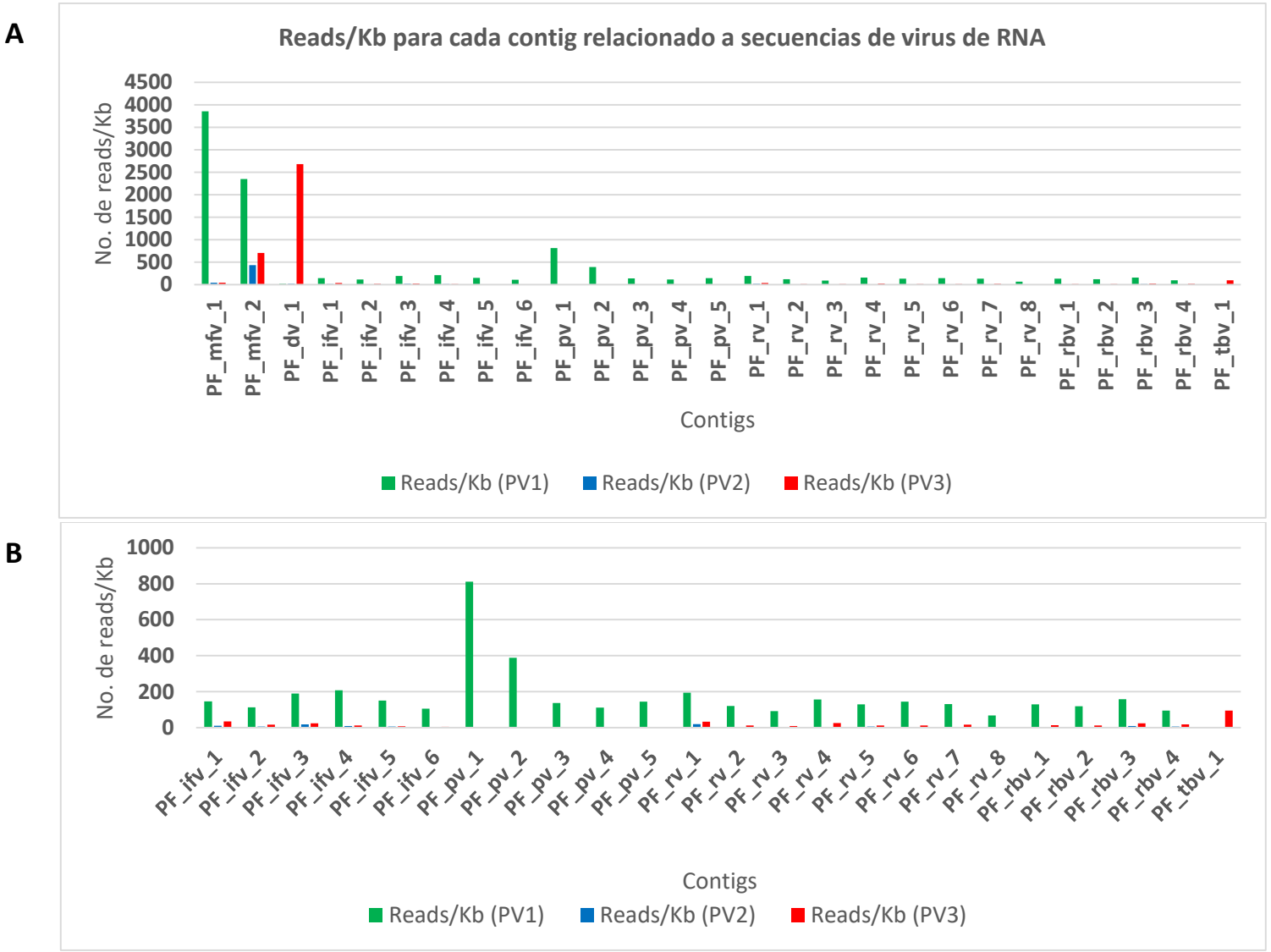

Figura 10. Gráficas reads/Kb de cada biblioteca mapeados a las secuencias seleccionadas. (A) Se muestran los reads/Kb de cada biblioteca para las 27 secuencias seleccionadas. (B) Se muestra lo mismo que en (A), omitiendo los contigs PF_mfv_1, PF_mfv_2 y PF_dv_1 para cambiar la escala.

Tabla 13. Resumen de las secuencias de virus de DNA putativos recuperadas en la selección manual y resultados de los análisis de similitud con BLASTP/RPS-BLAST/HMMER. Las secuencias *subject* que se presentan corresponden al mejor hit de las bases de datos NR y RefSeq. Los espacios con "-" indican que la secuencia no obtuvo hit con las bases de datos (los valores de alinemiento no aplican).

Taxonomía tentativa		Contig (Long. en nt)	No. de ORF / long. en aa de proteína predicha	BLASTP vs NR				BLASTP vs RefSeq				RPS-BLAST/HMMER	
				Descripción del hit	qcov	id	e	Descripción del hit	qcov	id	e	Dominio; ID	e
Baculoviridae (dsDNA)	*Betabaculovirus*	PF_bac_1 (596)	ORF/198	YP_006340.1: ORF4 [Agrotis segetum granulovirus]	95	30	9.00E-12	YP_009513034.1: hypothetical protein AsGV004 [Agrotis segetum granulovirus]	95	30	2.00E-14	-	-
	Alphabaculovirus	PF_bac_2 (894)	ORF/101	NP_818684.1: inhibitor of apoptosis protein 4 [Adoxophyes honmai nucleopolyhedrovirus]	85	41	5.00E-09	NP_818684.1: inhibitor of apoptosis protein 4 [Adoxophyes honmai nucleopolyhedrovirus]	85	41	1.00E-11	-	-
Poxviridae (dsDNA)	*Betaentomopoxvirus*	PF_pox_1 (2676)	ORF/780	YP_008003527.1: unknown similar to AgseGV orf4 [Mythimna separata entomopoxvirus 'L']	97	27	8.00E-71	YP_008003527.1: unknown similar to AgseGV orf4 [Mythimna separata entomopoxvirus 'L']	97	27	8.00E-71	-	-
		PF_pox_2 (1090)	ORF/307	XP_008699194.1: PREDICTED: kelch repeat and BTB domain-containing protein 4 isoform X2 [Ursus maritimus]	46	26	1.14E-06	YP_004821521.1: kelch-like protein [Yokapox virus]	55	30	1E-05	-	-

Taxonomía tentativa		Contig (Long. en nt)	No. de ORF / long. en aa de proteína predicha	BLASTP vs NR				BLASTP vs RefSeq				RPS-BLAST/HMMER	
				Descripción del hit	qcov	id	e	Descripción del hit	qcov	id	e	Dominio; ID	e
Baculoviridae (dsDNA)	*Betaentomopoxvirus*	PF_pox_3 (528)	ORF/159	YP_008003948.1: inhibitor of apoptosis 4 [Adoxophyes honmai entomopoxvirus 'L']	67	38	0	YP_008003948.1: inhibitor of apoptosis 4 [Adoxophyes honmai entomopoxvirus 'L']	67	38	1.80E-21	-	-
	Betaentomopoxvirus	PF_pox_4 (1037)	ORF/240	NP_064801.1: hypothetical protein AMV019 [Amsacta moorei entomopoxvirus]	94	32	4.55E-17	NP_064801.1: hypothetical protein AMV019 [Amsacta moorei entomopoxvirus]	94	32	9.84E-20	-	-
	Entomopoxvirinae	PF_pox_5 (828)	ORF/224	NP_048132.1: ORF MSV061 putative LINE reverse transcriptase, similar to Caenorhabditis elegans GB:U00063 [Melanoplus sanguinipes entomopoxvirus]	91	34	2.00E-32	NP_048132.1: ORF MSV061 putative LINE reverse transcriptase, similar to Caenorhabditis elegans GB:U00063 [Melanoplus sanguinipes entomopoxvirus]	91	34	2.00E-32	-	-
Polydnaviridae (dsDNA)	*Bracovirus*	PF_pol_1 (2384)	ORF/232	YP_184883.1: unnamed protein product [Cotesia congregata bracovirus]	91	42	4.00E-47	YP_184883.1: unnamed protein product [Cotesia congregata bracovirus]	91	42	4.00E-47	-	-

3.4.1. Secuencias virales putativas relacionadas con la familia *Dicistroviridae*

El primer genoma viral putativo de RNA con relación al orden *Picornavirales* es el contig PF_dv_1, de 9902 nt de longitud. En su secuencia se detectaron tres ORFs no empalmados, siendo uno de ellos <100 aa, por lo que no será considerado en lo siguiente. En dirección 5' → 3' y tomando como referencia la secuencia del contig, el primer ORF considerado (ORF1, que codifica una proteína predicha de 1925 aa de longitud) abarca la región desde el nt 588 al nt 6362, mientras que el segundo ORF (ORF2, que codifica una proteína predicha de 844 aa de longitud) abarca la región del nt 7029 al nt 9560. Los ORFs anteriores están separados entre sí por una región intergénica (IGR) de 665 nt de longitud, y separados de los extremos del contig por regiones no codificantes (NCR); la primera de 587 nt, en el extremo 5', y la segunda de 342 nt, en el extremo 3' (**figura 11**).

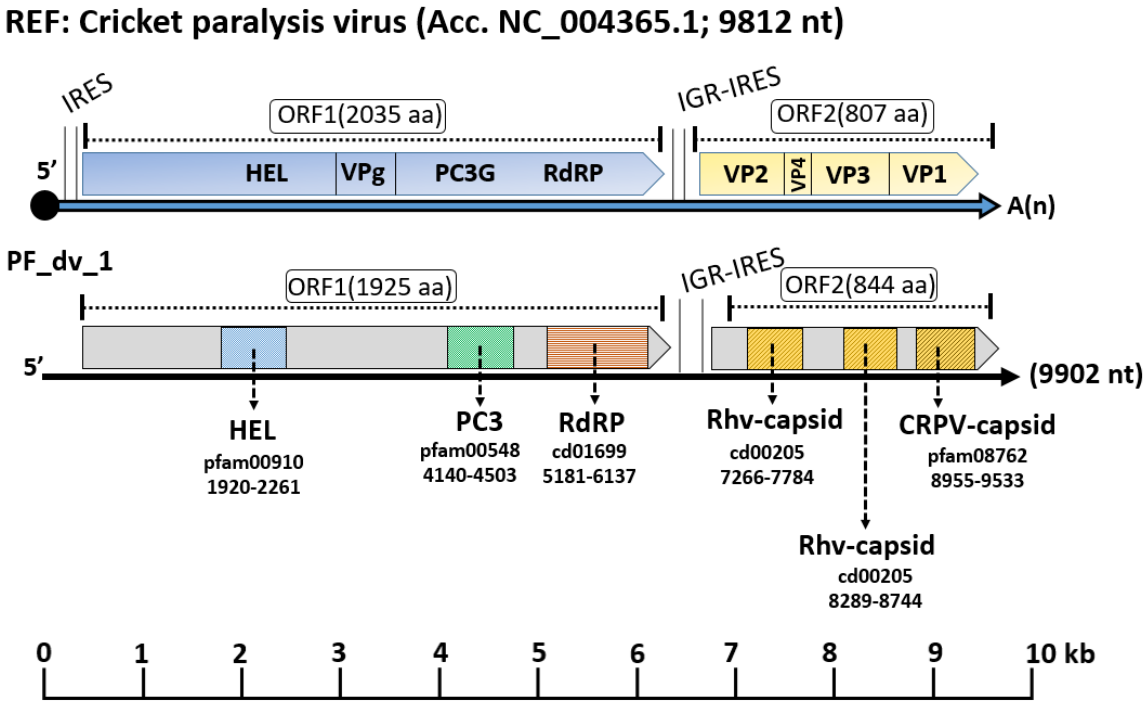

Figura 11. Organización genómica predicha para el contig PF_dv_1. Se muestra el contig (flecha negra sólida), los ORFs encontrados en su secuencia (cajas grises sobre el contig), así como el ID de cada dominio encontrado en las secuencias de proteína predichas para los ORFs. Las coordenadas de los dominios indican en que región de la secuencia nucleotídica se encuentran codificados. Se muestra también la IGR donde se haya un IRES tipo 1 predicho con base en la similitud de secuencias con otros IREs de dicistroviridos. Por comparación, se incluye la organización genómica de la especie tipo de los cripavirus (Cricket paralysis virus, flecha azul), su número de acceso en el Genebank, la longitud del genoma completo y los ORFs con las proteínas que codifcan (cajas de color sobre el genoma). A(n): cola poli-A en el extremo 3'; proteína VPg en el extremo 5' representada con un círculo negro sólido.

El análisis de BLASTP contra la base de datos de RefSeq mostró similitud, aunque limitada, del producto proteico predicho del ORF1 de PF_dv_1 con la poliproteína no estructural del ORF1 de Wuhan Insect virus 33 (WiV13; id=32%, qcov=98% y e=0), con la poliproteína del ORF1 de aphid lethal paralysis virus (ALPV; id=32.4%, qcov=78% y e=9E-151), y con la poliproteína del ORF1 de Big Sioux River virus (BSRV; id=32.4%, qcov=78, e=1.6E-149) (**tabla 12**). También se observaron hits del producto del ORF1 de PF_dv_1 con las poliproteínas no estructurales respectivas de Hubei picorna-like virus 14 (HpV14), del ORF1 de Hubei picorna-like virus 15 (HpV15) y del ORF1 de cricket paralysis virus (CRPV). Todos los *subjects* mencionados representan taxa picorna-like con genomas bicistrónicos, cuya secuencia y organización genómica está relacionada con la familia *Dicistroviridae* (pero que no son miembros reconocidos de esta), por lo que se encuentran agrupados en el clado "*Dicistroviridae*" descrito por Shi et al. (2016), que incluye a los miembros reconocidos de la familia *Dicistroviridae* (la cual incluye mayormente a virus que han sido caracterizados previamente como ISVs) y otros virus picorna-like relacionados (los cuales se han encontrado predominantemente a partir de muestras de insectos).

Las búsquedas de BLASTP contra la base de datos RefSeq hechas con la secuencia de proteína predicha del ORF2 de PF_dv_1 mostraron similitud de esta con la secuencia de poliproteína estructural del ORF2 de aphid lethal paralysis virus (ALPV; id=35, qcov=96 y e=7E-151), con la poliproteína del ORF2 de Anopheles C virus (ACV; id=33, qcov=91 y e=9.7E-141) y con la poliproteína del ORF2 de Aphis dicistrovirus (id=36, qcov=91 y e=9.6E-140) (**tabla 12**). Otras secuencias de poliproteína de los genomas de HpV15, Aphis glycines virus 3 (ApGlV3) y CRPV también fueron hits. Al igual que en el caso de la proteína del ORF1, los hits de la proteína del ORF2 de PF_dv_1 son secuencias de taxa picorna-like relacionados a dicistroviridos[2].

En conjunto, los resultados descritos de BLASTP con la base de datos RefSeq indicaron inicialmente que el contig PF_dv_1 es más semejante en su secuencia y organización genómica a los picorna-like relacionados con dicistroviridos, que a cualquier otro grupo de virus picorna-like. Presenta similitud significativa (si bien con identidad baja) a los *subjects* de las bases de datos usadas en las búsquedas iniciales de secuencias virales, lo que permitió recuperarlo durane la selección manual. Por otro lado, los análisis de BLASTP contra

[2] Nota: En ocasiones, los nombres de dos taxa virales supra-específicos tienen la misma raíz (por ejemplo, el género *Flavivirus* y la familia *Flaviviridae),* por lo que puede existir ambigüedad al usar el nombre colectivo "flavivirus", ya que puede entenderse como una referencia para ambos grupos. Por lo tanto, en la comunidad virológica se ha adoptado el uso de sufijos específicos para cada taxa; así, los términos "raíz-viridos" y "raiz-virus" hacen referencia, respectivamente, a los miembros de la familia y del género (siguiendo el ejemplo anterior, los miembros de la familia *Flaviviridae* son referidos como flaviviridos y los miembros del género *Flavivirus* como flavivirus). Por lo tanto, en el presente trabajo se usará siempre el sufijo "-viridos" para referirse a los miembros de una familia, reservando el uso del sufijo "-virus" para referirse a los géneros, independientemente de si hay o no una raíz común para los nombres de dichos taxa (ICTV, 2020b).

la base de datos NR con cada una de las proteínas de los ORFs de PF_dv_1 detectaron taxa virales cuyas secuencias mostraron una similitud mucho más elevada en comparación con la obtenida con los hits de RefSeq (**tabla 12**). Las secuencias de proteínas predichas en ambos ORFs de PF_dv_1 presentaron como mejores hits a las poliproteínas hipotéticas de Leibnitzia anandria dicistrovirus (LeDV), un genoma viral putativo picorna-like relacionado con la familia *Dicistroviridae*, reportado recientemente por Zhao et al. (2020, sin publicar). La similitud de la proteína del ORF2 de PF_dv_1 con la proteína del ORF2 de LeDV fue más elevada (id=92%, qcov=100% y e=0) que la obtenida por la proteína del ORF1 de PF_dv_1 con la proteína del ORF1 de LeDV (id=82%, qcov=99% y e=0). El segundo mejor hit de las proteínas del ORF1 y el ORF2 de PV_dv_1 fueron, respectivamente, las proteínas del ORF1 y ORF2 de otro genoma dicistrónico picorna-like relacionado a dicistroviridos descrito recientemente por Luria et al. (2020), denominado Phenacoccus solenopsis virus (PhSoV). De nuevo, la similitud del producto del ORF2 (qcov=96%, id=89.6%, e=0) fue más alta que la obtenida por el producto de ORF1 (qcov=99%, id=79.6%, e=0) con sus respectivos *subjects*. Al examinar el heatmap de la matriz de porcentajes de indentidad pareada (que indica la similitud en función de la densidad del color) entre las proteínas correspondientes de ambos ORFs de dicistroviridos, se puede observar una identidad entre las proteínas del ORF2 de PF_dv_1, PhSov y LeDV (89-92%) más alta que la observada para las proteínas del ORF1 de dichos taxa (80-83%). Además, en dicha matriz se puede observar que la similitud de las proteínas de ambos ORFs de PF_dv_1, PhSov y LeDV es alta en el contexto de las especies de virus de la familia *Dicistroviridae* y taxa relacionados (**figura 12**). Dada la similitud entre los productos proteícos de los ORFs de PF_dv_1, PhSoV y LeDV (específicamente la similitud de ~90% entre las proteínas del ORF2 de dichos taxa, que se considera como un limite inferior de similitud en la demarcación de aislados de una especie de dicistrovirido; ver sección correspondiente del siguiente capítulo), y debido a que PhSoV y LeDV se encontraron en especies diferentes de organismos respecto a PF_dv_1 (y en el caso de PhSoV, en individuos del piojo harinoso *Ph. solenopsis* colectados en otras regiones del mundo), es posible que este último corresponda a un virus previamente no reportado, i.e., distinto de PhSoV y LeDV.

Un análisis adicional de BLASTN (algoritmo blastn) en la plataforma del NCBI, considerando la secuencia completa de PF_dv_1, LeDV y PhSoV para verificar la similitud entre sus genomas, resultó nuevamente en una mayor identidad entre las secuencias de PF_dv_1 y LeDV (qcov=90%, id=76% y e=0) que entre PF_dv_1 con PhSoV (qcov=98%, id=73% y e=0), aunque en este último caso hubo mayor cobertura del *query*. De nuevo, al considerar la identidad moderada entre las secuencias nucleotídicas completas de PF_dv_1, PhSoV y LeDV de ~70% (aunado a la similitud entre sus proteínas estructurales y no estructurales descrita anteriormente), cabe hipotetizar a PF_dv_1 como el genoma casi completo de un virus de RNA novedoso. Respecto a las regiones no codificantes, la secuencia de LeDV presenta una NCR 5'-proximal de 58 nt, una

NCR 3’ proximal de 187 nt, y una IGR de 642 nt. Por su parte, PhSoV tiene una NCR 5’-proximal de 554 nt, una NCR 3’-proximal de 349 nt, y una IGR de 723 nt. Por tanto, todas las NCR de PF_dv_1 son de una longitud más similar a las regiones correspondientes de PhSoV que a las de LeDV. Sin embargo, es importante mencionar que la secuencia de LeDV no es un genoma terminado (i.e., las secuencias de sus NCRs terminales no están completas).

Respecto a la anotación de las secuencias de proteína de los ORFs de PF_dv_1, se encontró que la proteína del ORF1 contiene dominios no estructurales de helicasa (HEL, pfam00910; del aa 445 al aa 559), proteasa C3 (PC3, pfam00548; del aa 1163 al aa 1327) y de polimerasa de RNA dependiende de RNA (RdRP, cd01699; del aa 1532 al aa 1850) en el orden HEL-PC3-RdRP. Los motivos A, B y C característicos de la superfamilia F3 de helicasas se encuentran presentes (con ligeras variaciones en el motivo C) en el dominio HEL (Yasmin et al., 2020). El motivo de las proteínas cisteína-proteasas (con el residuo activo de cisteina característico) se encontró en el dominio PC3 (Yasmin et al., 2020). En cuanto al dominio de RdRP, se observaron los motivos A, B y C, altamente conservados del subdominio de “palma αβ” (Gorbalenya et al., 2002). Por su parte, la poliproteína del ORF2 de PF_dv_1 presentó tres dominios de cápside JRC; dos dominios de cápside de Rhv_capsid (cd00205; el primero del aa 80 al aa 252 y el segundo del aa 421 al aa 572) y un dominio de cápside CRPV_capsid (pfam08762; del aa 643 al aa 835). Las regiones de dichos dominios estructurales coinciden en el alineamiento, respectivamente, con la posición de las proteínas de cápside VP2, VP3(-VP4) y VP1 en la poliproteína de referencia. De acuerdo con Chen et al. (2012a), en la mayoría de las especies de dicistroviridos las proteínas VP4 y VP3 se forman por escición pos-traduccional del precursor denominado VP0, por lo que es probable que el segundo dominio Rhv_capsid del producto proteíco estructural predicho de PF_dv_1 corresponda a un precursor VP0. Sin embargo, el genoma de CRPV (el taxa representativo de los cripavirus) presenta un dominio anotado como Dicistro_VP4 (cd288362; independiente de los dominios de cápside correspondientes a VP1, VP2 y VP3), el cual no fue encontrado en el producto del ORF2 de PF_dv_1. Por otra parte, en los análisis de BLASTP o HMMER no se detectó algúna región que correspondiera a la proteína VPg (**figura 11**).

De manera correspondiente con los análisis de similitud, las relaciones filogenéticas resueltas con las proteínas de ambos ORFs de PF_dv_1 por separado (**figuras 13** y **14**) lo agrupan en un clado compartido entre taxa del género *Cripavirus* como ALPV o Rhopalosiphum padi virus (RhPV) y taxa picorna-like relacionados a dicistroviridos como PhSoV, LeDV, Solenopsis invicta virus 6 (SIV6) o Bundaberg bee virus 1 (BbV1). Si bien en ambas filogenias la rama de PF_dv_1 muestra una relación más estrecha con los taxa del género *Cripavirus*, únicamente en la filogenia de la poliproteína del ORF2 se agrupa en clados separados a los miembros de cada género de dicistroviridos; aparavirus, triatovirus y cripavirus (**figura 14**).

Considerando la relevancia extensamente reportada de la IGR y el IRES que contiene, necesaria para la traducción de la proteína estructural de los genomas de dicistroviridos (Jan, 2006; Hertz y Thompson, 2011), además de que los análisis filogenéticos indican que PF_dv_1 se relacionan más estrechamente con los miembros del género *Cripavirus*, se realizó un alineamiento múltiple de las IGRs de taxa selectos de dicho género y la IGR de PF_dv_1 para verificar la presencia de los sitios conservados característicos de las IRES tipo 1 en este último (**figura 15**).

Para los taxa incluidos en este análisis (excepto para PF_dv_1 y LeDV) se consideró la región del genoma comprendida desde el primer nucleótido después de la terminación del ORF1 hasta un nucleótido antes del inicio del ORF2, según la anotación de cada genoma en el Genebank. Respecto a PF_pv_1 se consideró la región del genoma mencionada anteriormente, más una secuencia de 100 nt después del inicio del ORF2. En cuanto a LeDV se consideró la región del genoma del nt 6150 al nt 6575 (316 nt después de la terminación del ORF1 y 100 nt después del inicio del ORF2). En el MSA de las secuencias anteriores se observaron sitios clave altamente conservados para la formación de las IRES de tipo 1 en las IGR de PF_dv_1. Específicamente, se observaron las secuencias invertidas complementrias (SIC) de las estructuras “pseudoknot” 1, 2 y 3, las secuencias correspondientes a los “stem-loop” 1 y 2, así como las secuencias correspondientes a dos “bulge” en la IGR de PF_dv_1. A excepción de una sustitución puntual de T por A en la secuencia del “Stem-loop 1” de PF_dv_1, las secuencias conservadas de la IRES de este último resultaron ser idénticas a las correspondientes en la IGR de PhSoV y LeDV (**figura 15**).

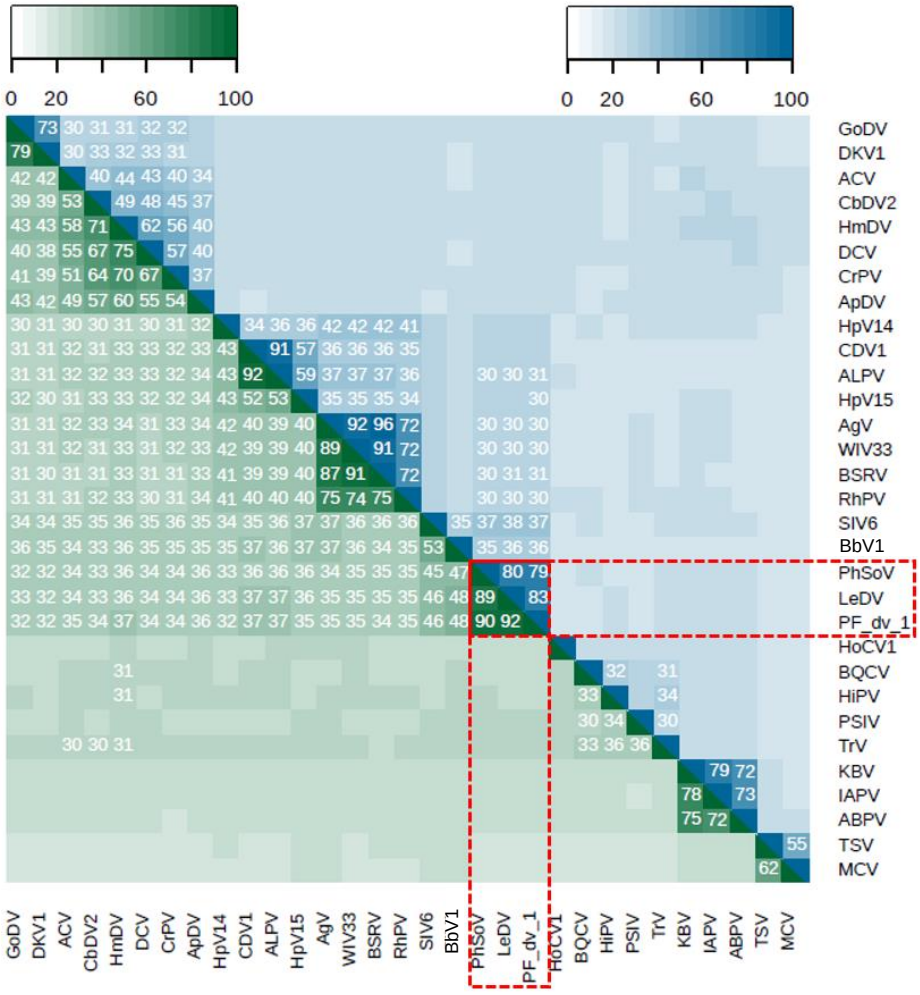

Figura 12. Matriz de porcentajes de identidad pareada de poliproteínas codificadas por dicistroviridos, PF_dv_1 y taxa de virus picorna-like. Los valores de %id correspondientes a la proteína del ORF2 se muestran debajo de la diagonal (verde) y los de la proteína del ORF1 arriba de la diagonal (azul). La intensidad del color de cada casilla va en función del valor de %id indicado dentro de cada casilla. Se muestran sólo los valores de %id ≥ 30. Se remarca en margen rojo los valores correspondientes a los ORFs de PF_dv_1, LeDV y PhSoV. Los taxa incluidos son: goose dicistrovirus (GoDV; YP_009221981.1), Drosophila kikkawai virus 1 (DKV1; AYQ66681.1), Anopheles C virus (ACV; YP_009252204.1), Caledonia beadlet anemone dicistro-like virus 2 (CbDV2; ASM93984.1), Hypsignathus monstrosus dicistrovirus (HmDV; AZR39355.1), Drosophila C virus (DCV; NP_044945.1), cricket paralysis virus (CrPV; NP_647481.1), Apis dicistrovirus (ApDV; YP_009388499.1), Hubei picorna-like virus 14 (HpV14; YP_009337313.1), Culex dicistrovirus 1 (CDV1; AXQ04775.1), aphid lethal paralysis virus (ALPV; NP_733845.1), Hubei picorna-like virus 15 (HpV15; YP_009336540.1), Aphis gossypii virus (AgV; AYH52676.1), Wuhan insect virus 33 (WIV33; YP_009345032.1), Big Sioux River virus (BSRV; YP_009389287.1), Rhopalosiphum padi virus (RhPV; NP_046155.1), Solenopsis invicta virus 6 (SIV6; QBL75887.1), Bundaberg bee virus 1 (BbV1; AWK77854.1), Phenacoccus solenopsis virus (PhSoV; QIU80542.1), Leibnitzia anandria dicistrovirus (LeDV; QKF95572.1), Homalodisca coagulata virus 1 (HoCV1; YP_610950.1), black queen cell virus (BQCV; NP_620564.1), Himetobi P virus (HiPV; NP_620560.1), Plautia stali intestine virus (PSIV; NP_620555.1), Triatoma virus (TrV; NP_620562.1), Kashmir bee virus (KBV; NP_851403.1), Israeli acute paralysis virus (IAPV; YP_001040002.1), acute bee paralysis virus (ABPV; NP_066241.1), Taura syndrome virus (TSV; NP_149057.1), mud crab virus (MCV; YP_004063985.1).

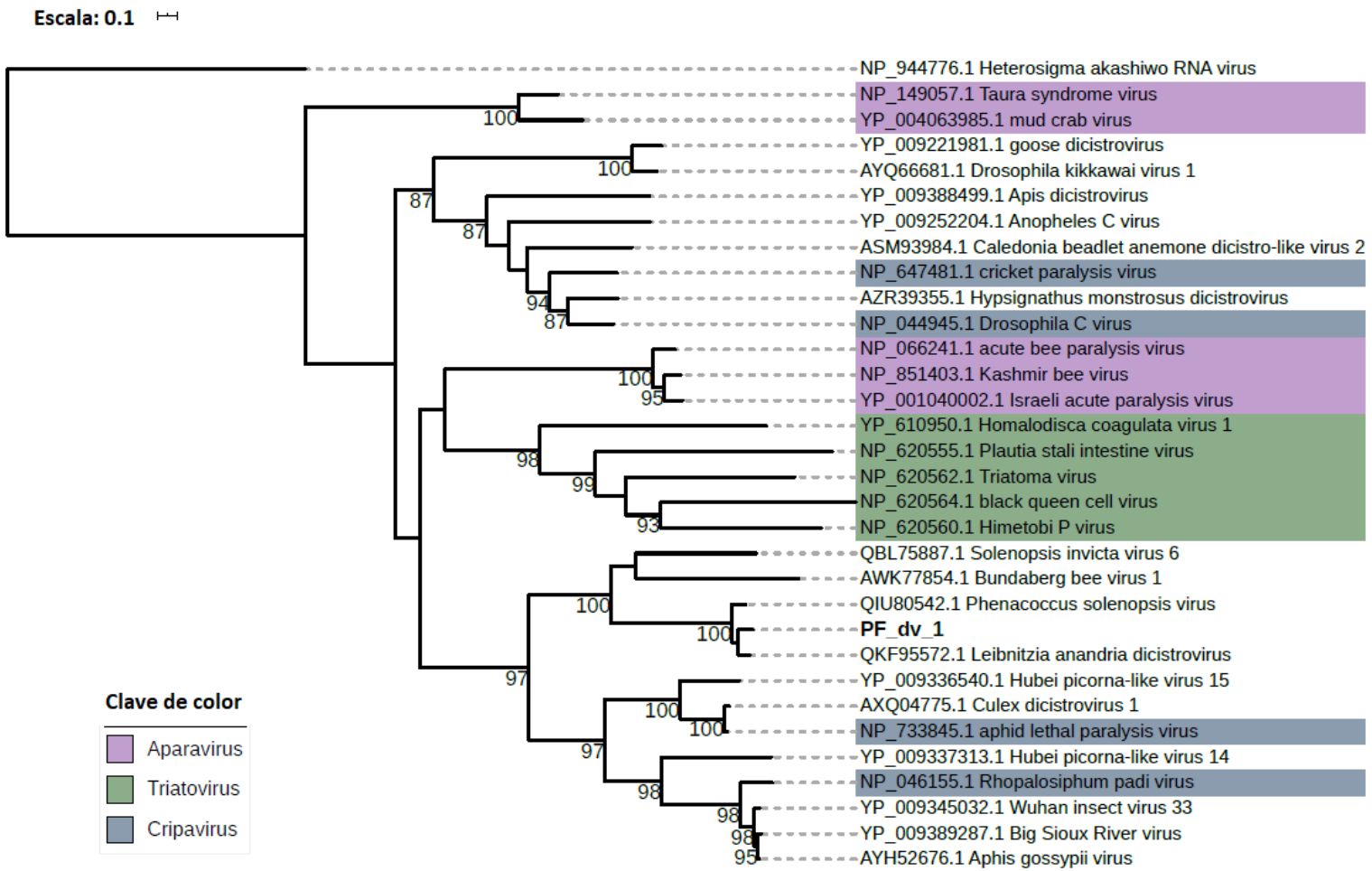

Figura 13. Árbol filogenético construido con las secuencias completas de poliproteína no estructural del ORF1 de taxa de dicistroviridos, PF_dv_1 y otros taxa de virus picorna-like. Los nombres de los taxa no clasificados no se remarcan en color. Se puede observar que no hay un único clado para los miembros de cada uno de los tres géneros de dicistroviridos, lo cual sí ocurre en la filogenia construida con la poliproteína del ORF2 (comparar con la figura 13). Se muestran los valores de BS de las ramas ≥ 70. En esta filogenia y las siguientes, la barra de escala representa 0.1 sustituciones de aminoácidos por sitio.

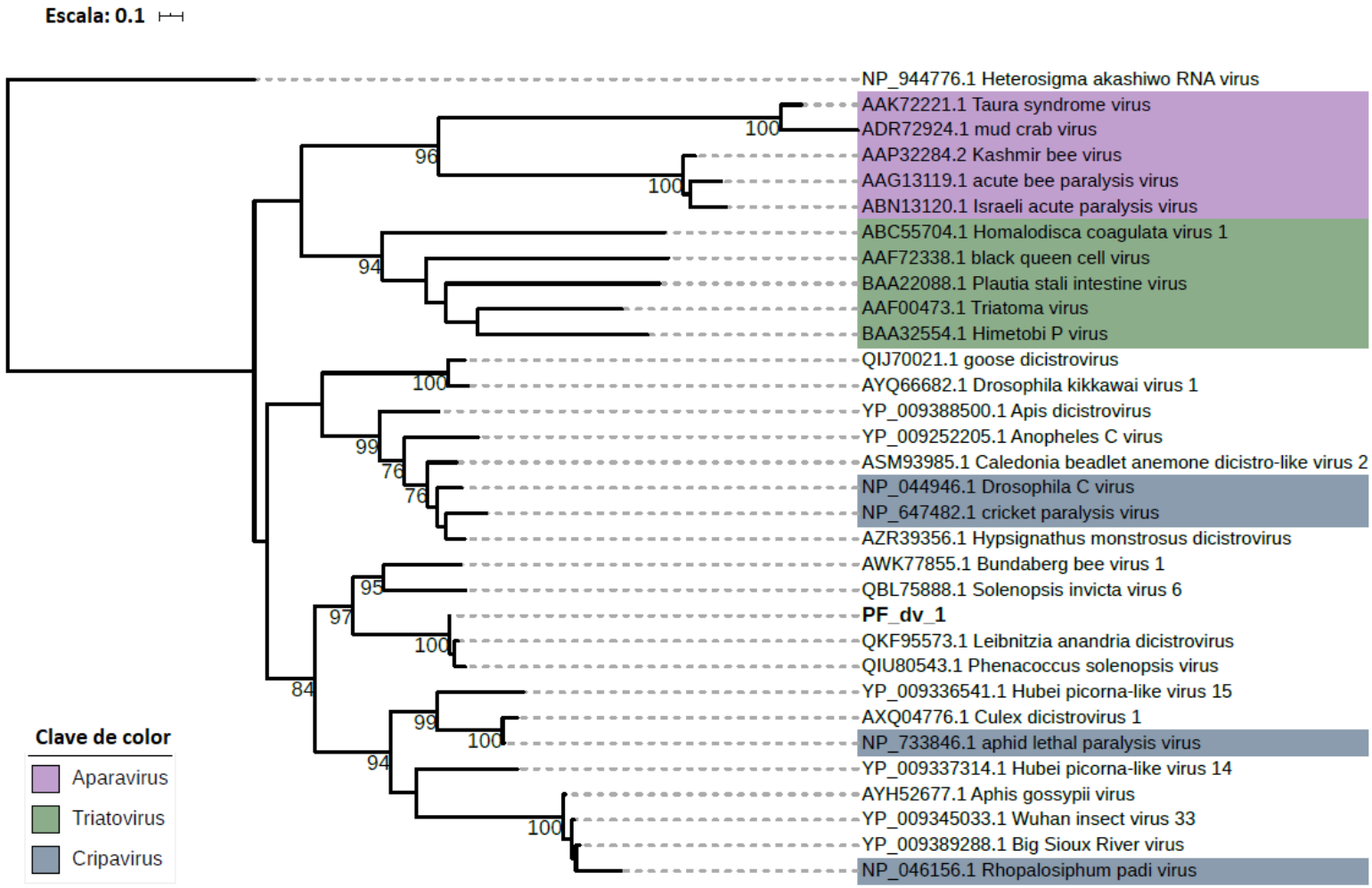

Figura 14. Árbol filogenético construido con las secuencias completas de poliproteína estructural del ORF2 de taxa de dicistroviridos, PF_dv_1 y otros taxa de virus picorna-like. Los nombres de los taxa no clasificados no se remarcan en color. Se puede observar que hay un clado para cada género de dicistroviridos, y específicamente, que los taxa de virus picorna-like incluidos se encuentran anidados en el clado que incluye a los cripavirus. Se muestran los valores de BS de las ramas ≥ 70.

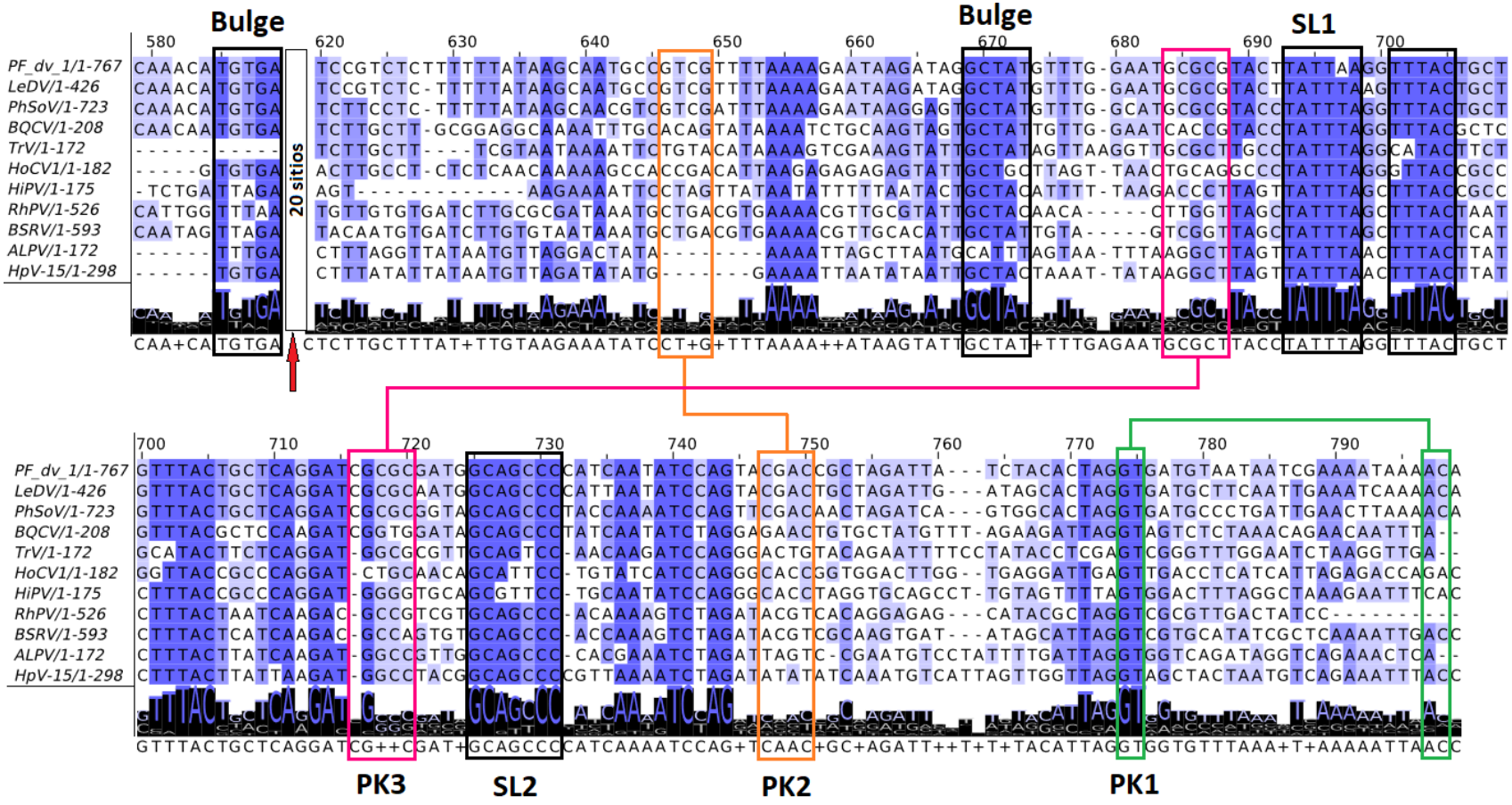

Figura 15. Alineamiento y logo de las IGR de taxa selectos de dicistroviridos, PF_dv_1 y otros taxa de virus picorna-like. Se resaltan en márgenes de color negro las secuencias clave correspondientes a los dos "bulges" y tres "stem loops" (SL1, SL2 y SL3). Se resaltan en márgenes de color verde, naranja y rosa los sitios correspondientes a los tres “pseudo-knoot" (PK1, PK2 y PK3). El logo (al pie del alineamiento) muestra la letra de cada nucleótido en un tamaño proporcional al grado de conservación de ese nucleótido para cada sitio de la secuencia. Para la identificación de los nucleótidos incluidos en las secuencias invertidas complementarias (SIC) aquí resaltadas en márgenes, se tomaron como referencia los trabajos de Jan (2006) y Luria et al. (2020). La flecha roja indica una región de 20 sitios del alineamiento que incluía gaps para la mayoría de los taxa, la cual fue recortada para mejorar la visualización del MSA. Los taxa incluidos son Leibnitzia anandria dicistrovirus (LeDV; MN723599.1), Phenacoccus solenopsis virus (PhSoV; MT176242.1), black queen cell virus (BQCV; NC_003784.1), Triatoma virus (TrV; NC_003783.1), Homalodisca coagulata virus 1 (HoCV; NC_008029.1), Himetobi P virus (HiPV; NC_003782.1), Rhopalosiphum padi virus (RhPV ; NC_001874.1), Big Sioux River virus (BSRV; GCF_002219505.1), aphid lethal paralysis virus (ALPV; GCF_000853305.1) y Hubei picorna-like virus 15 (HpV-15; GCF_001966835.1).

Los análisis de mapeo y cobertura mostraron que la biblioteca con mayor número de reads mapeados al contig PF_dv_1 fue la PV3 (**figura 16**) y por ende el valor de reads/Kb contig alcanzó también su máximo en PV3 (**figura 10**). La cobertura de la secuencia con los reads de las otras bibliotecas fue prácticamente nula.

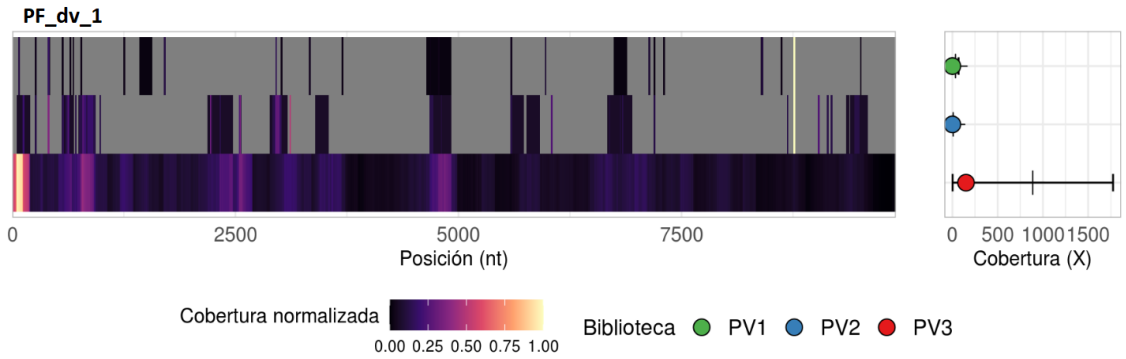

Figura 16. Heatmaps y gráfica de cobertura absoluta para el contig PF_dv_1. En el panel izquierdo, los heatmaps representan la cobertura normalizada para cada sitio nucleotídico de la secuencia de PF_dv_1 en función de un gradiente de color. Los valores 0, 0.5 y 1 del gradiente corresponden, respectivamente, a los valores mínimo (1X), intermedio (indicado como una franja vertical en el rango del min-max del diagrama de cobertura X) o máximo de la cobertura absoluta para cada sitio nucleotídico. Las ausencias de cobertura se representan en color gris. En la gráfica de cobertura en X por biblioteca (panel derecho) se muestra un rango definido por los valores máximo y mínmo de cobertura por sitio nucleotídico para cada biblioteca (segmentos de línea horizontales en negro). Se muestra también el valor de la mediana de cobertura (i.e., el valor que se encuentra en medio de los valores de cobertura de cada sitio nucleotídico cuando estos se ordenan de menor a mayor) señalado como un círculo de color para cada biblioteca. La línea vertical en cada rango min-máx indica el valor intermedio del rango.

Al graficar la cobertura absoluta en unidades X (las unidades de cobertura X hacen referencia al número de veces que un nucleótido está representado en una posición específica o cubierto por los reads mapeados) se incluyó el valor de la mediana (**figura 16**). El valor de la mediana permite definir un rango más estrecho de valores donde oscila la cobertura X para cada sitio de la secuencia; así, los valores de cobertura para el 50% de los sitios de la secuencia de PF_dv_1 se encuentran en el rango definido entre el valor de mediana y el mínimo (siempre es de 1X), mientras que los valores de cobertura para el 50% restante de los sitios de la secuencia se encuentran en el rango definido entre el valor de la mediana y el máximo

Se determinó que el contig PF_dv_1 estuvo presente sólo en la biblioteca PV3 (i.e., la biblioteca de origen; sólo con el mapeo de los reads de esta biblioteca se tiene una cobertura mínima de 1X a lo largo de una

porción contínua de la secuencia ≥70%), donde el valor de la mediana de cobertura fue de 150X y el valor máximo de 1775X. La cobertura en X unidades de PF_dv_1 con los reads de PV3 fue relativamente heterogénea ya que el 50% de los sitios no presentaron una cobertura mayor a 150X, mientras que una gran porción del 50% restante de los sitios nucleotídicos no superó un valor de ~800X (el valor internedio entre el rango definido por el mínimo y máximo de cobertura X, que corresponde a la densidad de color de 0.5 en el heatmap y a la franja vertical de la gráfica de cobetura X). De hecho, como se observa en el heatmap de PV3, hay una abundancia de la densidad de color (negro) que corresponde al mínimo (1X) de cobertura absoluta. Esta información es relevante ya que entre mayor sea la cobertura X de cada sitio, tanto mayor es la certeza en la identidad de los sitios nucleotídicos de la secuencia. Por otra parte, puede observarse que una región estrecha cercana al extremo 5' de PF_dv_1 obtuvo la mayor cobertura, que corresponde a valores cercanos a 1800X. Sin embargo, como puede observarse en la **figura 16**, la mayoría de los sitios tienen una cobertura mucho menor a dicho máximo.

3.4.2. Secuencias virales putativas relacionadas con la familia *Iflaviridae*

Los contigs PF_ifv_1, PF_ifv_2, PF_ifv_3, PF_ifv_4, PF_ifv_5 y PF_ifv_6 representan un grupo de secuencias que mostraron relación con taxa virales agrupados en el clado "*Iflaviridae-Secoviridae"* descrito por Shi et al. 2016, el cual contiene a los miembros reconocidos de dichas familias y otros picorna-like monocistrónicos afines (**figura 17**). En lo siguiente se hará referencia a los seis contigs en conjunto como "contigs ifv".

Es importante mencionar que la secuencia de los contigs PF_ifv_1, PF_ifv_3, PF_ifv_4, PF_ifv_5, representa el reverso complementario de las correspondientes secuencias recuperadas en la selección manual. Sin embargo, esto no afecta la información de anotación y coordenadas de los dominios en las secuencias de proteínas predichas, y la información de los ORFs y otras características a nivel de secuencia nucleotídica se han ajustado a las secuencias que aquí se presentan (i.e., los contigs ifv).

En cada uno de los contigs ifv se detectó un único ORF, que cubren desde un 70% (PF_ifv_1, PF_ifv_3 y PF_ifv_4) al 100% (PF_ifv_2 y PF_ifv_5) de la secuencia de los contigs (**figura 17** y **anexo C**). Las búsquedas de BLASTP, tanto con RefSeq como con NR, indicaron similitud de las proteínas de los ORFs de PF_ifv_1 (id=35%, qcov=96% y e=7E-68), PF_ifv_2 (id=35%, qcov=96% y e=7E-68) y PF_ifv_3 (id=35%, qcov=96% y e=7E-68), a diferentes regiones de la secuencia de poliproteína del ORF único de Wuhan coneheads virus

1 (WCV1); cada uno de dichos *queries* se alineo a regiones distintas de la porción estructural de dicha proteína de referencia, cubriendo las 3 regiones consecutivas y no empalmadas con dominio de cápside, que corresponden a las proteínas VP2, VP3 y VP1, respectivamente (**tabla 12** y **figura 17**).

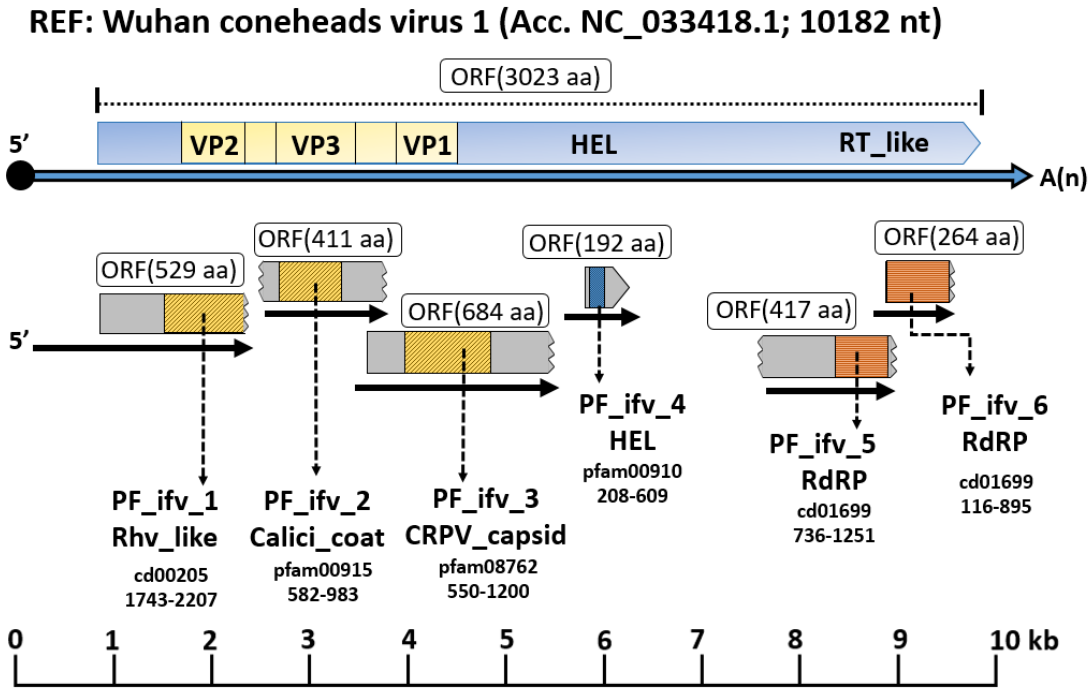

Figura 17. Organización genómica predicha para los contigs ifv. Se muestran los contigs (flechas negras sólidas), los ORFs encontrados en sus secuencias (cajas grises), así como el ID de cada dominio encontrado en las secuencias de proteína predichas para cada uno de los ORFs. Las coordenadas de los dominios indican en qué región de la secuencia nucleotídica se encuentran codificados. Los ORFs sin codón de paro o de inicio se muestran con bordes dentados. Por comparación, se incluye la organización genómica de un virus picorna-like relacionado a iflaviridos (flecha azul), su número de acceso del Genebank, la longitud del genoma completo y los ORFs con las proteínas que codifican (cajas de color sobre el genoma). A(n): cola poli-A en el extremo 3'; proteína VPg en el extremo 5' representada como un círculo negro sólido.

Así, respecto a la secuencia de animoácidos de la poliproteína de WCV1 (el mejor hit con RefSeq), las proteínas de los ORFs de PF_ifv_1, PF_ifv_2 y PF_ifv_3 se alinearon a las regiones del sitio 4 al 510, del sitio 507 al 896 y del sitio 907 al 1548, respectivamente. En correlación con esto, se detectó un dominio de cápside rhv_like (cd00205; del aa 375 al aa 529) en la proteína del ORF de PF_ifv_1, un dominio Calici_coat (pfam00915; del aa 147 al aa 327) en la proteína del ORF de PV_ifv_2 y un dominio CRPV_capsid (pfam08762; del aa 149 al aa 365) en la proteína del ORF de PF_ifv_3. Así mismo, los tres *queries* presentaron hits signifcativos con la porción estructural de la poliproteína única de otros virus picorna-like relacionados a iflaviridos como Dinocampus coccinellae paralysis virus (DCPV), Lysiphlebus fabarum rna-virus tipo A y B (LFRV-A y LFRV-B, respectivamente) o Isahaya Culex iflavirus (IsCV). Se pudo observar que en el caso de los hits con DCPV, los tres *queries* se alinean, al igual que ocurre con WCV1, a la porción

estructural de la poliproteína única del genoma. Sin embargo, DCPV solo presenta dos dominios de cápside, y se observó que las secuencias de proteínas de PF_ifv_1 y PF_ifv_2 cubren dichos dominios, siendo la región a la que se alinea la proteína de PF_ifv_3 carente de dominios estructurales.

En cuanto a las proteínas de los ORFs de PF_ifv_4, PF_ifv_5 y PF_ifv_6, se encontró similitud media (id=49-60% y qcov=96-100%) a diferentes regiones de la porción no estructural de la poliproteína de los WCV1, DCPV, FBRV-A y LFRV-B y algunos otros como Wuhan insect virus 13 (WiV13) o Hubei picorna-like virus 30 (HpV30). Así, respecto a la poliproteína codificada en el ORF de WCV1, las proteínas de los ORFs de PF_ifv_4, PF_ifv_5 y PF_ifv_6 se alinearon del sitio 1584 al 1763, del sitio 2273 al 2672 y del sitio 2655 al 2921, respectivamente. En correspondencia con lo anterior, se detectó un dominio helicasa (pfam00910; del aa 44 al aa 148, con los motivos A, B y C de las helicasas de virus de +ssRNA) en la proteína del ORF de PF_ifv_4, mientras que en las proteínas de los ORFs de PF_ifv_5 y PF_ifv_6 se detectaron dominios parciales de RdRP_1 (pfam00680); del aa 246 al aa 417 en la proteína de PF_ifv_5, con los motivos A y B del subdominio "palma αβ" (Gorbalenya et al., 2002), y del aa 12 al aa 264 en la proteína de PF_ifv_6 con los motivos A, B y C del subdominio anterior.

Durante la revisión del alineamiento pareado de las proteínas codificadas en los contigs ifv con las poliproteínas de los taxa de referencia, pudo advertirse que, en algunos casos, la región alineada del *subject* con un *query* se empalmaba ligeramente con la región de alineamiento del mismo *subject* con otro *query* (como en el caso de los alineamientos de las proteínas de PF_ifv_1 y PF_ifv_2, o de las proteínas de PF_ifv_5 y PF_ifv_6 con la poliproteína de WCV1) (**figura 17**). Para verificar si este empalme podía llevar a una unión (extensión) de los contigs correspondientes en una sola secuencia más larga (y por tanto de los ORFs y las proteínas predichas), se inspeccionó la región del empalme a nivel de nucleótidos. Sin embargo, en cada caso las secuencias nucleotídicas involucradas en el empalme presentaron elevadas disimilitudes que no permitieron la unión de los contigs. Además, considerando la información de los dominios, como en el caso de las proteínas de PF_ifv_5 y PF_ifv_6, donde los dominios de RdRP contienen los mismos motivos de RdRP, es improbable que dichos contigs sean fragmentos adyacentes de una misma proteína.

Las relaciones filogenéticas realizadas con las proteínas de los ORFs de PF_ifv_5 y PF_ifv_6 (con dominio parcial RdRP_1) sitúan a ambas secuencias con el grupo de taxa que comprende únicamente a picorna-like relacionados con iflaviridos y a taxa del género *Iflavirus* (que corresponden a los miembros de la familia *Iflaviridae*, que sólo incluye un género); esto es así ya que durante las búsquedas de BLASTP con las proteínas de PF_ifv_5 y PF_ifv_6 y usando la base de datos NR no se obtuvieron hits con algún otro grupo taxonómico (**figura 18** y **19**). Cabe recalcar que no fue posible construir una sola filogenia que

comprendiera a ambas secuencias ya que no cubren la misma región de la poliproteína de referencia (**figura 17**). Además, los taxa incluidos en cada filogenia cambian debido a que los *queries* no siempre presentaban similitudes con los mismos taxa.

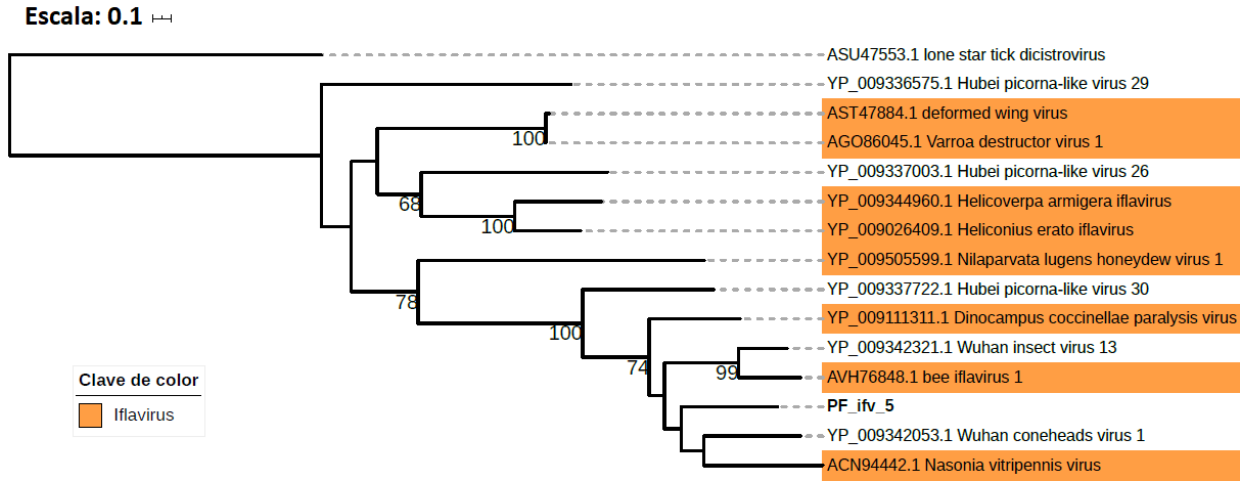

Figura 18. Árbol filogenético construido con la secuencia de proteína del ORF de PF_ifv_5 y secuencias parciales de poliproteína no estructural de taxa de virus picorna-like. Los nombres de los taxa no clasificados no se remarcan en color. Se observa como PF_ifv_5 y otros taxa picorna-like se anidan entre los taxa de iflavirus. Se muestran los valores de BS de las ramas ≥ 50.

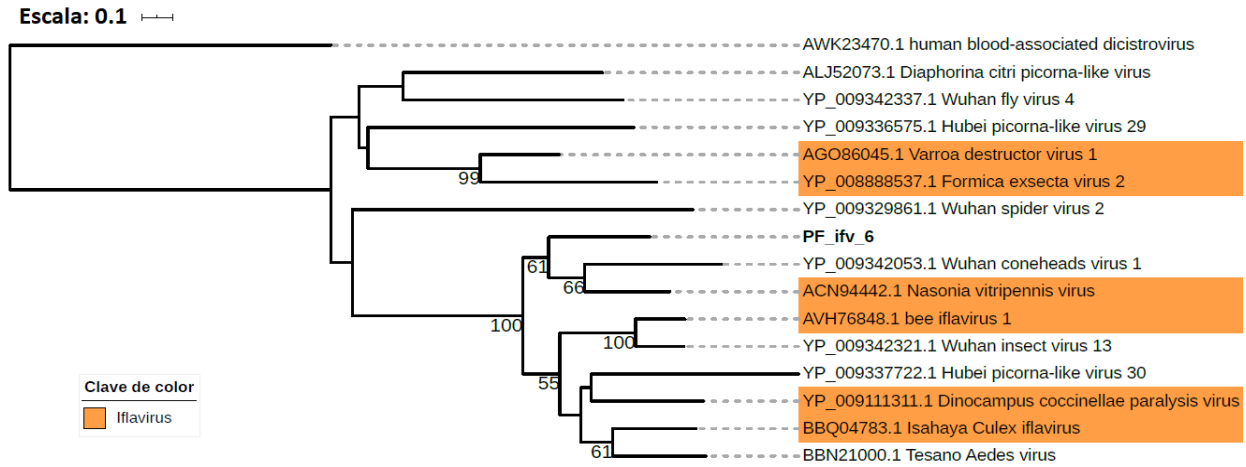

Figura 19. Árbol filogenético construido con la secuencia de proteína del ORF de PF_ifv_6 y secuencias parciales de poliproteína no estructural de taxa picorna-like relacionados. Los nombres de los taxa no clasificados no se remarcan en color. Al igual que ocurre con PF_ifv_5, se observa que PF_ifv_6 se anida entre los taxa de iflavirus. Se muestran los valores de BS de las ramas ≥ 50.

En sus filogenias correspondientes, un nodo une a PF_ifv_5 con (WCV1 + NaV) y a PF_ifv_6 con (WCV1 + NaV) (aunque con valores de BS muy disntintos, de 28 y 61, respectivamente). A su vez, en sus respectivas filogenias, el clado (PF_ifv_5 + (WCV1 + NaV)) y el clado (PF_ifv_6 + (WCV1 + NaV)) se relacionan más estrechamente con los taxa WiV13, BiV1, DCPV, IsCV, o Hubei picorna-like virus 30 (HpV30). Al considerar lo anterior, aunado al valor de similitud de las proteínas de PF_ifv_5 y Pf_ifv_6 con los taxa más estrechamente relacionados en las filogenias (id=40-60%), se sabe que la secuencia de PF_ifv_5 y PF_ifv_6 son considerablemente divergentes respecto a las secuencias de referencia.

Por su parte, la filogenia realizada con la proteína del ORF de PF_ifv_3 (con dominio CRPV_capsid y el más grande de los 3 dominios estructurales) (**figura 20**) muestra una relación estrecha de PF_ifv_3 únicamente con WCV1 (con un un valor de BS de 70), ambos distantes del resto de los picorna-like no clasificados (entre ellos WIV13 y BiV1) con un fuerte valor de soporte (BS de 100). Por su parte, los hits de BLASTP de la proteína predicha de PF_ifv_3 permitieron considerar en la filogenia a una diversidad más amplia de miembros de la familia *Iflaviridae*, aunque no a taxa de otras familias.

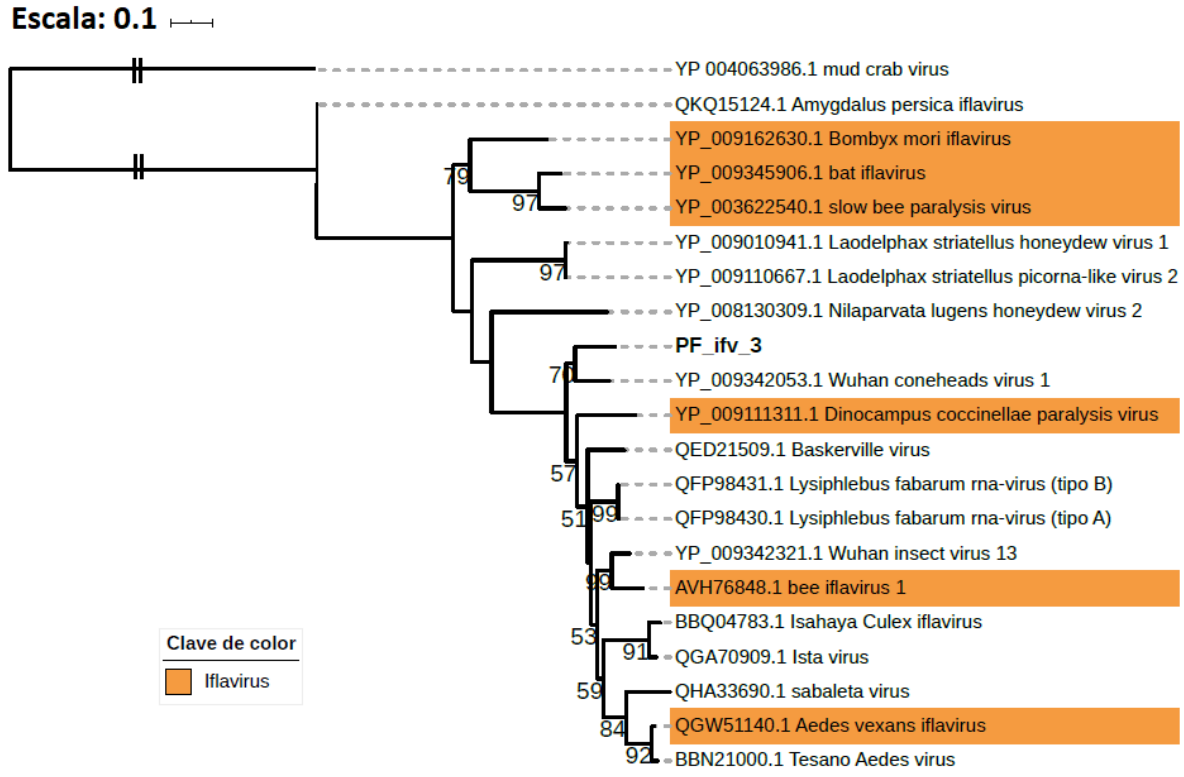

Figura 20. Árbol filogenético construido con la secuencia de proteína del ORF de PF_ifv_3 y secuencias parciales de poliproteína estructural de virus picorna-like. Los nombres de los taxa no clasificados no se remarcan en color. Se muestran los valores de BS de las ramas ≥ 50. Las ramas marcadas con (ll) se recortaron a la mitad de su tamaño.

En conjunto, los análisis de similitud y de relaciones filogenéticas indican que los contigs ifv representan secuencias virales putativas que divergen considerablemente de los genomas de taxa virales relacionados a iflaviridos previamente reportados, y por tanto pueden representar a uno o más virus de RNA novedosos. Sin embargo, debe recalcarse que sólo la porción de los contigs que es codificante (la porción de secuencia que abarcan los ORFs) pudo ser evaluada con los análisis de similitud a nivel de aminoácidos.

En cuanto a los resultados del mapeo, se determinó que los contigs ifv estuvieron presentes en una o más bibliotecas, siendo mayor la cobertura X para los seis contigs con los reads de PV1 (la biblioteca de origen), donde los máximos de cobertura oscilan entre 50X y 30X, excepto para el contig PF_ifv_6, donde el máximo es apenas de 20X (**figura 21**). Esta variación entre máximos es sustancialmente menor que la observada para otros contigs recuperados (ver secciones siguientes). Por otra parte, los valores de la mediana de cobertura nunca superan los 20X, y solo para los contigs PF_ifv_1 y PF_ifv_6 coincide aproximadamente con el valor intermedio definido por el rango min-máx. Lo anterior puede observarse en los heatmaps como una porción notablemente mayor de los sitios con cobertura por debajo del valor intermedio de cobertura (0.5).

Respecto a las bibliotecas secuendarias, se determinó que los contigs PF_ifv_2 a PF_ifv_5 también se encuentran representados en la biblioteca PV3. Por otro lado, sólo el contig PF_ifv_4 se consideró representado en las tres bibliotecas. Como puede observarse en los heatmaps de las bibliotecas PV3 y PV2 (**figura 21**), el mapeo de sus reads a los contigs representados en cada una no cubren la totalidad de los contig (se observan regiones en gris con ausencia de reads mapeados). Lo anterior es relevante ya que sugiere que los reads de estas bibliotecas, mapeados a los contigs ifv, pudieran haber reconstruido contigs similares en secuencia a los contigs ifv, pero más pequeños, y que no fueron considerados durante la selección manual. Por otra parte, debe notarse que en algunos casos, se observa ausencia de reads mapeados a pequeñas fracciones de la secuencias de los contigs, incluso en el mapeo de los reads de las bibliotecas de origen (por ejemplo, ver heatmap de PF_ifv_2 en PV1). Esto puede deberse a que los reads inicialmente mapeados a dichas porciones sin cobertura de los contigs fueron descartados durante el filtrado por calidad del mapeo (aunque dichos reads se usaran en el ensamble). En relación con lo anterior, debe considerarse la cobertura del contig PF_ifv_1 con los reads de PV3, que no fue en una porción contínua ≥70% de la secuencia; puede observarse que con el mapeo de los reads se cubren dos porciones de PF_ifv_1 (una de ~600 nt y otra de ~1200 nt), separadas entre si por una pequeña porción sin mapeo. Lo anterior puede deberse a que los reads mapeados a dicha porción sin cobertura no pasaron el filtrado por calidad del mapeo. En tal caso, los reads de PV3 puedieron haber ensamblado una secuencia (que contiene a las dos porciones de secuencia cubiertas mencionadas de PF_ifv_1 y separadas por la región

sin cobertura) que no fue seleccionada. Por lo tanto, se consideró que PF_ifv_1 también está representado en PV3.

Adicionalmente, se pudo determinar que hay heterogeneidad en la distribución de la cobertura de acuerdo a los heatmaps de las bibliotecas PV1 y PV3 de los contigs representados en cada una (**figura 21**). Además, los máximos de cobertura X en la biblioteca PV3 oscilan entre 10X y 35X, y en cada caso, son similares o menores a los máximos de la cobertura X con los reads de PV1. Cabe mencionar que la cobertura de los contigs ifv es ≤200 reads/Kb de PV1 (**figura 10**), lo cual es equiparable a la cobertura observada para lor contigs rv, rbv tbv, y algunos contigs pv (ver secciones siguientes).

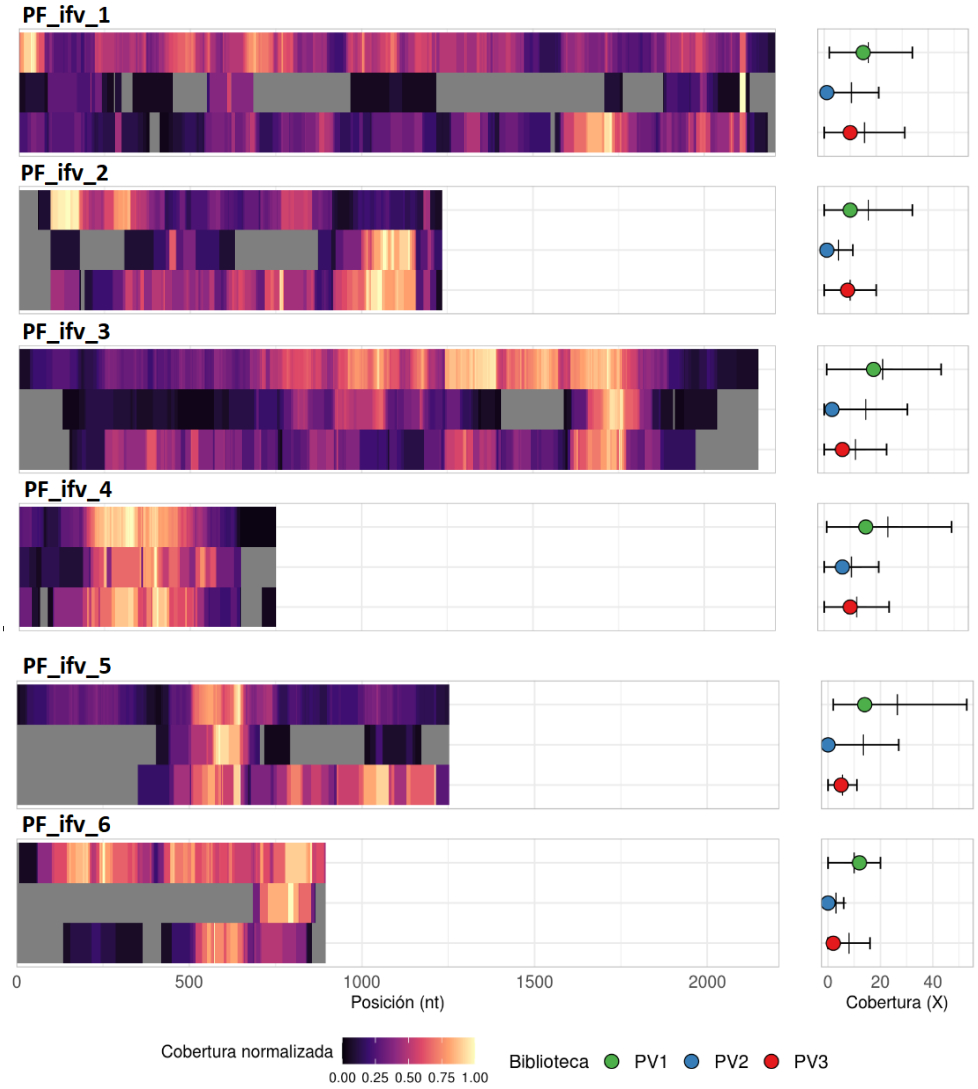

Figura 21. Heatmaps y gráficas de cobertura absoluta para los contigs ifv. La descripción de esta figura es la misma dada para la figura 16.

3.4.3. Secuencias virales putativas relacionadas con *Picornavirales* no clasificados

Los contigs PF_pv_1, PF_pv_2, PF_pv_3, PF_pv_4 y PF_pv_5 representan un grupo de secuencias cuyas proteínas predichas mostraron similitud significativa con un grupo de taxa picorna-like especialmente divergentes, cuya máxima resolución taxonómica permite relacionarlos únicamente con el orden *Picornavirales* (**figura 22**). Dichos taxa no clasificados se agrupan en el clado "*Picorna-Calici"* descrito por Shi et al. (2016), el cual contiene a los miembros oficiales del orden *Picornavirales* (las familias *Dicistroviridae*, *Iflaviridae*, *Secoviridae*, *Picornaviridae*, *Marnaviridae* y *Calicivirdae*) y otros picorna-like que no pueden agruparse, o se apartan en la topología del árbol, de los clados que contienen a los miembros de cada familia y sus relacionados (i.e., son aún más divergentes que los picorna-like relacionados con alguna familia). En lo siguiente se hace referencia al conjunto de los cinco contigs como "contigs pv".

La secuencia de de los contigs PF_pv_1, PF_pv_3, PF_pv_4 y PF_pv_5 representa el reverso complementario de las correspondientes secuencias recuperadas en la selección manual. Al igual que en el caso de los contigs ifv, la información de los ORFs y otras características a nivel de secuencia nucleotídica se han ajustado a las secuencias que aquí se presentan.

Cuatro de los 5 contigs pv presentaron un único ORF; solo PF_pv_2 presentó dos ORFs. En general, los ORFs cubren una longitud variable de los contigs correspondientes (60-99%; **figura 22** y **anexo C**). Al igual que las proteínas predichas de los ORFs que mostraron hit con picorna-like relacionados a iflaviridos, las proteínas hipotéticas codificadas por los contigs pv cubrieron únicamente regiones parciales de las poliproteínas de sus hits. Los resultados de BLASTP, tanto con RefSeq como con NR, muestran a cherry virus Trakiya (CVT), Hubei picorna-like virus 51 (HpV51), maize associated virus (MaPV), Aphis glycines virus 1 (ApGV1) y Tetranychus urticae-associated picorna-like virus (TaPV1) como los hits con mayor identidad (**tabla 12**). Las proteínas del ORF único de PF_pv_1 y del ORF1 de PF_pv_2 se alinearon con la poliproteína estructural (ORF1), mientras que las proteínas de los ORFs de PF_pv_3, PF_pv_4 y PF_pv_5 se alinearon con la poliproteína no estructural (ORF2), de los taxa de referencia mencionados (**figura 22**). La proteína del ORF2 de PF_pv_2 no presentó similitud con ninguna secuencia de las bases de datos.

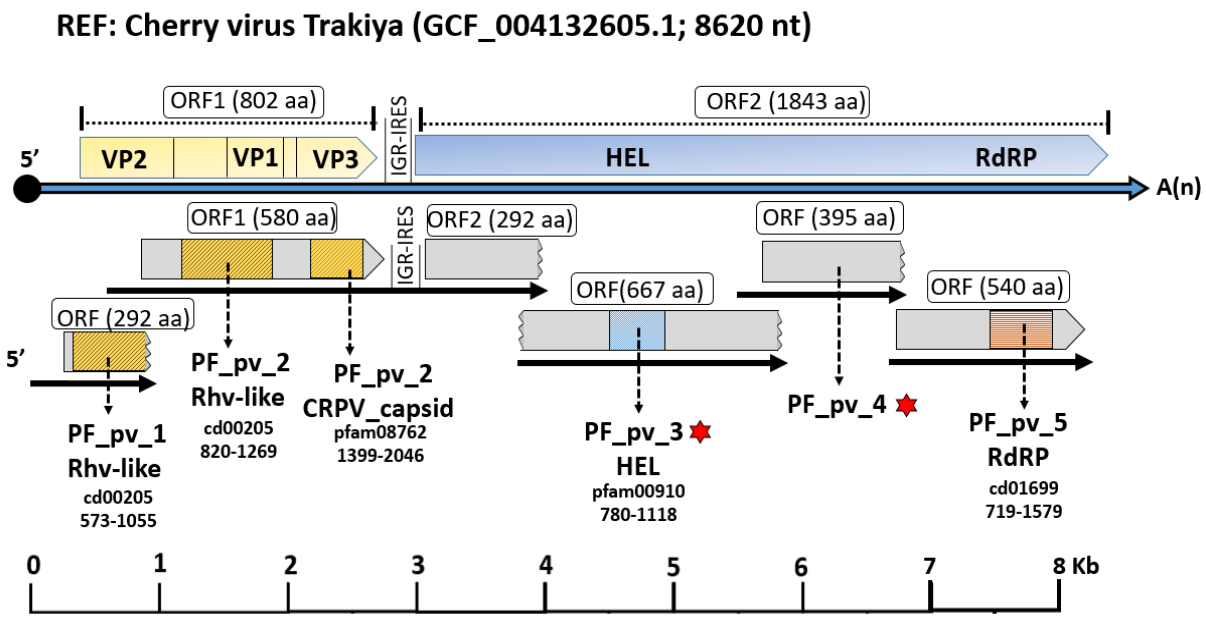

Figura 22. Organización genómica predicha para los contigs pv. Se muestran los contigs (con flechas negras solidas) y sus ORFs (cajas grises sobre los contigs), así como el ID de cada dominio encontrado en las secuencias de proteína predichas para cada uno de los ORFs. Las coordenadas de los dominios indican en qué región de la secuencia nucleotídica se encuentran codificados. Los ORFs sin codón de paro o de inicio se muestran con bordes dentados. Con asterisco rojo se marcan aquellas secuencias cuyos hits de BLASTP tuvieron un valor de qcov <90%. Por comparación, se incluye la organización genómica de un virus picorna-like (flecha azul), su número de acceso del Genebank, la longitud del genoma completo y los ORFs con las proteínas que codifican (cajas de color sobre el genoma). A(n): cola poli-A en el extremo 3'; proteína VPg en el extremo 5' representada como un círculo negro sólido.

Específicamente, las secuencias de proteínas del ORF de PF_pv_1 y del ORF1 de PF_pv_2 se alinearon, respectivamente, con la región del sitio 1 al 222 y la región del sitio 237 al 796 de la secuencia de referencia (proteína estructural de CVT), cubriendo así los 3 dominios de cápside del *subject*. En correspondencia con esto, se detectó un dominio rhv_like (cd00205; aa 53-213) en la proteína del ORF de PF_pv_1, así como un dominio rhv_like (cd00205; aa 163-312) y uno dominio CRPV_capsid (pfam08762; aa 414-571) en la proteína del ORF1 de PF_pv_2. Dichos dominios estructurales codificados en ambos contigs corresponden a dominios de cápside JRC. Respecto a los contigs PF_pv_3, PF_pv_4 y PF_pv_5, los alineamientos con el ORF2 de CVT (el mejor hit de RefSeq, excepto para PF_pv_4) abarcan las regiones del sitio 313 al 685, del 947 al 1257, y del 1252 al 1763, respectivamente. Estas regiones de la referencia contienen dominios de helicasa (HEL) y de polimerasa (RdRP_1). En correspondencia, la secuencia de los ORFs de PF_pv_3 y PF_pv_5 se detectaron, respectivamente, dominios de helicasa (pfam00910; aa 260-372) y RdRP_1 (pfam00680; aa 70-530). Por su parte, en la proteína del ORF de Pf_pv_4 se detectó un dominio de peptidasa C3G en dos regiones (pfam12381; del aa 174 al aa 279 y del aa 295 al aa 375), sin embargo, no

se mencionan en la tabla 12 ya que su valor de e=0.31 y e=0.17, para la primera y segunda región, respectivamente, quedaron por arriba del valor de reporte especificado en el análisis.

El contig PF_pv_2 es de particular interés, ya que en su secuencia se encontraron dos ORFs, separados por una NCR. Si bien el producto predicho del ORF2 de PF_pv_2 no presentó homología con ninguna secuencia o presencia de dominios, el alineamiento de la proteína predicha del ORF1 con la poliproteína estructural de CVT (la porción más proximal al extremo 5') reveló que la IGR del primero coincide aproximadamente en posición con la IGR del segundo.

Considerando que los picorna-like no clasificados como ApGlV1 presentan un IRES tipo 1 en la IGR similar al de los dicistroviridos, se inspeccionó el alineamiento de una porción de la secuencia de la IGR y el ORF2 de PF_pv_2 con la IGR de ApGlV1, CVT y otros dicistroviridos (**figura 23**). Exceptuando a PF_pv_2, en el MSA se incluyeron las secuencias de la IGR de los taxa incluidos (desde el primer nucleótido después del término del ORF1 hasta el nucleótido antes del inicio del ORF2). En el caso de PF_pv_2 la secuencia incluida corresponde a la región comprendida desde el nt 2073 (el nucleótido después del término del ORF1) hasta el nt 2610 (85 nt después del inicio del ORF2). Al examinar el MSA se pudieron encontrar las secuencias conservadas correspondientes a dos estructuras "bulge" y dos de "stem loop", así como las SIC de tres de "pseudo-knot", que conforman una IRES tipo 1 putativa en la región de secuencia considerada en el MSA de PF_pv_2 (**figura 23**). Cabe mencionar que selección de dicha región fue la que permitió encontrar todas las SIC, puesto que se realizaron otros MSA donde se consideraban diferentes regiones del contig PF_pv_2 (secuencia completa o la región IGR sin incluir porciones de secuencia de los ORFs) donde sólo se reconocían algunas de las SIC y otras secuencias conservadas.

Las relaciones filogenéticas modeladas con la proteína del ORF de PF_pv_2 (con el dominio JRC de CRPV_capsid) muestran dos clados principales, separados desde la base del árbol, cada uno con altos valores de soporte (>80) (**figura 24**). El primer clado agrupa a todos los taxa de picorna-like no clasificados, relacionados con *Picornavirales*, que incluye a PF_pv_2, mientras que el segundo clado contiene a los taxa incluidos de dicistroviridos.

Por su parte, la filogenia resuelta con la proteína del ORF de PF_pv_5 (con el dominio de RdRP_1) muestra una separación en clados distintos de los taxa incluidos de dicistroviridos, secoviridos, iflaviridos y relacionados, y picorna-like no clasificados (aunque con valores de soporte moderados, de entre 50 y 80) (**figura 25**). Es importante notar que en el caso de la filogenia de PF_pv_5, las secuencias que se pudieron incluir representan un espectro más amplio de grupos taxonómicos que los que pudieron incluirse en la

filogenia de PF_pv_2. Además, en sus respectivas filogenias, el clado que agrupa a los secoviridos está más estrechamente relacionado al clado de los picorna-like no clasificados (donde se incluye a PF_pv_5), mientras que en la filogenia de PF_pv_2 el clado más cercano a los picorna-like no clasificados es el de dicistroviridos. Sin embargo, el valor de BS que une el clado de PF_pv_5 con el clado de secoviridos presenta un valor de BS < 70. Por su parte, el valor de soporte de la rama que une el clado de PF_pv_2 con el clado de dicistroviridos no se conoce dado que es la misma que lleva al grupo externo (al hacer esto, el programa RAxML no calcula valor de BS, puesto que el usuario ha establecido esa polaridad por defecto).

En síntesis, de acuerdo con los análisis de similitud y de relaciones filogenéticas, se puede hipotetizar que los contigs pv representan secuencias genómicas parciales de uno o más virus picorna-like, que divergen de las secuencias genómicas de otros taxa picorna-like previamente reportados. Por otra parte, cabe resaltar que al igual que ocurre con los contigs ifv, sólo la fracción codificante de los contigs pv pudo ser evaluada con los anáisis de similitud.

En los resultados del mapeo se observó que los contigs pv se encontraron representados únicamente en la biblioteca PV1 (la biblioteca de origen), mientras que muy pocos reads de las otras bibliotecas mapearon, i.e., no hay bibliotecas secundarias para estos contigs (**figura 26**). Como puede observarse en los heatmaps, únicamente con el mapeo de los reads de PV1 se obtuvo una cobertura total de las secuencias, mientras que en el mapeo de los reads de las bibliotecas restantes la cobertura se concentró únicamente en regiones estrechas y dispersas, de donde se sigue que los contigs pv pudieron ser generados únicamente a partir de los reads de la biblioteca PV1.

Respecto a dicha biblioteca, se puede observar que para algunas secuencias como PF_pv_2 y PF_pv_5, la cobertura es mayormente heterogénea, mientras que para algunas como PF_pv_1, PF_pv_3 y PF_pv_4, la cobertura normalizada se observa mayormente por arriba del intermedio del rango min-máx de cobertura. Por otra parte, puede notarse que los máximos de cobertura varían hasta por un orden de magnitud. PF_pv_1 presenta un máximo de hasta 175X, con una mediana ligeramente por arriba de 100X (que es el valor intermedio del rango min-máx para este contig), mientras que otros como PF_pv_3, PF_pv_4 y PF_pv_5 presentaron máximos de entre 25X y 50X, con medianas de ~25X. Por su parte el contig PF_pv_2, si bien presenta un máximo mayor al de PF_pv_1 (> 175X), su mediana es de 50X, lo cual se corresponde en el heatmap con abundancia de sitios por debajo de la densidad intermedia de color (ver **figura 10**). Cabe mencionar que los contigs PF_pv_1 y PF_pv_2 presentan una cobertura de ~800 y ~400 reads/Kb de PV1, lo cual es superior a la cobertura (de ~200 reads/Kb de PV1) para el resto de los contigs pv y los contigs ifv.

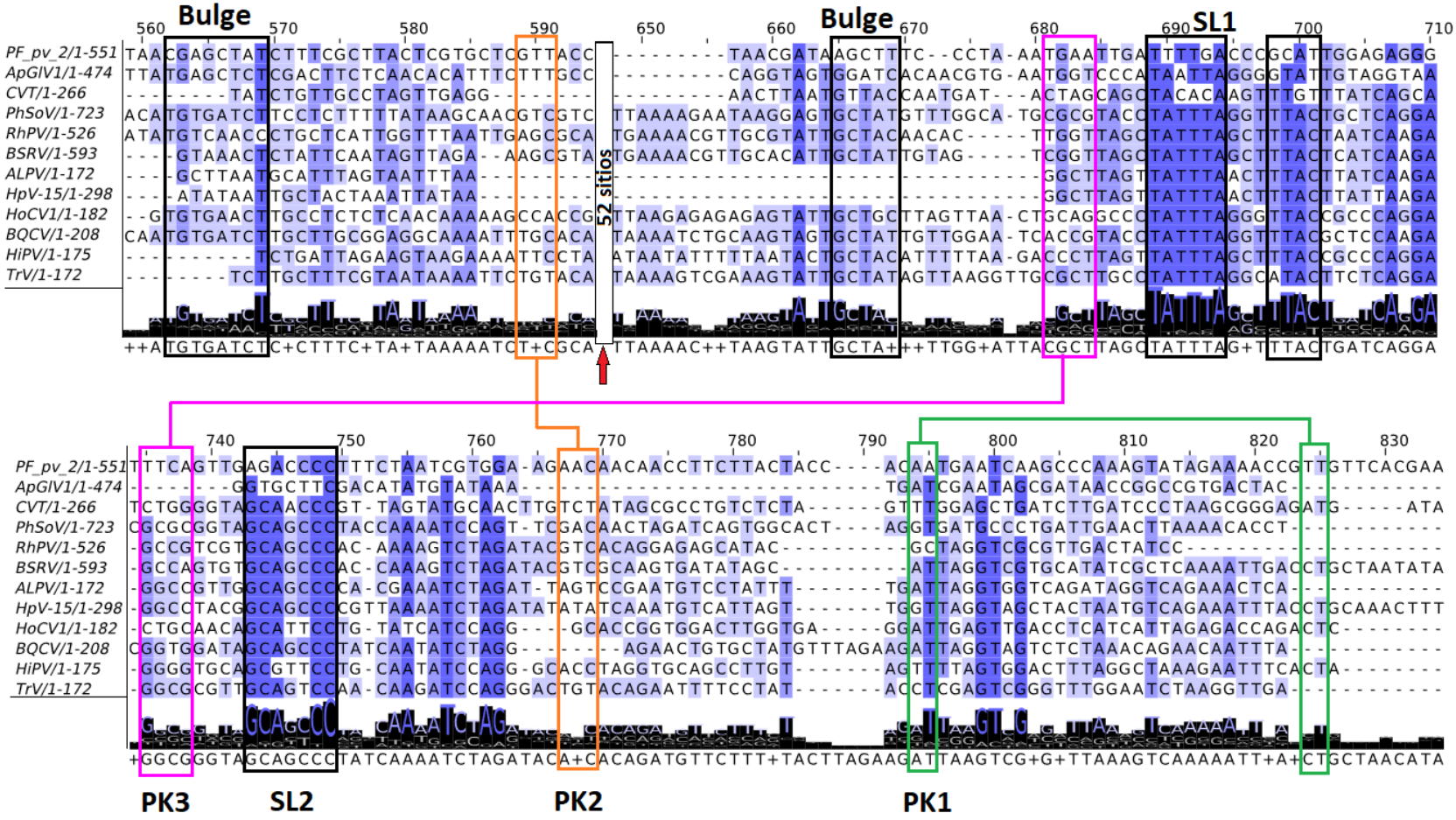

Figura 23. Alineamiento y logo de la IGR de PF_pv_2 y taxa selectos del orden *Picornavirales*. Se resaltan en márgenes de color negro las secuencias clave correspondientes a los dos "bulges" y tres "stem loops" (SL1, SL2 y SL3). Se resaltan en márgenes de color verde, naranja y rosa las SIC correspondientes a los tres "pseudo-knoot" (PK1, PK2 y PK3). El logo muestra el grado de conservación de las SIC y otras secuencias conservadas en las IRES tipo 1. Para la identificación de las SIC y otras secuencias clave de la IRES se tomaron como referencia los trabajos de Jan (2006) y de Luiria et al., (2020). La flecha roja indica una región de 52 sitios del alineamiento que incluía gaps para la mayoría de los taxa, y fue recortada para mejorar la visualización del MSA. Los taxa inluidos son los mencionados en la figura 17, además de cherry virus Trakiya (CVT; GCF_004132605.1) y aphis glycines virus 1 (ApGlV1; KM015260.2).

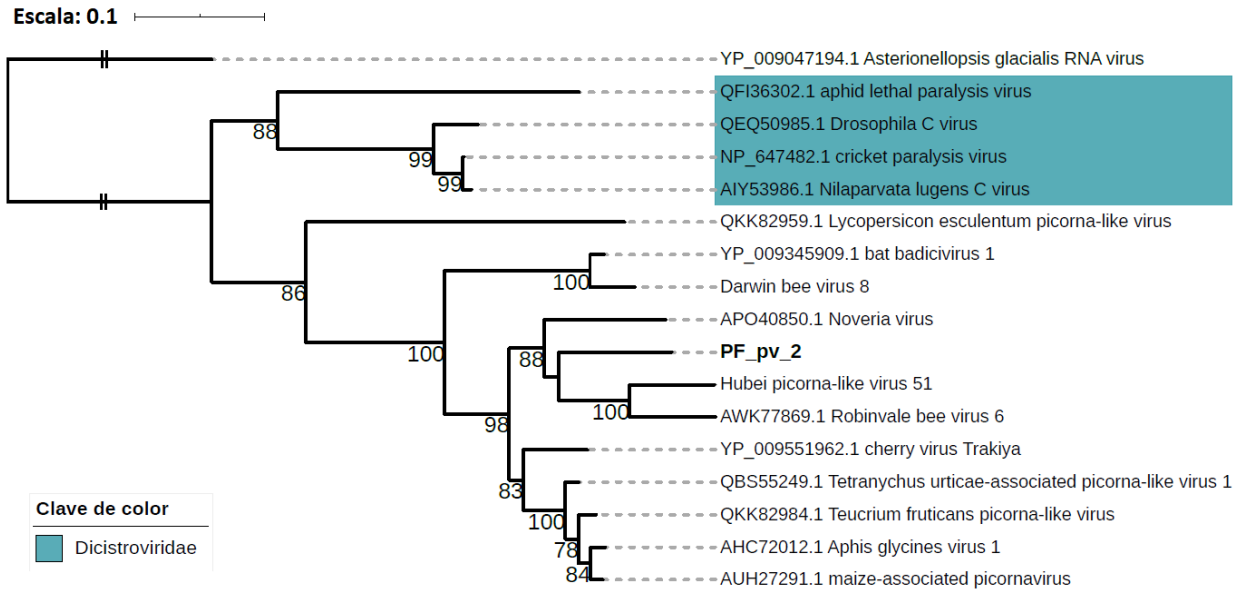

Figura 24. Árbol filogenético construido con la secuencia de proteína del ORF1 de PF_pv_2 y secuencias parciales de poliproteína estructural de virus picorna-like. Los taxa picorna-like no clasificados no se remarcan en color. Se muestran los valores de BS de las ramas ≥ 70. Las ramas marcadas con (ll) se recortaron a la mitad de su tamaño. El grupo externo empleado es un miembro de a familia *Marnaviridae*.

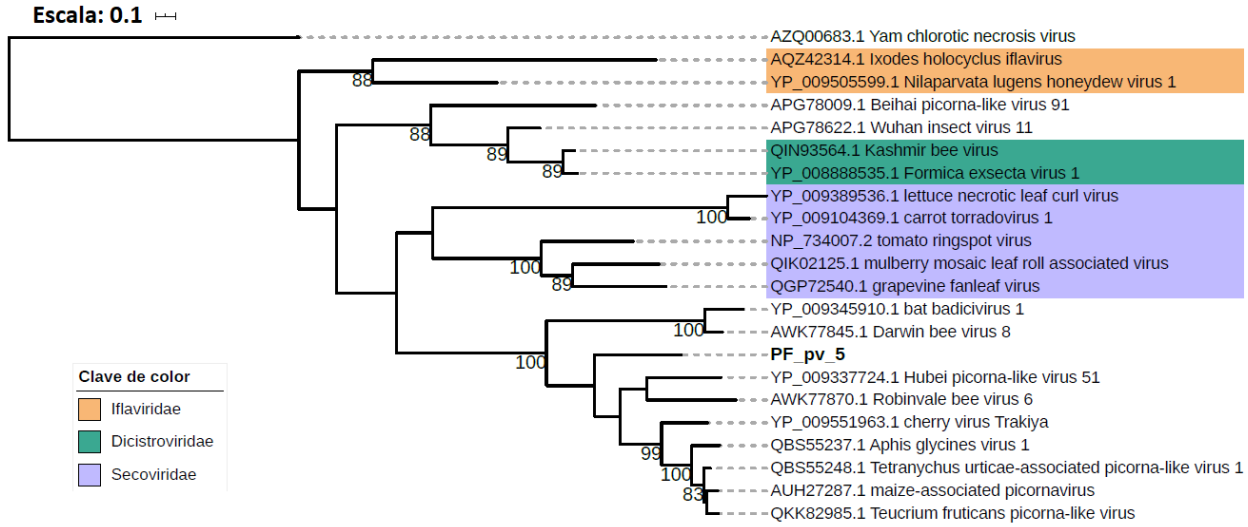

Figura 25. Árbol filogenético construido con la secuencia de proteína del ORF de PF_pv_5 y secuencias parciales de poliproteína no estructural de virus picorna-like. Los nombres de los taxa no clasificados no se remarcan en color. Se muestran los valores de BS de las ramas ≥ 70.

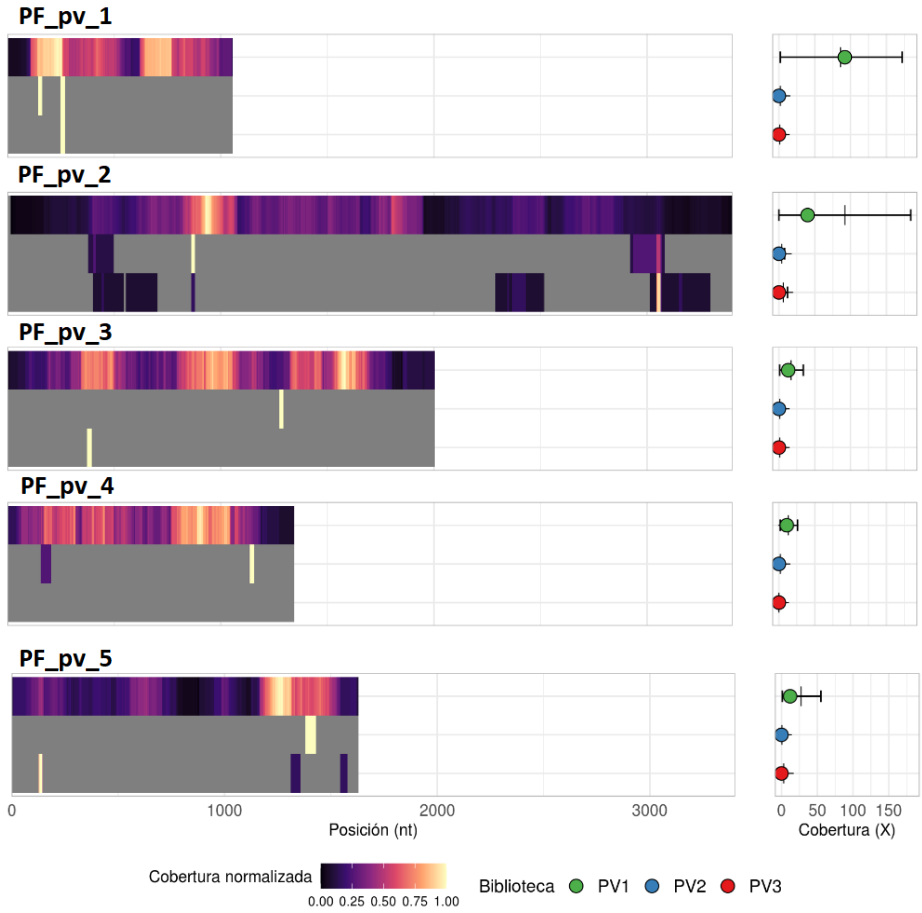

Figura 26. Heatmaps y gráficas de cobertura absoluta para los contigs pv. La descripción de esta figura es la misma dada para la figura 16.

3.4.4. Secuencias virales putativas relacionadas con la familia *Reoviridae*

Los contigs PF_rv_1 a PF_rv_8 representan un grupo de secuencias con homología tentativa a diferentes segmentos del genoma multipartita de virus de dsRNA relacionados a la familia *Reovididae*, algunos de los cuales se agrupan dentro del clado "Reo" descrito por Shi et al. (2016), que contiene a los miembros reconocidos de la familia *Reoviridae* y otros taxa relacionados. Cada uno de los 8 contigs contiene un solo ORF, que en todos los casos, abarca una fracción ≥80% del total de su secuencia (**figura 27** y **anexo C**). En lo siguiente se hará referencia a los ocho contigs en conjunto como "contigs rv". La secuencia de los contigs PF_rv_2 es el reverso complementario de secuencia recuperada en la selección manual. Al igual que en el caso de los contigs anteriores, se ajusta la información de los ORFs y otras características de nucleótidos a las secuencias que aquí se presentan.

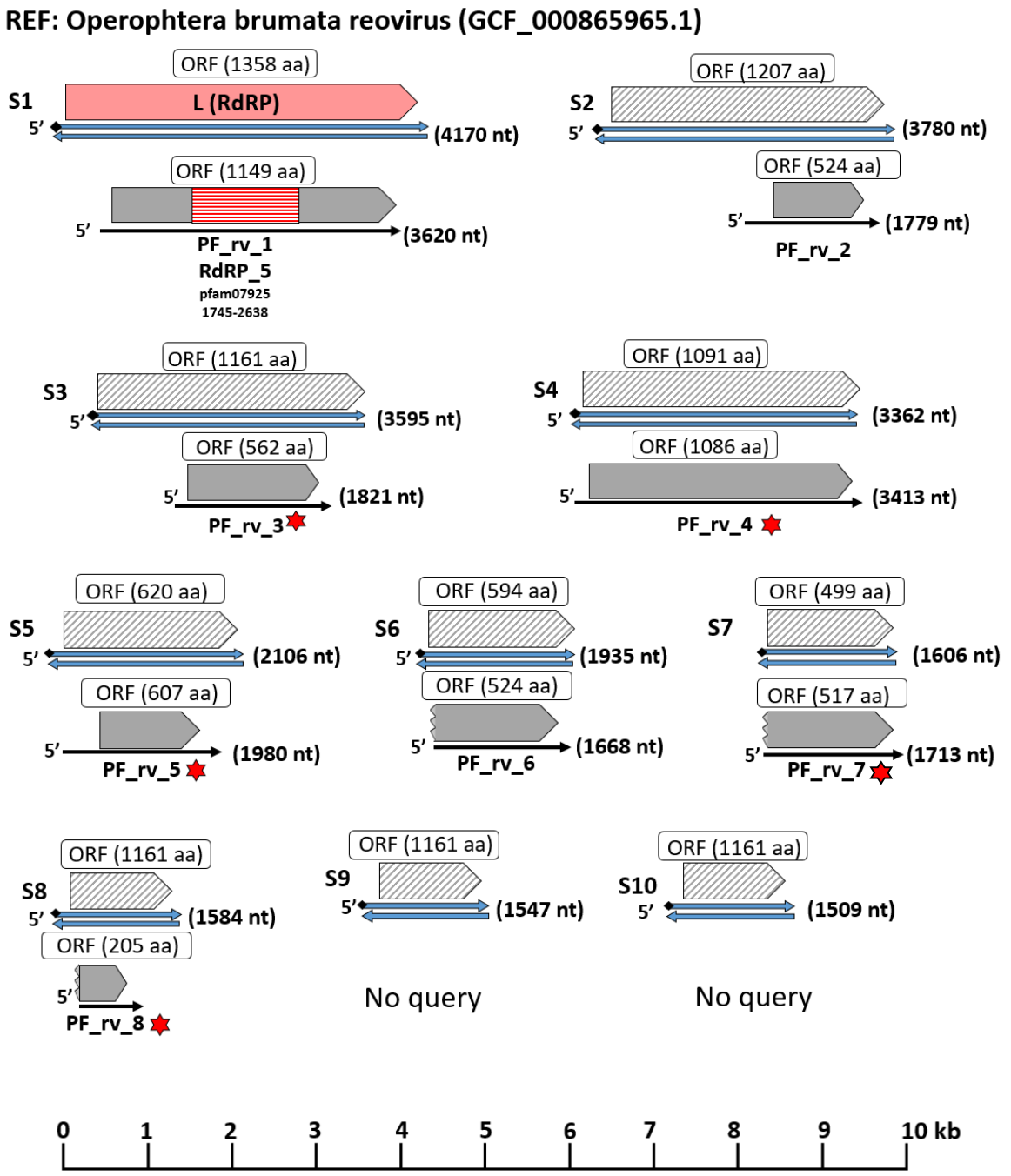

Figura 27. Organización genómica predicha para los contigs rv. Se muestran los contigs (flechas negras sólidas) y sus ORFs (cajas grises/coloreadas sobre los contigs), así como el el ID de cada dominio encontrado en las secuencias de proteína predichas para cada uno de los ORFs. Las coordenadas de los dominios indican en qué región de la secuencia nucleotídica se encuentran codificados. Los ORFs sin codón de inicio se muestran con bordes dentados. Con asterisco rojo se marcan aquellas secuencias cuyos hits de BLASTP tuvieron un valor de qcov <90%. Por comparación, se muestran los segmentos (con su longitud en nt) de un genoma de referencia de reovirus (flechas azules) y su número de acceso en el Genebank. Se muestran los ORFs con las proteínas que codifican (cajas sobre los segmentos). Las proteínas sin dominios o función predicha del genoma de referencia se muestran en cajas con franjas grises en diagonal. Los extremos cap-5′ se representan con diamantes negros.

Inicialmente, las búsquedas de BLASTP vs RefSeq indicaron que cada uno de las proteínas de los ocho ORFs obtuvo una similitud limitada con las proteínas codificadas en 8 de los 10 segmentos monocistrónicos del genoma multipartita de Operophtera brumata reovirus (OpBRV), un miembro propuesto, aunque no oficialmente aceptado, del género *Idnoreovirus* (Graham et al., 2008) (**tabla 12** y **figura 27**). Únicamente las proteínas de los ORFs de PF_rv_1, PF_rv_2, PF_rv_4, PF_rv_6 y PF_rv_7 tienen una longitud que es ≥80% de la longitud de secuencia de su respectivo mejor hit con la base de datos RefSeq. Respecto a los valores del alineamiento, los hits de los ocho *queries* con las proteínas de OpBRV contrastan ampliamente en la cobertura del *query* (desde 35 hasta 98%) y el porcentaje de identidad (desde 19% hasta 37%), siendo los hits de las proteínas de PF_rv_3, PF_rv_4, PF_rv_5, PF_rv_7 y PF_rv_8 los que obtuvieron un a cobertura del *query* < 90%.

Por su parte, los análisis de BLASTP con NR mostraron hits para 7 de las 8 proteínas *queries* con las proteínas de taxa virales no clasificados, relacionados a reoviridos, como Shelly beach virus (SRV), Hubei diptera virus 21 (HDV21), bat reovirus (BatRV) y Eccles virus (ECV), cuyos valores de cobertura e identidad son ligeramente mayores (en un máximo de 10%) a los obtenidos con RefSeq (**tabla 12**). Sólo en el caso del ORF del contig PF_rv_7 se obtuvo un único hit con una proteína hipotética de *Tubulinosema ratisbonensis*, con mejores valores de alineamiento (qcov=55%, id=25% y e=4E-17) que los obtenidos con la proteína del segmento S7 de OpRV (qcov=35%, id=20% y e=4E-4).

Respecto a la anotación de las secuencias, únicamente la proteína del ORF de PF_rv_1 presentó un dominio de RdRP_5 (pfam07925; del aa 546 al aa 843), con los motivos A, B y C del subdominio “palma αβ” de RdRP (Gorbalenya et al., 2002). En correspondencia, dicha proteína presentó similitud con las RdRP de OpBRV y otros taxa relacionados a reoviridos (**tabla 12**). En el ORF de PF_rv_4 se encontró un dominio TBCC_N de chaperona C específico de tubulina (pfam16752; del aa 678 al aa 716) y obtuvo hit con la proteína del segmento 4 de OpBRV, cuya anotación en la base de datos NR es “polihedrina”. Sin embargo, dicho *subject* es descrito por Graham et al. (2008) únicamente como proteína hipotética. Los productos proteicos de los ORFs de los 6 contigs restantes sólo obtuvieron hit con proteínas del genoma de OpRV sin función conocida (i.e., hipotéticas), y no se encontraron dominios conservados en sus secuencias.

La filogenia reconstruida con las secuencias de RdRP de taxa de reoviridos y otros relacionados (**figura 28**) muestran una agrupación de PF_rv_1 con SBV, ECV, BatRV y HDV21 en un clado (con soporte BS=100) separado del resto de los taxa miembros de la familia *Reoviridae*, los cuales se observan separados en clados que definen a los diferentes géneros. Como cabe esperarse, algunos taxa no clasificados se relacionan más estrechamente con los taxa de algún género particular, y por tanto se observan anidados

dentro de sus respectivos clados. Sin embargo, se observó una estrecha relación entre el único miembro incluido del género *Idnoreovirus* (OpBRV) con el clado que contiene a PF_rv_1 (clado "no clasificados"). En nodos más basales, el clado (Idnoreovirus + "no clasificados") se une con el género *Fijivirus*, aunque con soporte bajo (<70). A su vez, el siguiente nodo más basal une a los anteriores con el grupo (*Mycovirus* + *Coltivirus*), pero igualmente con valores bajos de soporte. Lo anterior es relevante debido a que si bien PF_rv_1 está en un clado filogenéticamente distante de taxa como los *Cypovirus* o *Dinovernavirus* (cuyos miembros infectan únicamente a insectos), está mas estrechamente relacionado con el género *Idnoreovirus* (cuyos miembros, aunque pocos, se sabe que infectan a insectos) (Attoui et al., 2011).

En consistencia con lo observado en la filogenia, la matriz de identidad pareada reconstruida para las secuencias de RdRP (las mismas de la filogenia) muestra valores de identidad variable (desde menos de 30% hasta un 86%) entre los taxa considerados (**figura 29**). Los valores de identidad pareada entre la proteína del ORF de PF_rv_1 y proteínas correspondientes de OpBRV, ECV, HDV21, BatRV o SRV son de entre 33-39%, siendo OpBRV el taxa más distante. Por su parte, ECV, HDV21, BatRV y SRV muestran identidades más elevadas entre sí (51-61%) que con PF_rv_1 o con OpBRV. Por otra parte, es notable que la identidad entre los taxa del clado que incluye a PF_rv_1 y el resto de los taxa considerados en la filogenia es <30% a nivel de aminoácidos, lo cual es una similitud de secuencias muy limitada, reportada previamente entre taxa de diferentes géneros de reoviridos o taxa relacionados (Attoui et al., 2011).

En conjunto, los análisis de similitud indican que los contigs rv presentan una similitud variable (de 23 a 42%) con sus correspondientes secuencias de referencia, que en algunos casos es comparable con la similitud que se observa entre proteínas de reoviridos de géneros distintos (Attoui et al., 2011). Además, la filogenia con secuencias de RdRP indica que PF_rv_1 diverge de otros taxa no clasificados, relacionados con reoviridos. Por lo tanto, se puede hipotetizar que los contigs rv representan segmentos de uno o más reoviridos putativos, distintos a los segmentos correspondientes de otros reoviridos previamente reportados.

Los resultados del mapeo de reads mostraron que los contigs rv están representados en la biblioteca PV1 (la biblioteca de origen), y en el caso de los contigs PF_rv_3, PF_rv_6 y PF_rv_7, también se encuentran representados en la biblioteca PV3 (**figura 30**). En contraste, muy pocos de los reads de la biblioteca PV2 mapearon a los contigs rv. Los máximos de cobertura X con los reads de PV1 son de entre 20X y 60X y los valores de la mediana ocilan entre 10X y 20X. Así mismo, en la mayoría de los casos, los valores de la mediana coincidieron aproximadamente con el valor intermedio del rango min-máx. de cobertura X, lo que se observa en el heatmap como porciones iguales de sitios con cobertura por arriba y por debajo de

la densidad de color intermedia (0.5). Respecto a los máximos de cobertura observados para los contigs PF_rv_3, PF_rv_6 y PF_rv_7 con los reads de PV3, puede observarse que son aproximadamente la mitad de los máximos de cobertura correspondientes a dichos contigs con los reads de PV1. Por otra parte, cabe destacar que los valores de cobertura en reads/Kb obtenidos para los contigs rv son comparables con los valores obtenidos por los contigs ifv y tres contigs pv (exceptuando a PF_pv_1 y PF_pv_2), pero no con los valores del contig PF_dv_1 (**figura 10**).

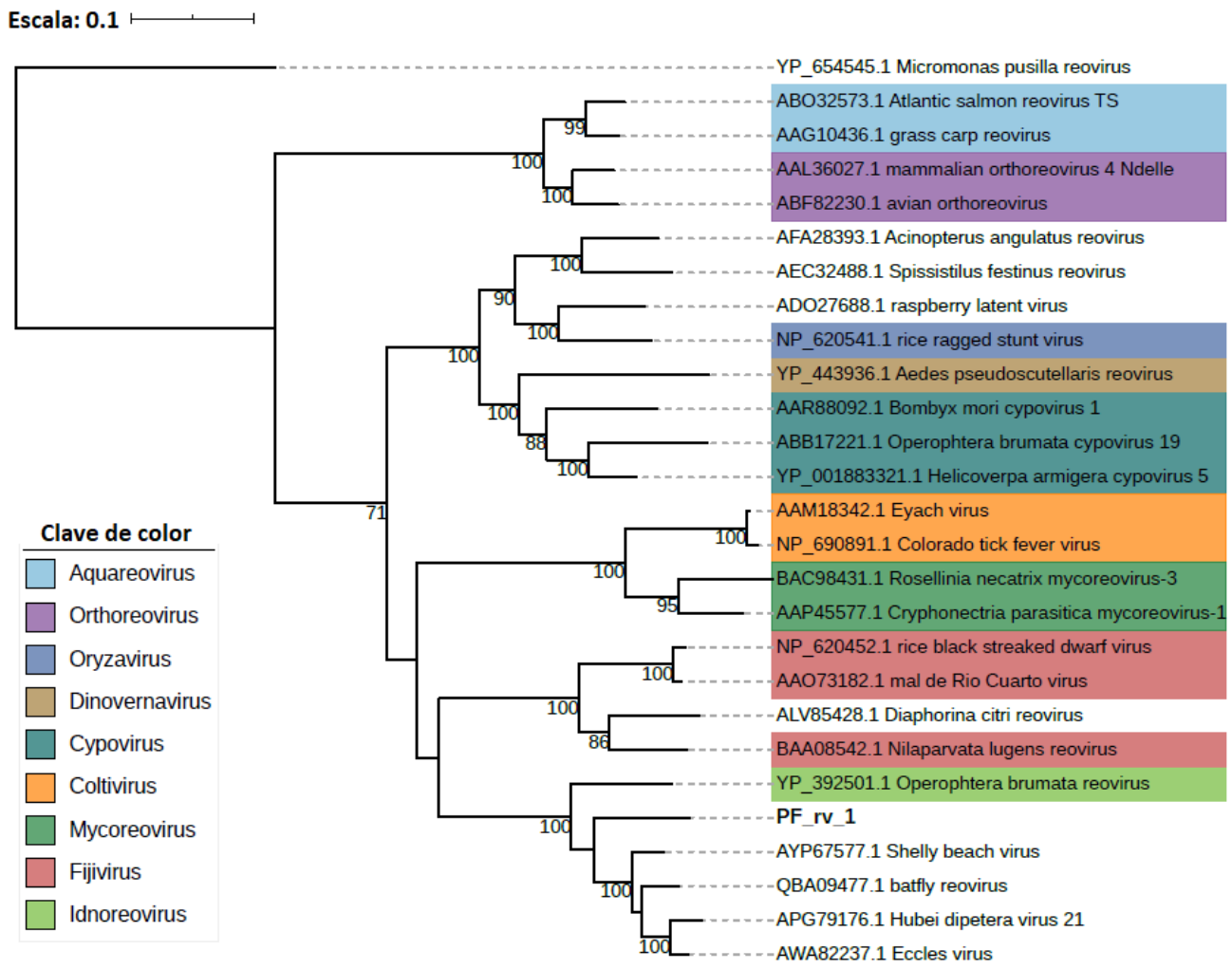

Figura 28. Árbol filogenético construido con la secuencia de proteína del ORF de PF_rv_1 y secuencias parciales de RdRP de reoviridos (subfamilia *Spinareovirinae*) y otros taxa de virus relacionados. Los taxa de virus no clasificados no se remarcan en color. Se muestran los valores de BS ≥ 70 para cada rama.

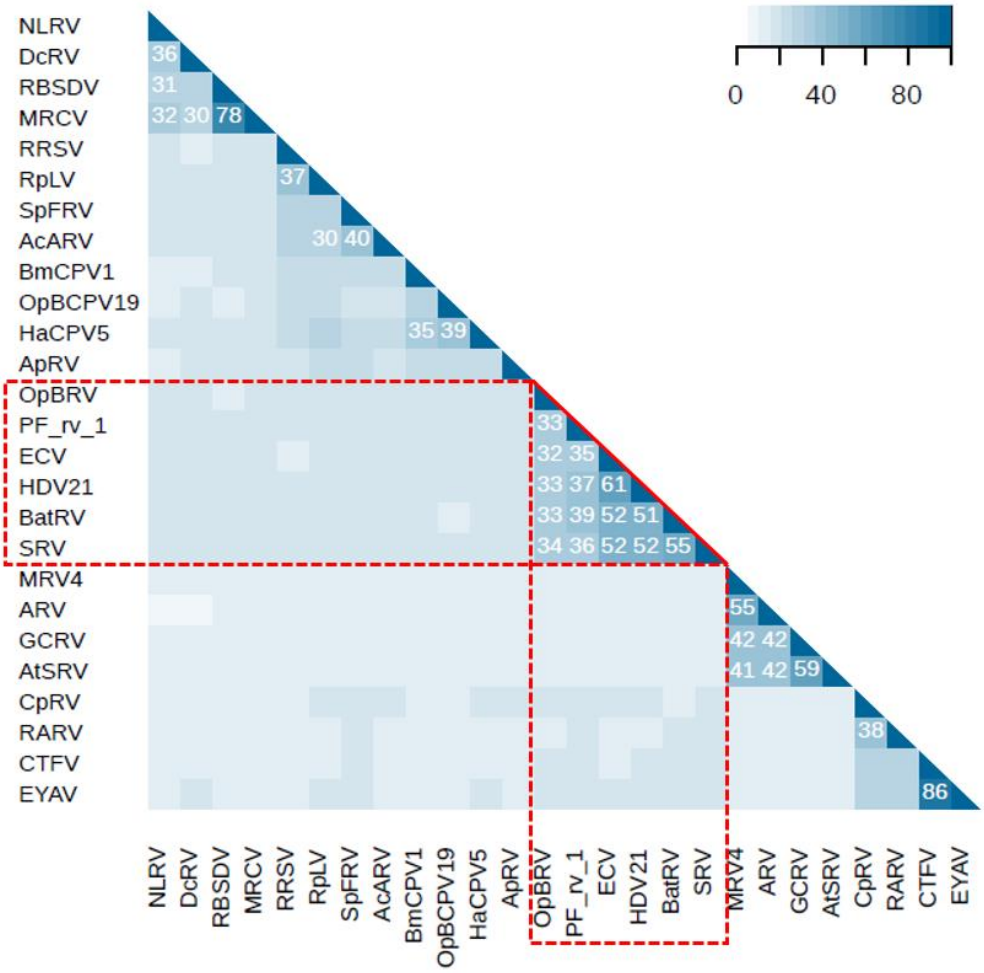

Figura 29. Matriz de porcentaje de identidad pareada de la proteína del ORF de PF_rv_1 y secuencias parciales de RdRP de reoviridos y otros taxa relacionados. Los valores de cada celda están correlacionados con la escala de color correspondiente (se muestran solo los valores ≥ 30%). En recuadro rojo se remarca las filas y columnas de los taxa con mayor identidad a PF_rv_2. Los taxa incluidos son Nilaparvata lugens reovirus (NLRV; BAA08542.1), Diaphorina citri reovirus (DcRV; ALV85428.1), rice black streaked dwarf virus (RBSDV; NP_620452.1), mal de Rio Cuarto virus (MRCV; AO73182.1), rice ragged stunt virus (RRSV; NP620541.1), raspberry latent virus (RpLV; ADO27688.1), Spissistilus festinus reovirus (SpFRV; AEC32488.1), Acinopterus angulatus reovirus (AcARV; AFA28393.1), Bombyx mori cypovirus 1 (BmCPV1; AAR88092.1), Operophtera brumata cypovirus 19 (OpBCPV19; ABB17221.1), Helicoverpa armigera cypovirus 5 (HaCPV5; YP_001883321.1), Aedes pseudoscutellaris reovirus (ApRV; YP_443936.1), Operophtera brumata reovirus (OpBRV; YP_392501.1), Eccles virus (ECV; AWA82237.1), Hubei diptera virus 21 (HDV21; APG79176.1), batfly reovirus (BatRV; QBA09477.1), Shelly beach virus (SRV; AYP67577.1), mammalian orthoreovirus 4 Ndelle (MRV4; AAL36027.1) avian orthoreovirus (ARV; ABF82230.1) grass carp reovirus (GCRV; AAG10436.1), Atlantic salmon reovirus TS (AtSRV; ABO32573.1), Cryphonectria parasitica mycoreovirus-1 (CpRV; AAP45577.1), Rosellinia necatrix mycoreovirus-3 (RARV; BAC98431.1), Colorado tick fever virus (CTFV; NP_690891.1) y Eyach virus (EYAV; AAM18342.1).

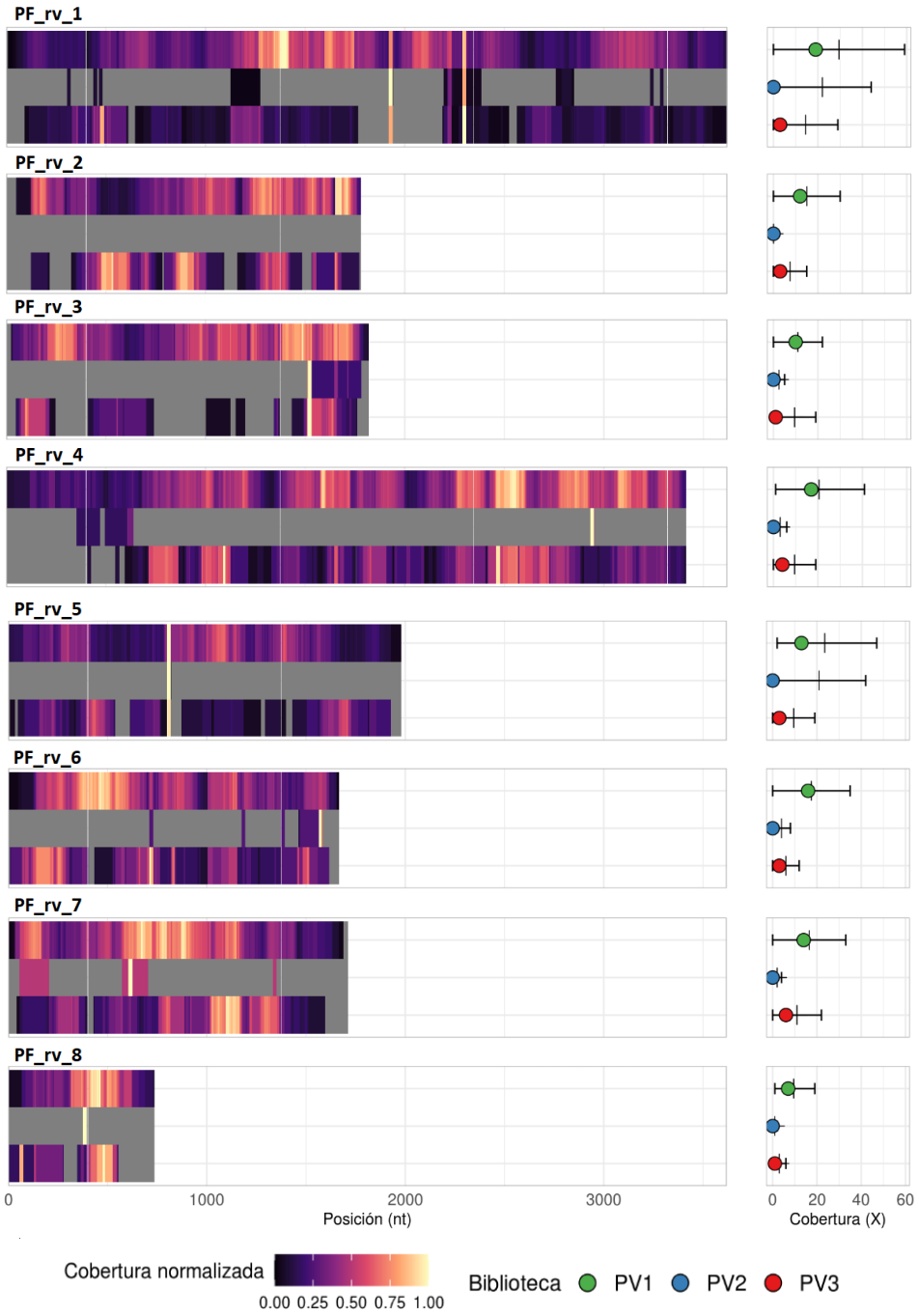

Figura 30. Heatmaps y gráficas de cobertura absoluta para los contigs rv. La descripción de esta figura es la misma dada para la figura 16.

3.4.5. Secuencias virales putativas relacionadas a la familia *Rhabdoviridae*

Los contigs PF_rbv_1, PF_rbv_2, PF_rbv_3 y PF_rbv_4 son un grupo de secuencias que presentaron homología con los miembros del supergrupo dimarhabdovirus, descrito y propuesto por Li et al. 2015

(actualmente no reconocido oficialmente por el ICTV), el cual contiene a su vez a algunos de los géneros reconocidos de la familia *Rhabdoviridae* y otros taxa relacionados. La organización genómica de los rhabdoviridos consisite en 5 ORFs canónicos, codificados en un genoma monopartita de -ssRNA, que producen, en dirección 5' → 3', proteínas de nucleocápside (N), fosfoproteína (P), proteína de matriz (M), glicoproteína (G) y proteína larga no estructural (L) (**figura 31**). En lo siguiente se hace referencia a los cuatro contigs en conjunto como "contigs rbv". La secuencia de los contigs PF_rbv_1 y PF_rbv_2 es el reverso complementario de las secuencias recuperadas en la selección manual, por lo que se han ajustado las características de los ORFs y otros datos de secuencia nucleotídica a las secuencias aquí presentadas.

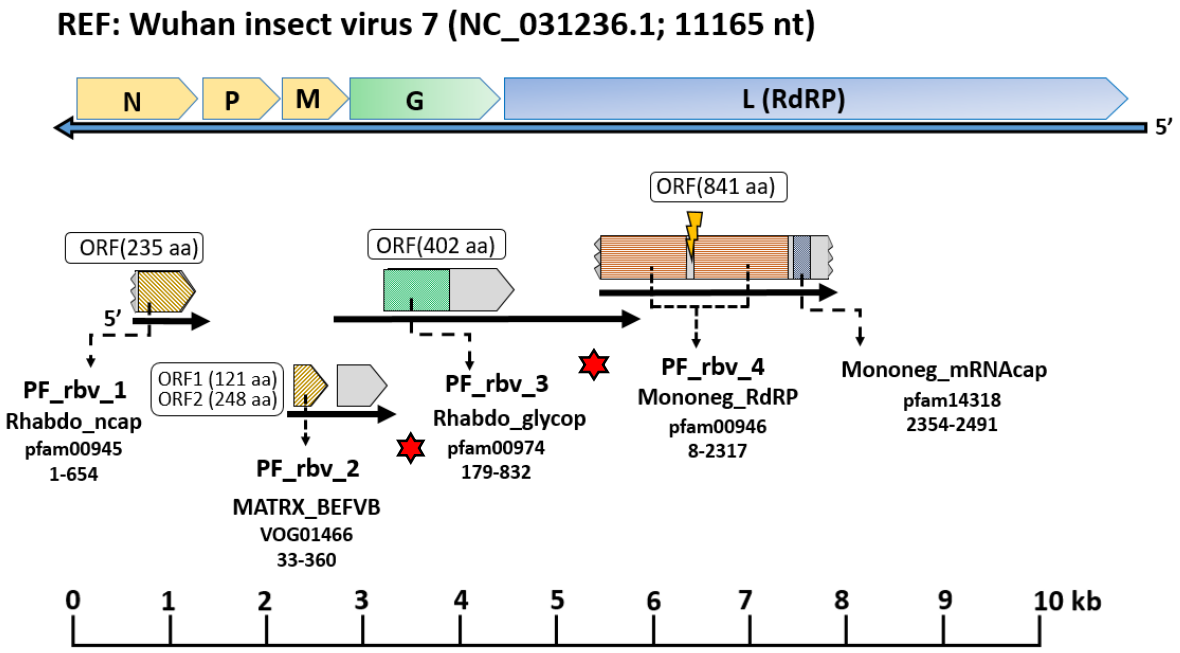

Figura 31. Organización genómica predicha para los contigs rbv. Se muestran los contigs (flechas negras sólidas), ORFs encontrados en cada secuencia (cajas grises/coloreadas sobre los contigs), así como el ID de cada dominio encontrado en las secuencias de proteína predichas para los ORFs. Las coordenadas de los dominios indican en qué región de la secuencia nucleotídica se encuentran codificados. Con asterisco rojo se marcan aquellas secuencias cuyos hits deBLASTP tuvieron un valor de qcov <90%. La interrupción del ORF en el contig PF_rbv_4 se representa con un rayo amarillo. Por comparación, se incluye la organización genómica de un virus del supergrupo dimarhabdovirus (flecha azul), su número de acceso en el Genebank, la longitud del genoma completo y las proteínas que codifica.

En los contigs PF_rbv_1 y PF_rbv_3 se detectó un único ORF que abarca, respectivamente, un 69% y 50% de la secuencia nucleotídica (**figura 31** y **anexo C**). En el contig PF_rbv_2 se encontraron dos ORFs que en conjunto abarcan un 92% de dicho contig. Por su parte, el contig PF_rbv_4 fue la única secuencia que se recuperó de un ensamble preliminar (utilizando el programa SPAdes v. 3.13.1; Nurk et al., 2013). Esta

secuencia nucleotídica contiene 10 posiciones indeterminadas (N) en la región comprendida entre los sitios 1001 al 1011. A su vez, se detectaron dos ORFs en la secuencia del contig PF_rbv_4 (que abarcan la totalidad del contig), separados por una región que consta de un codón de paro (TAG), seguido por las 10 posiciones N. Lo anterior sugiere que hay un ORF en PF_rbv_4 que está interrumpido, posiblemente por falta de información en la secuencia del contig que se completó con 10 Ns (aunque esta región indeterminada no necesariamente corresponde a 10 bases), debido a que el contig PF_rbv_4 es producto del proceso de *scaffolding* del ensamblador SPAdes, en el cual dos contigs se unieron en sus extremos adyacentes completando las bases ambiguas como Ns. Por lo tanto, se consideró a dichos ORFs iniciales (inmediatamente adyacentes) como uno sólo ORF, asumiendo que la interrupción del mismo ocurre por la presencia de bases indeterminadas en su secuencia nucleotídica, y por tanto, el producto proteíco (con posiciones indeterminadas X en su secuencia de aminoácidos, correspondientes a las posiciones N de su ORF) de ORF se sometió a los análisis de similitud (**figura 31**).

Los análisis de BLASTP, tanto con RefSeq como con NR, de las proteínas predichas en los ORFs considerados presentaron homología tentativa, con identidad limitada, con las proteínas N, M, G y L de los genomas de Hubei dimarhabdovirus 2 (HdRV2), Tetrastichus brontispae RNA virus 1 (TBRV1) y Wuhan insect virus 7 (WiV7) (todos miembros del supergrupo dimarhabdovirus). En la selección manual de secuencias no se encontró ninguna secuencia con homología a fosfoproteína, que mostrara afiliación a la proteína P de rhabdoviridos. La proteína del ORF de PF_rbv_1 (de 235 aa de longitud) obtuvo el mejor hit con la proteína N de HdRV2 (qcov=99%, id=51% y e =1E-74), alineándose con la región del *subject* que va del sitio 187 al 420, cubriendo así aprox. el 50% de su secuencia. La proteína del ORF de PF_rbv_2 (de 121 aa de longitud), obtuvo hit con la proteína M de TBRV1 (qcov=56%, id=37% y e=8E-6), alineándose con la región que va del sitio 60 al 127 de dicho *subject*, cubriendo aprox. el 80% de la secuencia. La proteína del ORF de PF_rbv_3 (de 402 aa de longitud) obtuvo como mejor hit a la proteína G de WiV7 (qcov=85%, id=38% y e =6E-83), alineándose en la región que va del sitio 132 al 470 de dicho *subject*. Por último, la proteína del ORF de PF_rbv_4 (de 841 aa de longitud) obtuvo como mejor hit a la proteína L de HdRV2 (qcov=99%, id=60% y e=0), alineándose en la región que va del sitio 287 al 1138 de dicho *subject*, cubriendo aprox. el 40% de la secuencia de dicho *subject* (**figura 31** y **tabla 12**). Así pues, considerando la similitud de los productos proteícos de los contigs rbv con sus respectivas secuencias proteícas de referencia, se puede considerar en principio que los contigs rbv corresponden a secuencias que divergen sustancialmente respecto a las secuencias de taxa relacionados con rabdoviridos. Sin embargo, se debe considerar que en el caso del contig PF_rbv_3 sólo una fracción de su secuencia (50%) es codificante, y en el caso del contig PF_rbv_2 sólo el producto del ORF1 (que abarca el 30% del contig) tuvo hit con alguna secuencia de las bases de datos. Por lo tanto, una porción sustancial de la secuencia de dichos contigs es indeterminada.

Respecto a la anotación de las proteínas hipotéticas (que representan secuencias parciales según sus relaciones taxonómicas), se encontraron dominios congruentes a los hits de BLASTP, excepto en la proteína de PF_rbv_2 con hit a proteína M, donde no se encontró ningún dominio. Sin embargo, dicho *query* obtuvo hit con el grupo de genes ortólogos MATRX_BEFVB (VOG01466; del aa 1 al aa 120), formado a partir se secuencias de proteínas M de virus de -ssRNA. En la proteína del ORF de PF_rbv_1 se encontró un dominio de nucleocápside Rhabdo_ncap (pfam00945; del aa 1 al aa 218), así como un dominio de glicoproteína Rhabdo_glyco en la proteína del ORF de PF_rbv_3 (pfam00974; del aa 13 al aa 214).

En la proteína del ORF de PF_rbv_4 se encontraron dominios parciales de polimerasa Mononeg-RNApol (RdRP) (cl15638; del aa 3 al aa 772) y de mRNA-capping (cl67796; del aa 785 al aa 830). Al inspeccionar el alineamiento entre dicho *query* y la secuencia consenso del dominio RdRP se observó que en la vecindad del alineamiento (de aproximadamente 50 aa río arriba) en la región correspondiente a las posiciones N del ORF de PF_rbv_4 se encuentran abundantes gaps. Lo anterior sugiere que las bases indeterminadas en el contig PF_rbv_4 corresponden a una región de más de 10 bases nucleotídicas.

La filogenia reconstruida con las secuencias parciales de RdRP de rhabdoviridos y la proteína putativa del ORF de PF_rbv_4 (**figura 32**) muestra una relación más estrecha de este último con HdRV2, WiV7 y TBRV1 en un clado (con valor de BS=100) filogenéticamente distante del resto de los clados que agrupan a los miembros de los distintos géneros. Sin embargo, la topología más basal (i.e., que une a los géneros) no siempre obtuvo valores de soporte elevados (< 70). La filogenia se reconstruyó incluyendo a miembros de varios géneros de rhabdoviridos para inspeccionar la relación entre PF_rbv_4 con los diferentes taxa no clasificados en el contexto de la familia *Rhabdoviridae*. A pesar del agrupamiento en un mismo clado, es posible que PF_rbv_4, HdRV2, WiV7 y TBRV1 sean filogenéticamente distantes (en el contexto de la familia), puesto que la longitud de las ramas que los unen es relativamente larga en comparación con las ramas de algunos clados que integran a los miembros de un género (como los taxa de sprivivirus, perhabdovirus, curiovirus o ephemerovirus). Sin embargo, otros géneros presentan ramas de longitud comparable (como tibrovirus o caligrhavirus).

En consistencia con lo anterior, una matriz de identidad pareada (**figura 33**) con los taxa incluidos en la filogenia de rhabdoviridos muestra que la identidad entre PF_rbv_4 y HdRV2, WiV7 y TBRV1 es de 59-62%, mientras que la identidad entre miembros de un mismo género es típicamente ≥ 70% (excepto para los géneros *Tibrovirus*, *Sigmavirus* y *Caligrhavirus* que muestran id=40-57% entre sus miembros, o el género *Ledantevirus*, que muestra id=62% entre algunos de sus miembros). Por otra parte, esta matriz muestra

también que el porcentaje de indentidad entre miembros de diferentes géneros próximos (i.e., unidos en clados inmediatamente más basales en la filogenia) disminuye a valores de 50-65%.

En los análisis de mapeo y cobertura se encontró que los cuatro contigs rbv están representados en la biblioteca PV1 (de origen) y en el caso de los contigs PF_rbv_1 y PF_rbv_2, también en la biblioteca PV3 (**figura 34**). Por su parte, la biblioteca PV2 fue la que presentó el menor número de reads mapeados a los contigs rbv, y de hecho se observa en los heatmaps ninguno está representado en dicha biblioteca. A diferencia de lo obtenido en el mapeo para grupos de contigs anteriores (dv, ifv, pv y rv), se observó que para el contig PF_rbv_1, la biblioteca PV3 presentó un máximo de cobertura X que supera al máximo de PV1 (pese a que el número de reads mapeados de PV3 es menor). Sin embargo, los heatmaps muestran que la cobertura del contig PF_rbv_1 está mejor distribuida en el caso del mapeo de los reads de PV1 (el valor de la mediana de cobertura es mayor en PV1 que en PV3 y coincide con el valor intermedio del rango min-máx, lo que no ocurre en PV3).

Respecto al contig PF_rbv_4, puede observarse que el máximo de cobertura de los reads de PV2 supera al de PV3 (pese a que el número de reads mapeados de PV2 es menor). En este caso, aunque la mediana para ambos rangos de cobertura es similar, la cobertura esta mejor distribuida para PV3 (que presenta menos regiones con ausencia de reads mapeados). Lo anterior ejemplifica que un máximo de cobertura no siempre se traduce en una mejor distribución de los reads (como ocurre en el caso del contig PF_dv_1). Por otra parte, se observa que hay una porción de ~1700 nt (comprendida aprox. desde el nt 200 hasta el nt 2000) del contig PF_rbv_4 (aprox. el 70% de su secuencia) que se encuentra cubierta de forma casi continua con los reads de PV3 (notese que al igual que en PV1, la región correspondiente a las posiciones N del contig muestran ausencia de cobertura). La ausencia de cobertura en algunas posiciones de dicha porción puede deberse a reads que fueron descartados en el filtrado por calidad del mapeo; en cuyo caso, una secuencia similar a PF_rbv_4 (aunque de menor tamaño) pudo haberse ensamblado a partir de los reads de PV3.

Por último, puede observarse que los contigs rbv obtuvieron valores de cobertura en reads/Kb de PV1 equiparables a los observados para los contigs ifv, rv, y algunos contigs pv (**figura 10**).

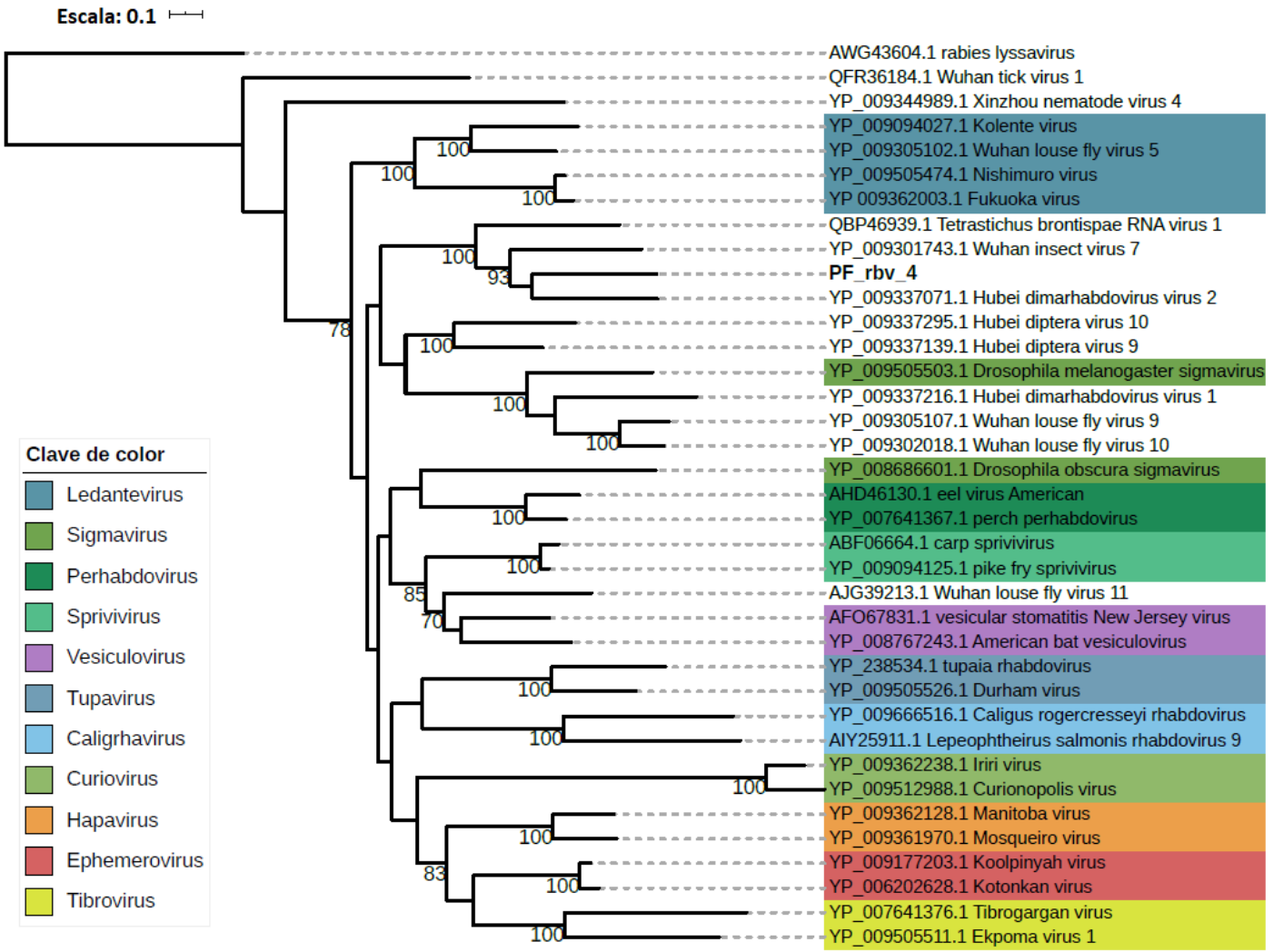

Figura 32. Árbol filogenético construido con la secuencia de proteína de PF_rbv_4 y RdRPs parciales de rhabdoviridos y otros taxa de virus relacionados. Se muestran los valores de BS de las ramas ≥ 70. Los taxa de virus no clasificados no se remarcan en color.

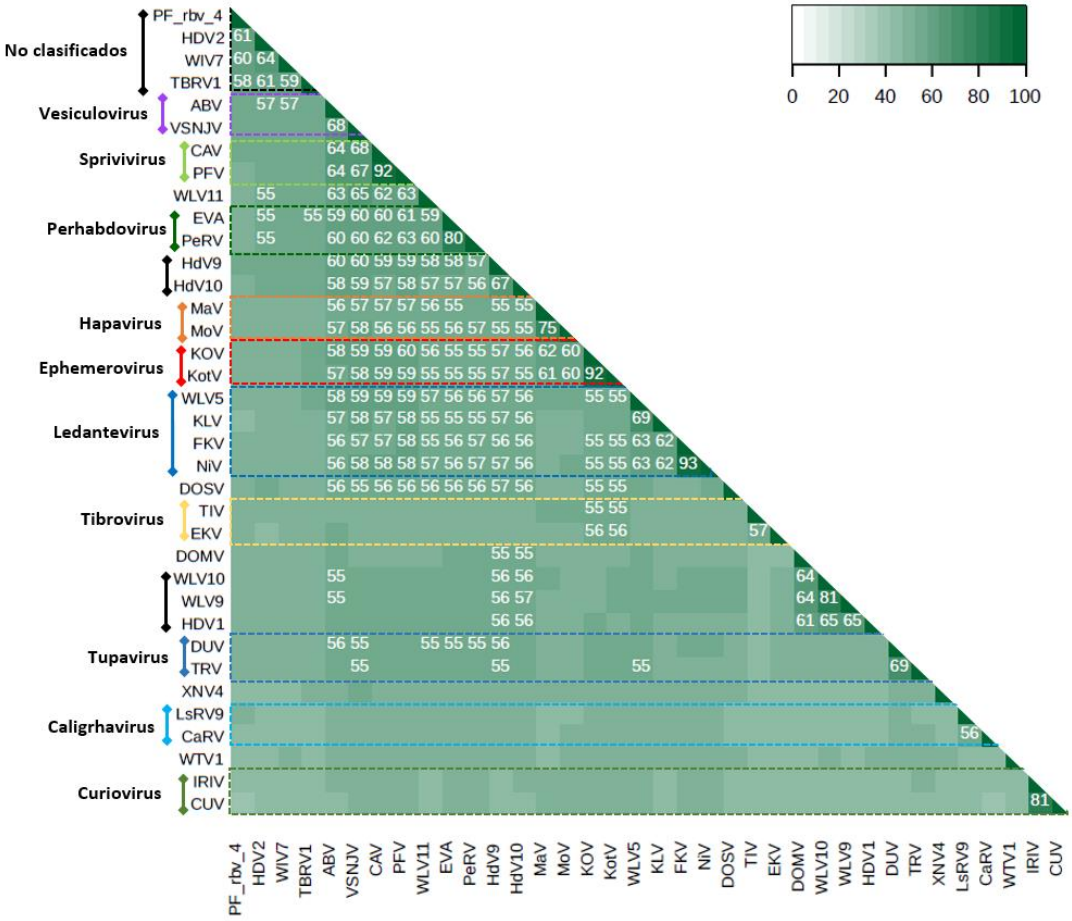

Figura 33. Matriz de porcentaje de identidad pareada de la proteína del ORF de PF_rbv_4 y secuencias parciales de RdRP de taxa de rhabdoviridos. Se muestran los valores ≥ 55%**.** Los taxa incluidos son Hubei dimarhabdovirus virus 2 (HDV2; YP_009337071.1), Wuhan Insect virus 7 (WIV7; YP_009301743.1), Tetrastichus brontispae RNA virus 1 (TBRV1; QBP46939.1), American bat vesiculovirus (ABV; YP_008767243.1), vesicular stomatitis New Jersey virus (VSNJV; AFO67831.1), carp sprivivirus (CAV; ABF06664.1), pike fry sprivivirus (PFV; YP_009094125.1), Wuhan Louse Fly Virus 11 (WLV11; AJG39213.1), eel virus American (EVA; AHD46130.1), perch perhabdovirus (PeRV; YP_007641367.1), Hubei diptera virus 9 (HdV9; YP_009337139.1), Hubei diptera virus 10 (HdV10; YP_009337295.1), Manitoba virus (MaV; YP_009362128.1), Mosqueiro virus (MoV; YP_009361970.1), Koolpinyah virus (KOV; YP_009177203.1), Kotonkan virus (KotV; YP_006202628.1), Wuhan Louse Fly Virus 5 (WLV5; YP_009305102.1), Kolente virus (KLV; YP_009094027.1), Fukuoka virus (FKV; YP_009362003.1), Nishimuro virus (NiV; YP_009505474.1), Drosophila obscura sigmavirus (DOSV; YP_008686601.1), Tibrogargan virus (TIV; YP_007641376.1), Ekpoma virus 1 (EKV; YP_009505511.1), Drosophila melanogaster sigmavirus (DOMV; YP_009505503.1), Wuhan Louse Fly Virus 10 (WLV10; YP_009302018.1), Wuhan Louse Fly Virus 9 (WLV9; YP_009305107.1), Hubei dimarhabdovirus virus 1 (HDV1; YP_009337216.1), Durham virus (DUV; YP_009505526.1), tupaia rhabdovirus (TRV; YP_238534.1), Xinzhou nematode virus 4 (XNV4; YP_009344989.1), Lepeophtheirus salmonis rhabdovirus 9 (LsRV9; AIY25911.1), Caligus rogercresseyi rhabdovirus (CaRV; YP_009666516.1), Wuhan Tick Virus 1 (WTV1; QFR36184.1), Iriri virus (IRIV; YP_009362238.1), Curionopolis virus (CUV; YP_009512988.1).

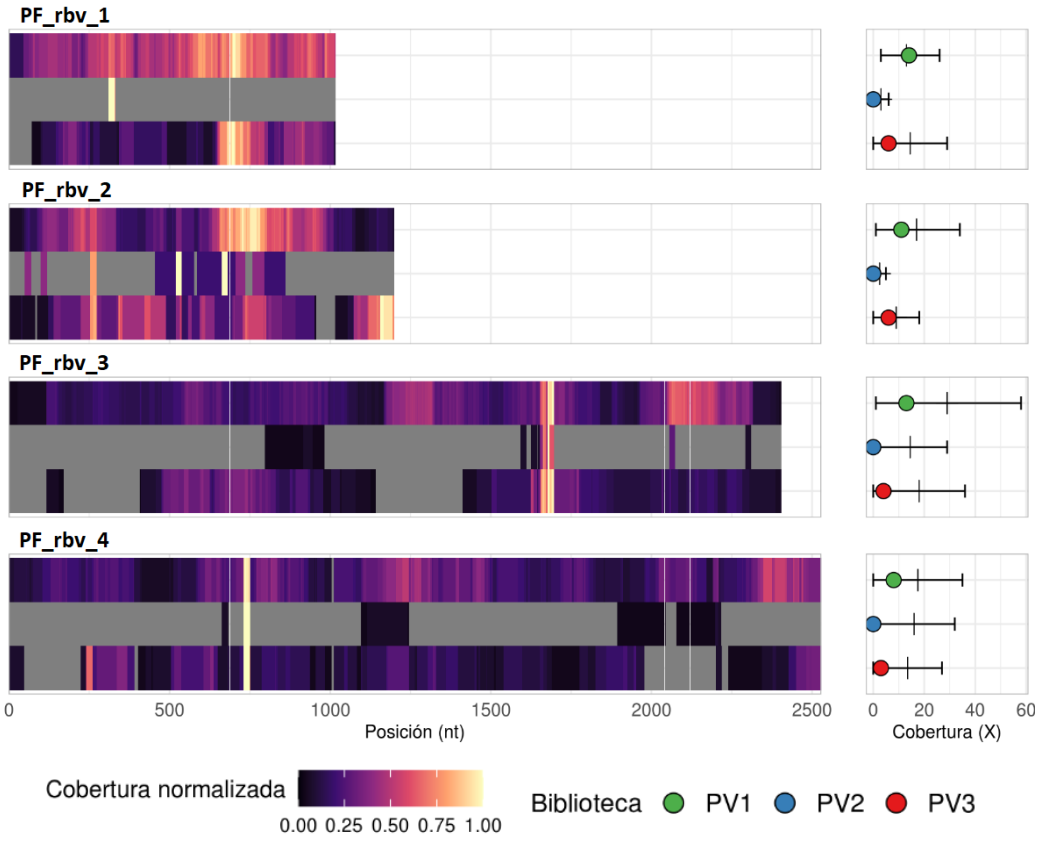

Figura 34. Heatmaps y gráficas de cobertura absoluta para los contigs rbv. La descripción de esta figura es la misma dada para la figura 16.

3.4.6. Secuencias virales putativas relacionadas con la familia *Tombusviridae*

El contig PF_tbv_1 es la única secuencia recuperada, cuyo análisis de similitud a nivel de aminoácidos mostraron homología a miembros de dos clados de RNA no relacionados. Por un lado, un ORF no estructural presentó homología a miembros del clado "Tombus-Noda" descrito por Shi et al. (2016), que incluye a las familias *Tombusviridae*, *Luteoviridae*, *Solemoviridae* y otros taxa relacionados, no clasificados. Por otra parte, un segundo ORF, con dominio estructural, presentó homología a miembros del clado "Permutotetra", también descrito por Shi et al., que incluye a los miembros de la familia *Permutotetraviridae* y otros taxa relacionados, no clasificados. Los dos ORFs detectados cubren colectivamente un 66% de la secuencia del contig PF_tbv_1 (**figura 35** y **anexo C**).

Los análisis de BLASTP con las bases de datos NR y RefSeq mostraron que la proteína del ORF1 (441 aa) de PF_tbv_1 tiene homología con identidad limitada con la proteína hipotética del ORF2 de Sanxia tombus-

like virus 5 (StV5; id=38%, qcov=94% y e=1.5E-77) y con la proteína hipotética no estructural del ORF2 de Wuhan spider virus 9 (WSV9; id=33%, qcov=95 y e=2.68E-65).

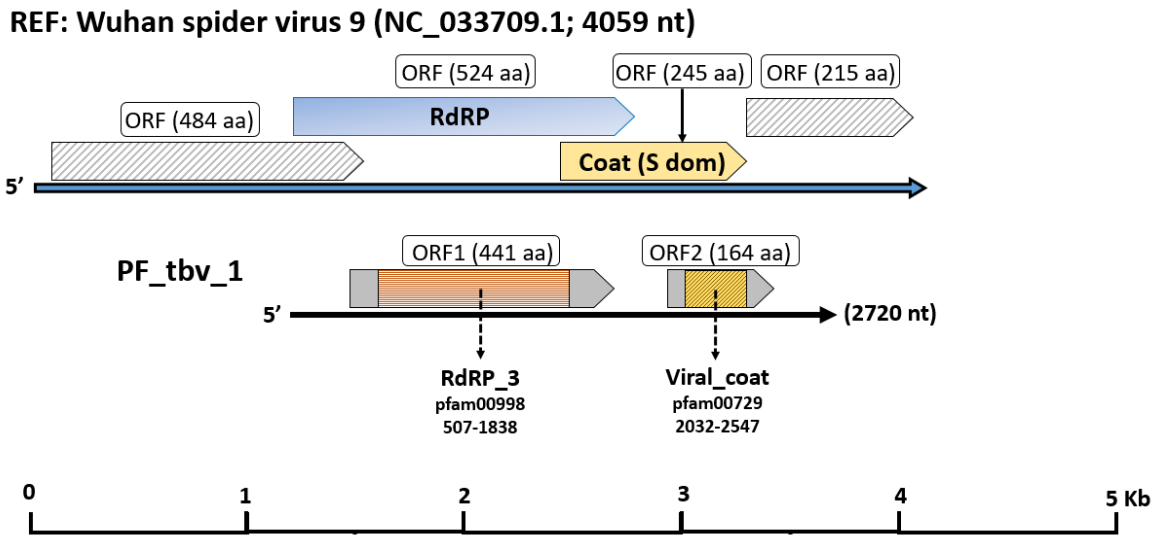

Figura 35. Organización genómica predicha para el contig PF_tbv_1. El contig (flecha negra sólida) y los ORFs (cajas grises sobre el contig), así como el ID de dominios conservados encontrados en las proteínas predichas a partir de los ORFs. Las coordenadas de los dominios indican la región del contig donde están codificados. Por comparación se muestra la organización genómica de un virus relacionado con tombusviridos (flecha azul), su número de acceso del Genebank, la longitu del genoma completo y las proteínas que codifica. Las proteínas sin dominios o función predicha del genoma de referencia se muestran en cajas con franjas grises en diagonal.

Ambos hits corresponden a taxa de virus no clasificados, relacionados con la familia *Tombusviridae*. Dado que la organización genómica de WSV9 es más acorde a la observada en el contig PF_tbv_1, se utilizó como referencia en la **figura 35**. Por su parte, la proteína del ORF2 (164 aa) de PF_tbv_1 mostró homología con la proteína putativa de cápside del ORF2 de Hubei permutotetra-like virus 4 (HpV4; %id=30, qcov=98% y e=1E-12), tanto en los análisis con la base de datos NR y RefSeq. También se obtuvieron hits con proteínas estructurales de virus no clasificados (relacionados con distintas familias) como la proteína putativa de cápside de Atrato Sobemo-like virus 5 (AtSV5; id=25%, qcov=96% y e=2E-10) o la proteína de cápside mayor (ApGlV2; id=25.5%, qcov=98% y e=6E-09) de Aphis glycines virus 2, codificada en el ORF3 de su genoma. AtSV5 es un taxa relacionado con el género *Sobemovirus*, de la familia *Solemoviridae*, mientras que ApGlV2 es un taxa con características de secuencia tanto de tetraviridos[3] como del género *Sobemovirus*.

[3]Nota: Los autores que describieron a ApGlV2 (Liu et al., 2016) utilizan el término tetravirus para hacer referencia a la familia *Tetraviridae*. Sin embargo, desde 2011, dicha familia fue oficialmente dividida en las familias *Alpha-*, *Carmo-* y *Permuto- tetraviridae* (King et al., 2011).

En correspondencia con los análisis de similitud, la proteína del ORF1 de PF_tbv_1 presentó un dominio de RdRp_3 (pfam00998; del aa 48 al aa 426) en la región del alineamiento con la proteína de WSV9 (también con dominio de RdRP_3 en la región del alineamiento), indicando que la proteína predicha correspondiente es de función no estructural. Por su parte, la proteína del ORF2 de PF_tbv_1 se alinea con la porción que va del sitio 47 al 215 de la secuencia del ORF2 de HpV4, cubriendo su respectivo dominio de cápside. En correspondencia, se detectó un dominio de cápside viral_coat (Pfam00729; del aa 47 al aa159) en el ORF2 de PF_tbv_1.

Es importante enfatizar que aunque el ORF1 y el ORF2 de PF_tbv_1 presentaron hits con proteínas no estructurales y estructurales, respectivamente, de grupos taxonómicos diferentes, tanto el genoma de referencia de WSV9 como el de ApGlV2 tienen una organización tal que presentan un ORF de proteína no estructural (con dominio de RdRP) rio arriba de un ORF de proteína estructural (con dominio de cápside), al igual que lo observado con el contig PF_tbv_1. No obstante, los ORFs de este último no se empalman, como sí ocurre con los ORFs correspondientes de WSV9 y ApGlV2.

Las relaciones filogenéticas reconstruidas con la secuencia del ORF1 de PF_tbv_1 (**figura 36**) lo sitúan entre los taxa no clasificados relacionados, que si bien presentan similitud de secuencia y/o organización genómica similar a la familia *Tombusviridae*, son filogenéticamente lejanos a culquiera de sus miembros incluidos (nótese la diferencia en la longitud de las ramas de los clados que incluyen los miembros de un género en relación con las ramas de los clados que incluyen a los taxa no clasificados). Cabe mencionar que si bien los miembros de la familia *Tombusviridae* se han descrito como virus que infectan a plantas, aquellos que se observan anidados más estrechamente con PF_tbv_1 en la filogenia (como STV6, STV5, HTV18, WSV9) se han asociado a insectos o arácnidos hospederos, excepto SWSV5, que se asocia con muestras de crustaceos. Así, al considerar las relaciones filogenéticas, aunado a los análisis de similitud, cabe hipotetizar a PF_tbv_1 como una secuencia de un virus de RNA no reportado.

Los resultados del mapeo mostraron que el contig PF_tbv_1 solo se encuentra representado en la biblioteca PV3 (la biblioteca de origen) y se observa una cobertura nula a dicho contig con los reads de PV1 y PV2 (2 y 1 reads mapeados, respectivamente) (**figura 37**). Con los reads de PV3, se alcanza un máximo de cobertura de 70X con una mediana de 20X, i.e., más del 50% de sitios se encuentran por debajo de una cobertura de ~35X (el valor intermedio del rango min-máx. de cobertura X). Por otra parte, se observa que una porción cerca del extremo 5’ presentó la cobertura mínima. El valor de cobertura en reads/Kb de PV1 obtenido por la secuencia PF_tbv_1 es equiparable a la cobertura de los contigs ifv, rv y rbv y tres contigs pv (PF_pv_3, PF_pv_4 y PF_pv_5) (**figura 10**).

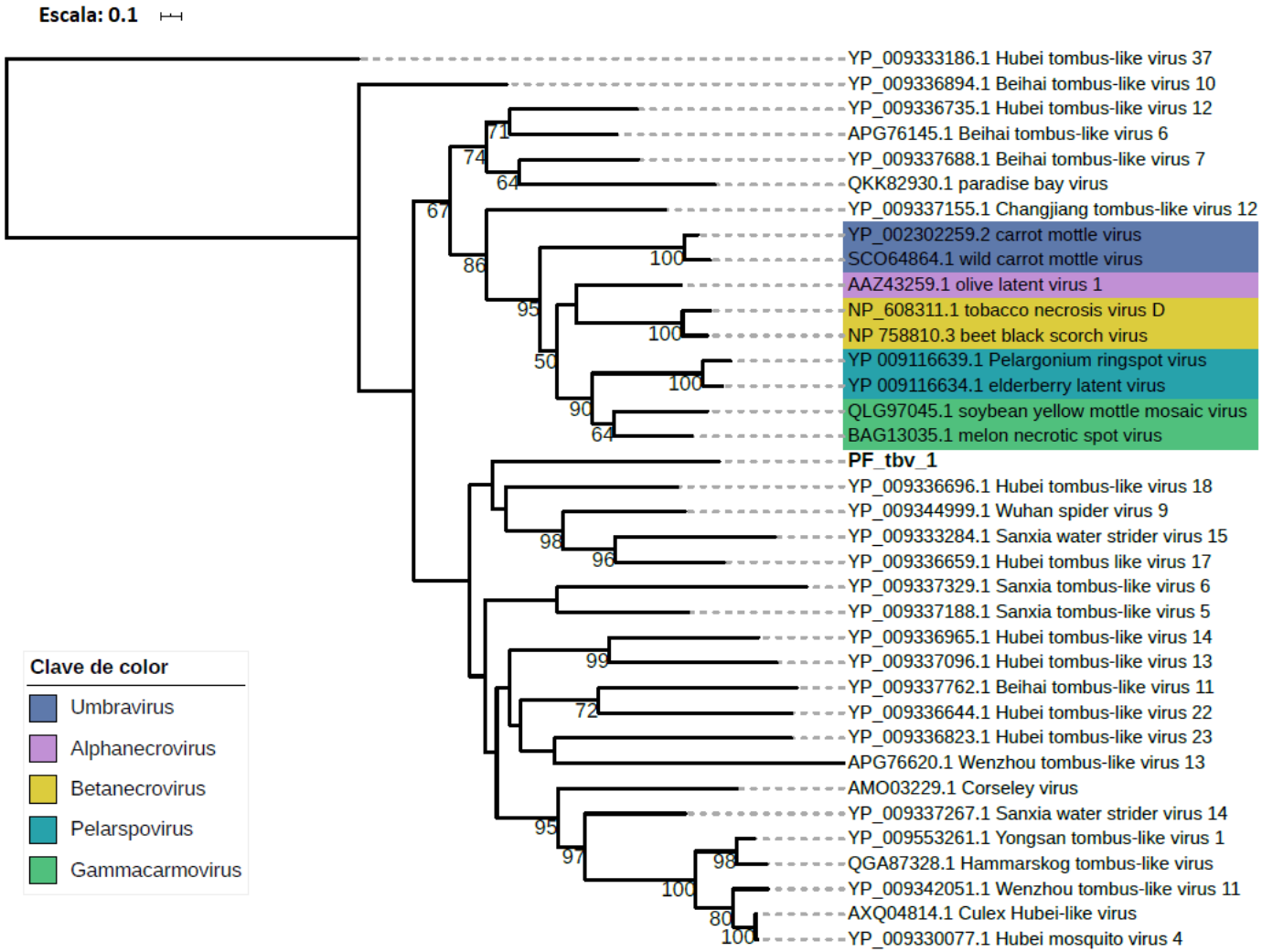

Figura 36. Árbol filogenético reconstruido con la secuencia de proteína del ORF1 de PF_tbv_1 y secuencias parciales de RdRP de tombusviridos y taxa de virus relacionados. Los taxa no clasificados no se remarcan con color. Se muestran los valores de BS de las ramas ≥ 50.

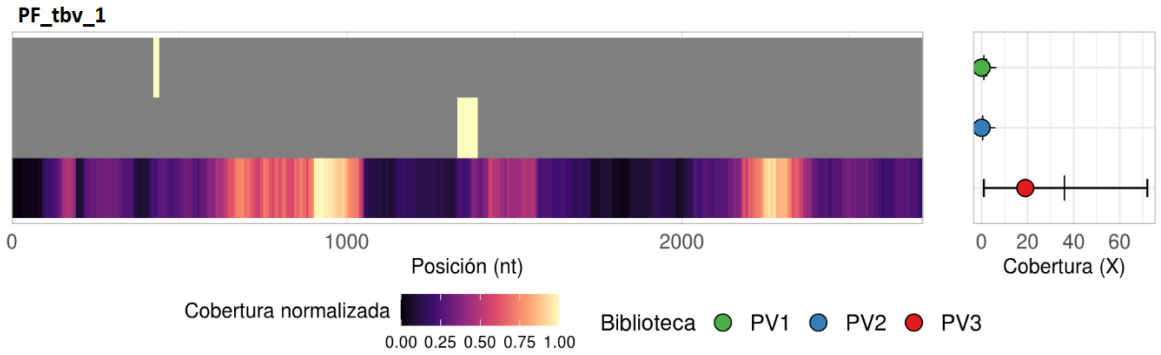

Figura 37. Heatmaps y gráficas de cobertura absoluta para el contig PF_tbv_1. La descripción de esta figura es la misma dada para la figura 16.

3.4.7. Secuencias virales putativas relacionadas con la familia *Tymoviridae*

Los contigs PF_mfv_1 y PF_mfv_2 presentaron homología con el género *Marafivirus* de la familia *Tymoviridae*. Para estos contigs ("contigs mfv"), la homología inferida con dichos taxa virales está fuertemente soportada por la alta similitud de secuencia y organización genómica que presentan. La secuencia del contig PF_mfv_2 es el inverso complementario a la secuencia recuperada durante la selección manual, por lo que las características de secuencia nucleotídica se han ajustado a la secuencia que aquí se presenta.

El contig PF_mfv_1, de 6423 nt de longitud, contiene un solo ORF a lo largo de toda su secuencia, el cual comienza a partir del nt 145 y se extiende hasta el nt 6423 y codifica una proteína hipotética de 2093 aa de longitud. Los análisis de BLASTP indican que la proteína predicha presentó identidad elevada a lo largo de toda su longitud (id=93%, qcov=98% y e=0) con la poliproteína del genoma de grapevine Syrah virus 1 (GSyV1), un virus del género *Marafivirus* de la familia *Tymoviridae* (**figura 38**). Cabe mencionar que durante el análisis de selección se encontraron 19 secuencias cuyo mejor hit de BLASTP fue GSyV-1, de las cuales el contig PF_mfv_1 se seleccionó al ser la secuencia más larga de todas.

De forma consistente, el análisis de dominios reveló que la secuencia de aminoácidos de la proteína de PF_mfv_1 presenta, en el mismo orden y posición, todos los dominios canónicos de la poliproteína única codificada en el genoma de GSyV1. Comenzando por el extremo 5', correspondiente al amino-terminal de la proteína predicha, tiene un dominio de metil-transferasa (MTR, pfam01660; del aa 152 al aa 433), un dominio de peptidasa (PRO, pfam05381; del aa 829 al aa 927), un dominio de helicasa (HEL, pfam01443; del aa 1022 al aa 1252), un dominio de polimerasa de RNA dependiente de RNA (RdRP2, pfam00978; del

aa 1558 al aa 1719) y un dominio de cápside (Tymo-coat, pfam00983; del aa 1939 al aa 2080). También se encontró un dominio de proteína larga de tegumento UL36 (LTP, PHA03247; del aa 624 al aa 830), en una región adyacente y río arriba del dominio HEL (**figura 38**).

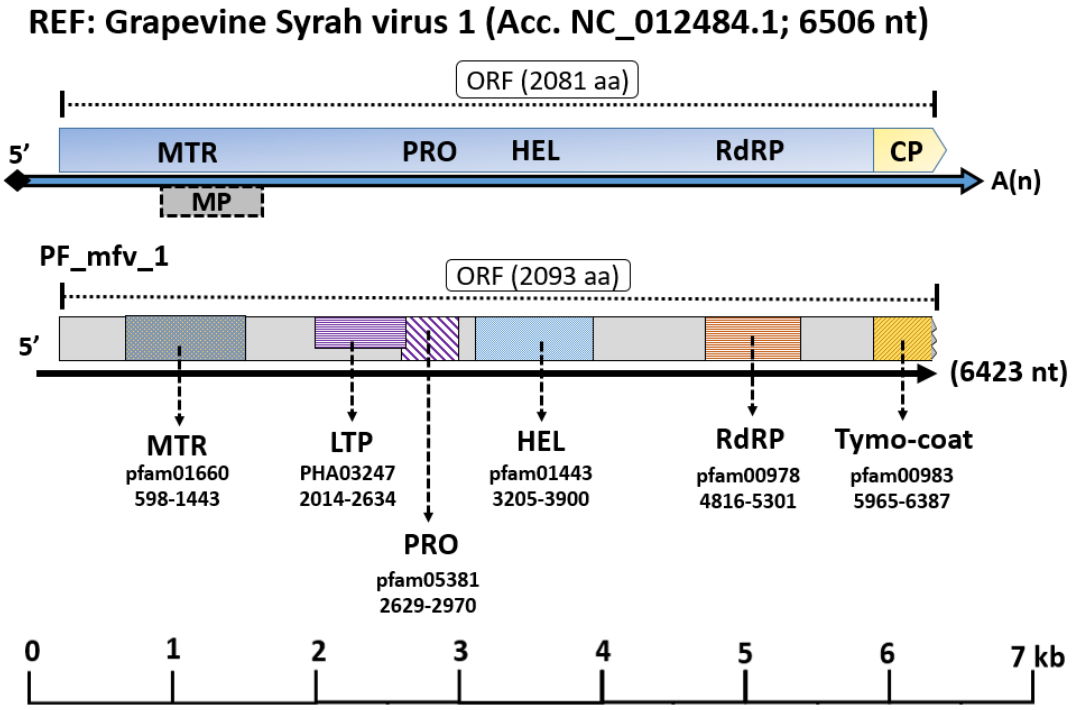

Figura 38. Organización genómica predicha para el contig PF_mfv_1. Se muestra el contig (flecha negra sólida) y el ORF (caja gris sobre el contig), así como el ID y coordenadas de los dominios estructurales y no estructurales encontrados en la proteína hipotética del ORF. Las coordenadas de los dominios indican la región de la secuencia nucleotídica donde se encuentran codificados. La ausencia de codón de paro en el ORF se indica con un borde dentado. Por comparación, se incluye la organización genómica de GSyV1 (flecha azul), su número de acceso del Genebank, la longitu del genoma completo, y el ORF con la poliproteína que codifca (caja de color sobre el genoma). El extremo cap-5' se representa con un diamante negro. A(n): cola poli-A en el extremo 3'.

El MSA de secuencias de poliproteína de diferentes aislados de GSyV1, que incluye la secuencia de aminoácidos del ORF de PF_mfv_1, mostró que esta última contiene una secuencia única de 36 aa codificada en la región proximal al extremo 5', que difiere del inicio "canónico" de los ORFs de los aislados de GSyV1 (**figura 39**, A). Más aún, la proteína de PF_mfv_1 también presenta una secuencia casi idéntica con el inicio del resto de los aislados de GSyV1, justo a continuación de dicha secuencia inicial única. La revisión de la secuencia de nucleótidos río arriba del ORF (con el marco de lectura correspondiente) de cada uno de los aislados de GSyV1 no muestra algun codón para metionina que sea anterior a la metionina anotada en el genoma como inicio de la poliproteína, lo que sí ocurre en el caso del ORF1 de PF_mfv_1.

Por otra parte, se observó que el ORF del contig PF_mfv_1 no está completo ya que no se encontró la señal de terminación en su secuencia correspondiente a un codón de paro en la secuencia del contig. El dominio de CP (el más cercano al extremo 3' del ORF) está truncado. En el MSA mencionado anteriormente también puede observarse una alta conservación de aminoácidos (>90%) a lo largo de todo el alineamiento, así como un gap de 22 sitios para la secuencia del ORF1 después de su último aminoácido, lo cual coincide con la terminación prematura de la secuencia codificante del contig (**figura 39**, B).

Respecto a lo anterior es importante mencionar que el contig PF_mfv_1 fue generado a partir de rensamble (**figura 39**, C); PF_mrv_1 se formó a partir del empalme de dos contigs (nombrados aquí por simplicidad como contig 1 y contig 2), en una región de 86 nt donde la similitud entre secuencias estuvo por encima del 98% (umbral del programa Cap3 durante el reensamble). El contig 1, que constituye el extremo 5’ del contig PF_mfv_1, contiene un ORF que comienza con la secuencia única de 36 aa mencionada anteriormente, seguida por el inicio canónico de los aislados de GSyV1. El ORF del contig 1 está truncado, y termina 122 aa río abajo del inicio del ORF de PF_mfv_1 (90 aa río abajo del inicio del ORF del resto de los aislados), mientras que el contig 2 (que constituye el resto del contig PF_mfv_1 hasta el extremo 3’) contiene un ORF codificado a partir de su sitio nucleotídico 100 (de su sitio 1 al 100 es donde se encuentra la región de empalme con el contig 1). El ORF del contig 2 comienza 123 aa río abajo respecto al ORF de PF_mfv_1 (91 aa río abajo del inicio del ORF1 del resto de los aislados) y su secuencia comienza (después del reensamble) da continuidad a la del ORF de PF_mfv_1.

La segunda secuencia recuperada con afiliación taxonómica a la familia *Tymoviridae* es el contig PF_mfv_2 de 4035 nt de longitud y con un ORF que abarca desde el nt 105 al nt 4241, y que codifica una proteína hipotética de 1379 aa (**figura 40**). Los análisis de BLASTP vs RefSeq/NR de la proteína del ORF de PF_mfv_2 indican homología tentativa con alta identidad (qcov=98%, id=97% y e=0), con la secuencia de grapevine rupestris vein feathering virus (GRVFV), otro miembro del género *Marafivirus*. A diferencia del ORF del contig PF_mfv_1, que abarca casi toda la secuencia de sus respectivo *subject* (la poliproteína de GSyV1), el ORF del contig PF_mfv_2 solo representa ~60% de la longitud de la poliproteína de aislados de GRVFV. Durante la selección manual se encontraron otras 65 secuencias con similitud a GRVFV, siendo PF_mfv_2 seleccionada por la longitud de su secuencia y la elevada similitud con dicho taxa.

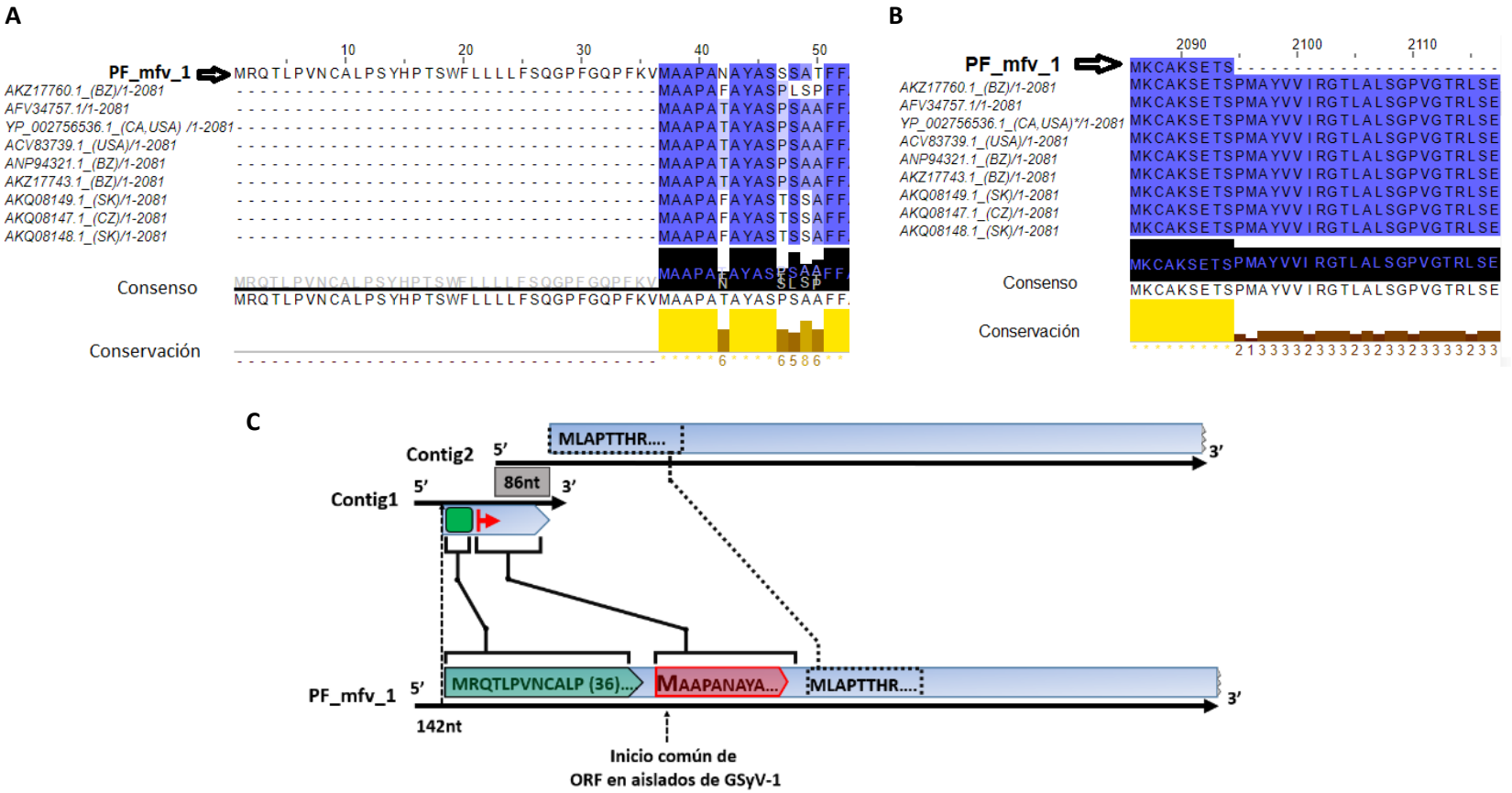

Figura 39. MSA y logo de las secuencias de poliproteína del ORF de PF_mfv_1 y los aislados de GSyV1. (A) Se observa una secuencia de 36 aa única al inicio de la secuencia del ORF de PF_mfv_1 (extremo 5'), seguida de una secuencia de inicio altamente conservada en los aislados de GSyV1. El logo (al pie del alineamiento) indica el grado de conservación de los nucleótidos en cada sitio en proporción al tamaño de las barras rectangulares. (B) En el extremo 3' de la secuencia del ORF1 se observa un gap de 22 sitios, indicando la terminación prematura del ORF1 de PF_mfv_1. (C) Esquema del empalme de las secuencias parentales del contig PF_mfv_1, llamadas aquí como Contig 1 y Contig 2, mostrando sus respectivos ORFs y la posición de estos respecto a la secuencia de del ORF de PF_mfv_1. Nótese la metionina (M) prematura en el contig reensamblado y la M del inicio canónico (que se encuentran solo en el ORF del contig 1), mientras el contig 2 no contiene esta última y su ORF comienza hasta una M codificada a partir del nt 87 de su secuencia, que corresponde al nt 410 respecto a la secuencia de PF_mfv_1.

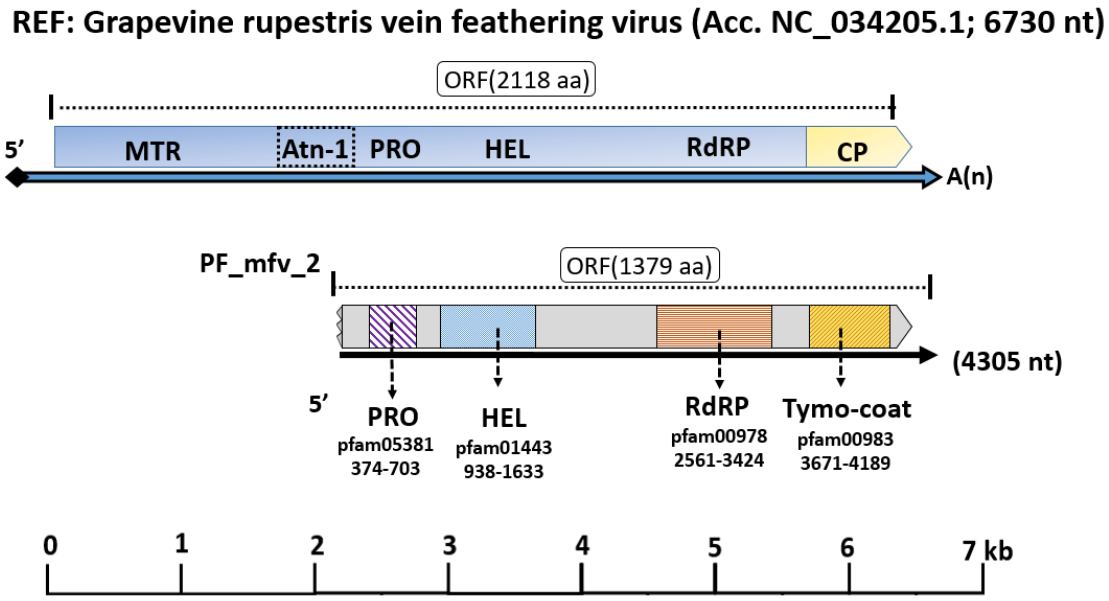

Figura 40. Organización genómica predicha para el contig PF_mfv_2. Se muestra el contig (flecha negra sólida) y el ORF (caja gris sobre el contig), así como el ID y coordenadas de los dominios estructurales y no estructurales encontrados en la proteína hipotética del ORF. Las coordenadas de los dominios indican la región de la secuencia nucleotídica donde se encuentran codificados. La ausencia de codón de inicio en el ORF se indica con un borde dentado. Por comparación, se incluye la organización genómica de GRVFV (flecha azul), su número de acceso del Genebank, la longitu del genoma completo, y el ORF con la poliproteína que codifca (caja de color sobre el genoma). El extremo cap-5' se representa con un diamante negro. A(n): cola poli-A en el extremo 3'.

De acuerdo con los análisis de similitud y búsqueda de dominios, la porción de poliproteína contenida en el ORF de PF_mfv_2 es la más proximal al extremo 3' del genoma de referencia (i.e., una fracción no estructural y la porción completa de la poliproteína estructural). En dirección 5' → 3', los dominios presentes son peptidasa (PRO, pfam05381; del aa 104 al aa 197), helicasa (HEL, pfam01443; del aa 293 al aa 523), RdRP (RdRP2, pfam00978; del aa 833 al aa 1120) y cápside de tymovirus (Tymo-coat, pfam00983; del aa 1212 al aa 1374).

La reconstrucción filogenética hecha con las secuencias completas de las poliproteínas de los ORFs de diferentes taxa virales del género *Marafivirus*, así como de PF_mfv_1 y PF_mfv_2, muestran que PF_mfv_1 se anida en un clado (con soporte de BS=100) que contiene a los aislados de GSyV1, mientras que PF_mfv_2 se ubica en un clado que contiene a los aislados GRVFV (con soporte de BS=100) (**figura 41**). Dada la alta identidad de secuencia que presentaron las proteínas *queries* en los análisis de similitud, el resultado anterior era esperado. Sin embargo, se incluyeron diferentes aislados de GSyV-1 y GRVFV (aquellos cuya secuencia era completa) para esclarecer más detalladamente la relación de estos con los contigs mfv. PF_mfv_1 está hermanado con un aislado de Brazil (lo cual es sobresaliente dado que ningún

otro aislado mostró un agrupamiento semejante). En contraste, se observó que PF_mfv_2 no se encuentra hermanado con algún otro aislado.

Los análisis de indentidad pareada mostraron que la secuencia de nucleótidos de PF_mfv_1 (**figura 42**) es más similar (84%) a los aislados BS3 de Brasil (KT037017.1) y ES2 de Eslovaquia (KP221256.1). Sin embargo, este último es más similar (92-99%) a otros aislados de Europa, Brasil y de EUA. Las secuencias de PF_mfv_1 y BS3 son los más divergentes respecto al resto de los aislados (83-84%), y a la vez, la similitud entre PF_mfv_1 y BS3 solo es ligeramente mayor (en 1%) de lo que es PF_mfv_1 con el resto de las secuencias. Los porcentajes de identidad a nivel de proteína mostraron resultados correspondientes, con la secuencia de PF_mfv_1 y BS3 siendo ligeramente más similares entre si (95%) que con el resto de los aislados (93-94%), y a su vez estos últimos siendo poco divergentes entre si (96-99%). Por su parte, la secuencia de nt de PF_mfv_2 (**figura 43**) obtuvo la mayor similitud con aislados S1 de Suiza (KY513702.1), F1 de Francia (NC_034205.1) y NZ1 de Nueva Zelanda (MF000326.1). Sin embargo, a nivel de aminoácidos, la secuencia de PF_mfv_2 mostró valores de similtud casi iguales con la mayoría de los aislados analizados (de 95-96%), excepto los aislados A7 y A8 australianos (QKI36478.1 y QKI36477.1), NZ2 de Nueva Zelanda (ATB20100.1) y EU1 de EUA (AAW33732.1). Por su parte, los aislados A7, A8, NZ2 y EU1 son más similares entre sí (79-94% en genoma y 91-99% en proteína), y los más divergentes respecto al resto de los aislados (78-79% en genoma y 88-90% en proteína). Estos resultados de similitud con los contigs mfv y el resto de los aislados indican que no hay una relación directa entre la similitud de las secuencias y la región de procedencia.

Los resultados del mapeo mostraron que los contigs mfv son los que obtuvieron el mayor número de reads/Kb de entre todas las secuencias seleccionadas, en las tres bibliotecas (**figura 44** y **figura 10**). Se determinó que los contigs mfv se encuentran representados en la biblioteca PV1 (la biblioteca de origen), donde hay un máximo de cobertura X por sitio nucleótidico de hasta 1000X y 625X para PF_mfv_1 y PF_mfv_2, respectivamente. Para ambos contigs el valor de la mediana de cobertura está cerca del valor intermedio del rango min-máx (i.e., al menos el 50% de los sitios de PF_mfv_1 y PF_mfv_2 presentaron coberturas de ~375X y 250X, respectivamente).

Por otra parte, se observó que mientras PF_mfv_1 no se encuentra representado en las bibliotecas PV2 y PV3, PF_mfv_2 es cubierto en gran parte de su secuencia por los reads de dichas bibliotecas (aunque no estrictamente en una porción contínua ≥ 70%). Respecto al mapeo de los reads de PV3, se observa una cobertura casi completa de la secuencia de PF_mfv_2, con una pequeña región (aprox. en la posición del nt 2400) que muestra ausencia de cobertura. Considerando que se recuperaron varios contigs con similitud

a GRVFV, es posible que los reads de PV3 mapeados a PF_mfv_2 representen a uno de los contigs de menor tamaño con similitud a dicho taxa (aunque no se verificó la biblioteca de origen para dichos contigs de menor tamaño). Por tanto, se consideró que PF_mfv_2 se encuentra representado en la bibioteca PV3. Respecto al mapeo de los reads de PV2 a PF_mfv_2, puede observarse una cobertura X y una distribución de los reads mapeados similar a la observada en PV3 (aunque con una mediana ligeramente menor). Por lo tanto, y tomando en cuenta que en el mapeo de los reads de PV2 y PV3 hay una cobertura en reads/Kb para PF_mfv_2 mayor que la correspondiente a los contigs ifv, rv, rbv o tbv, se determinó que PF_mfv_2 está representado en la biblioteca PV2.

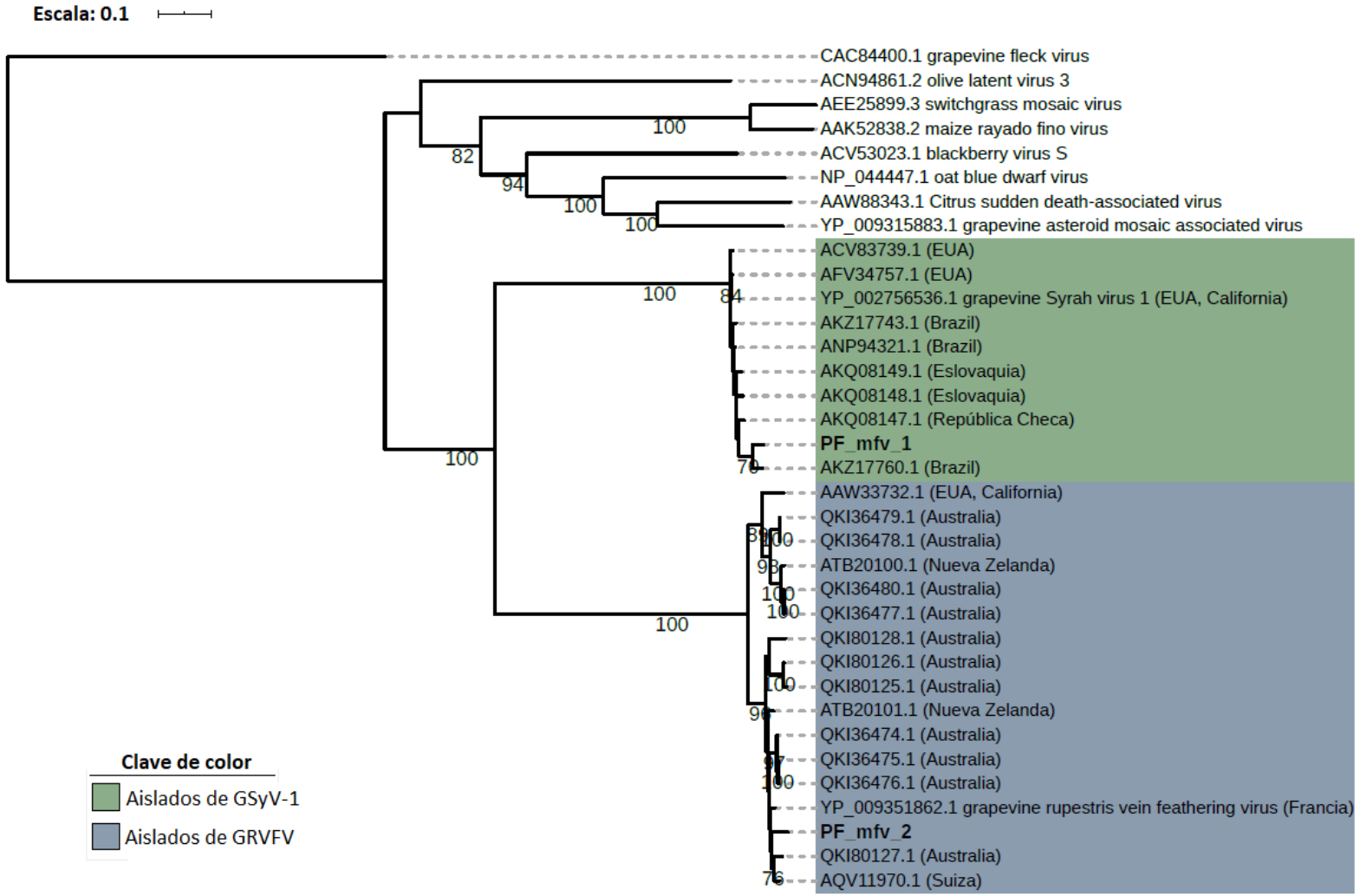

Figura 41. Árbol filogenético construido las secuencias de poliproteína del ORF de PF_mfv_1, del ORF de PF_mfv_2 y de taxa del género *Marafivirus*. Se remarcan en color los aislados de GSyV-1 y GRVFV (el nombre completo se da para la secuencia de la base de datos RefSeq). Se muestran los valores de BS de las ramas ≥ 70.

Genoma	83	84	>84
Proteína	93-94	95	>95

		KP221255.1 (R. Checa)	KP221257.1 (Eslovaquia)	KP221256.1 (Eslovaquia)	KR153306.1 (Brasil)	KX130754.1 (Brasil)	FJ977041.1 (USA)	NC_012484.1 (EUA)	JX513896.1 (EUA)	KT037017.1 (Brasil)	**PF-mfv-1**
RC1	AKQ08147.1 (R. Checa)	-	93	93	92	93	93	93	92	84	83
ES1	AKQ08149.1 (Eslovaquia)	97	-	93	92	92	93	93	92	83	83
ES2	AKQ08148.1 (Eslovaquia)	97	97	-	92	93	93	93	93	83	84
BS1	AKZ17743.1 (Brasil)	96	97	96	-	95	95	95	95	83	83
BS2	NP94321.1 (Brasil)	97	97	97	98	-	96	97	96	83	83
EU1	ACV83739.1 (EUA)	96	96	96	98	98	-	99	98	83	83
EU2	YP_002756536.1 (EUA)	96	97	97	98	99	99	-	99	83	83
EU3	AFV34757.1 (EUA)	96	96	96	98	98	99	99	-	83	83
BS3	AKZ17760.1 (Brasil)	93	93	93	93	93	93	93	93	-	84
	PF-mfv-1	94	93	93	94	94	93	94	94	95	-

Figura 42. Matriz de identidad pareada a nivel de genoma y poliproteína completa para PF_mfv_1 y aislados de GSyV-1. Los intervalos de color se establecieron en función del mímino (83% para genoma y 93% para proteína) y máximo (84% genoma y 95% proteína) de identidad de PF_mfv_1 con los otros aislados. Los nombres resaltados en **negrita** corresponden al ID asignado a cada aislado.

Genoma	< 80	80	85	>85
Proteína	< 91	91-95	96	>96

		MN974274.1 (Australia)	MN974273.1 (Australia)	MN974275.1 (Australia)	KY513702.1 (Suiza)	MT084807.1 (Australia)	MT084809.1 (Australia)	NC_034205.1 (Francia)	MF000326.1 (N. Zelanda)	MN974276.1 (Australia)	**PF-mfv-2**	MT084812.1 (Australia)	MT084811.1 (Australia)	MF000325.1 (N. Zelanda)	AY706994.1 (EUA)
A1	QKI80126.1 (Australia)	-	94	83	83	82	82	83	83	82	82	79	79	79	78
A2	QKI80125.1 (Australia)	99	-	83	83	82	82	82	82	82	82	79	79	78	78
A3	QKI80127.1 (Australia)	94	94	-	86	85	85	84	85	83	84	79	79	79	79
S1	AQV11970.1 (Suiza)	94	94	97	-	85	85	85	84	84	85	79	79	79	78
A4	QKI36474.1 (Australia)	94	93	96	96	-	99	85	84	83	83	78	79	79	78
A5	QKI36476.1 (Australia)	94	93	96	96	99	-	84	84	83	84	78	79	79	78
F1	YP_009351862.1 (Francia)	94	94	96	97	96	96	-	86	83	85	79	79	79	78
NZ1	ATB20101.1 (N. Zelanda)	94	94	96	96	96	96	97	-	83	85	79	79	79	78
A6	QKI80128.1 (Australia)	93	93	94	95	93	93	94	94	-	83	79	79	79	78
	PF-mfv-2	95	95	96	96	96	96	96	96	96	-	80	80	80	79
A7	QKI36478.1 (Australia)	90	90	90	90	90	90	90	90	89	91	-	83	83	80
A8	QKI36477.1 (Australia)	90	90	90	90	90	90	90	90	89	91	96	-	94	79
NZ2	ATB20100.1 (N. Zelanda)	90	90	90	90	90	89	90	90	89	91	96	99	-	80
EU1	AAW33732.1 (EUA)	89	89	89	90	89	89	90	89	88	91	92	92	92	-

Figura 43. Matriz de identidad pareada a nivel de genoma y poliproteína completa para PF_mfv_2 y aislados de GRVFV. Los intervalos de color se establecieron en función del mímino (80% para genoma y 91% para proteína) y máximo (85% para genoma y 96% para proteína) de identidad de PF_mfv_2 con los otros aislados. Los nombres resaltados en **negrita** corresponden al ID asignado a cada aislado.

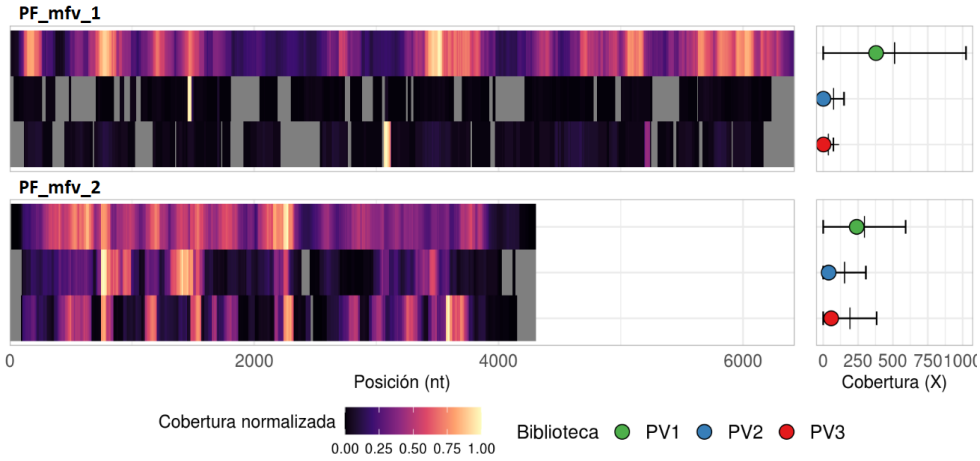

Figura 44. Heatmaps y gráficas de cobertura absoluta para los contigs mfv. La descripción de esta figura es la misma dada para la figura 15.

Respecto al mapeo de los reads de las bibliotecas PV2 y PV3, se observaron máximos de cobertura mucho menores (de hasta 125X para PF_mfv_1 con los reads de PV2 y 375X para PF_mfv_2 con los reads de PV3) a los observados con los reads de PV1 para ambos contigs mfv. Por otro lado, los valores de la mediana de cobertura en estas bibliotecas son cercanos a cero, lo que de acuerdo con los heatmaps, indica un mínimo de cobertura casi en todos los sitios de la secuencia de PF_mfv_1 y diferentes regiones a lo largo de la secuencia de PF_mfv_2, además de regiones dispersas con ausencia de cobertura en ambos contigs, en ambas bibliotecas.

3.5. Análisis de partículas tipo virus (VLPs)

El análisis de VLPs se realizó con la finalidad de buscar evidencia adicional sobre la presencia de virus putativos asociados a *P. ficus*. Este tipo de análisis (independiente de secuencias virales) es importante ya que en algunos casos se ha demostrado que algunos ISVs que se replican activamente y producen viriones o estructuras derivadas (como los OBs), no siempre generan síntomas evidentes de infección (Thiem, 1999). Por otra parte, si bien la identificación de especies virales se puede llevar a cabo de forma más derminante por medio del análisis de secuencias, los análisis de VLPs proveen evidencia física directa de la presencia de virus (i.e., ayuda a confirmar las especies designadas como putativas encontradas en los análisis de secuencia). Además, para algunos de los grupos de ISVs de mayor interés, las características morfológicas de los viriones y otros derivados de la replicación (OBs, esferoides, estructuras

paracristalinas, fabricas virales, etc.) se hayan extensamente caracterizados (Thiem, 1999; Attoui et al., 2011; Flint et al, 2015; Williams, 2017). Por lo tanto, la detección e identificación de virus que pueden producir dichas estructuras en alta cantidad en las muestras puede realizarse de forma bastante más rápida y directa en relación a lo que requiere un análisis completo de virómica. Por otra parte, cabe mencionar que una alternativa que combina ambas estrategias es la secuenciación de RNA obtenido a partir de VLPs, ya que por un lado se enriquecen VLPs (de lo cual se puede analizar una submuestra por microscopía electrónica) y a la vez utilizar dicho material para la extracción de ácidos nucleícos, maximizando la abundancia de secuencias virales y reduciendo el rendimiento necesario de una plataforma de secuenciación (Roossnick, 2015).

3.5.1. Observación de VLPs al TEM.

En las micrografías obtenidas del TEM pudieron observarse VLPs con potencial origen viral de diferente tamaño y forma para las tres muestras tratadas en el enriquecimiento de VLPs.

En las micrografías de la muestra P1 se logró captar una posible cápside viral bastoniforme (**figura 45,** A y B). Las dimensiones de dicha partícula fueron de aprox. 10 nm x 2000 nm, lo cual se encuentra dentro del rago esperado para los viriones de algunos ampelovirus, que son virus de plantas cuyos vectores son insectos como el piojo harinoso. Así mismo, se lograron observar partículas con un contorno circular y con una apariencia curveada en sus uperficie, lo cual sugiere cuerpo esferoide para dicha partícula. Dichas partículas alcanzan un diámetro de ~150 nm (**figura 45**, C). No obstante, la forma de estas últimas partículas no corresponde a lo esperado para los viriones de los grupos de virus de dsDNA de interés, aunque su tamaño sí se encuentra en el rango de tamaño de dichos virus. Por otra parte, un acercamiento hacia la superficie de las partículas esferoides de P1 mencionadas anteriormente, revela una apariencia granular semejante a la de las estructuras amorfas más pequeñas e irregulares que las rodenan, por lo cual, dichas partículas esferoides pueden corresponder a agregados de las partículas más pequeñas (**figura 45**, C y D).

Tanto las partículas bastoniformes como las esféricas observadas en la muestra P1 se encuentran rodeadas por las partículas amorfas de menor tamaño que se observan alrededor, las cuales además se encontraron distribuidas en casi toda el área de observación, como se aprecia en la **figura 45**, D. Esto también ocurre en los campos de observación de las muestas P5 y P16.

Respecto a las partículas observadas en la muestra P16, algunas de las más notables presentaron un contorno hexagonal, que sugieren un cuerpo poliedrico para las partículas, con un tamaño de ~10 µm entre los vértices opuestos (**figura 46**, A, B y C). Dichas partículas fueron relativamente abundantes (en comparación con las VLPs observadas en P1) y aunque su contorno conservaba una simetría hexagonal, algunas presentaron aristas irregulares, así como variaciones en el tamaño respecto a las partículas aledañas (**figura 46**, B y C). La observación de estas partículas es relevante ya que su rango de tamaño y forma aproximadamente regular es consistente con lo que se observa en los cuerpos de oclusión de algunos virus cuyos hospederos son insectos (p. ej. reovirus o baculovirus), donde se ha reportado poliedrosis citoplasmática o nuclear.

Por otra parte, en la muestra P16 también se observaron partículas de contorno aproximadamente circular, cuyas superficies sugieren cuerpos esferoides, semejantes a las partículas observadas en la muestra P1 (compárese **figura 45**, C con **figura 46**, D), y con una superficie clara pero con un halo de material denso alrededor. Sin embargo, no se pudo corroborar la similitud de las dimensiones de dichas partículas circulares entre muestras ya que desafortunadamente, no se cuenta con la escala para las micrografías de estas partícualas en la muestra P16.

En cuanto a la muestra P5, se observaron partículas densas (denotadas por el color oscuro que presentan), de contorno aproximadamente circular, lo que sugiere un cuerpo esferoide para dicha partícula; su dimámetro fue de ~500-700 nm (**figura 47,** A y B). Considerando que en la muestra P16 se observaron partículas hexagonales de bordes irregulares, algnas de las cuales asemejaban también a formas ovaladas (**figura 47**, C y D), es posible que las partículas de P5 tengan el mismo origen que las partículas "deformadas" de P16. Por otra parte, es importante mencionar que durante las observaciones de las muestras P16 y P5 al TEM, se encontraban partículas granulares y amorfas en gran parte del campo de observación, similares a las partículas granulares observadas en la muestra P1. Estas partículas granulares amorfas "de fondo" pueden provenir de distintas fuentes, aunque dada su abundancia, puede ser agregados proteícos provenientes del hospedero que no se removieron en el enriquecimiento de VLPs.

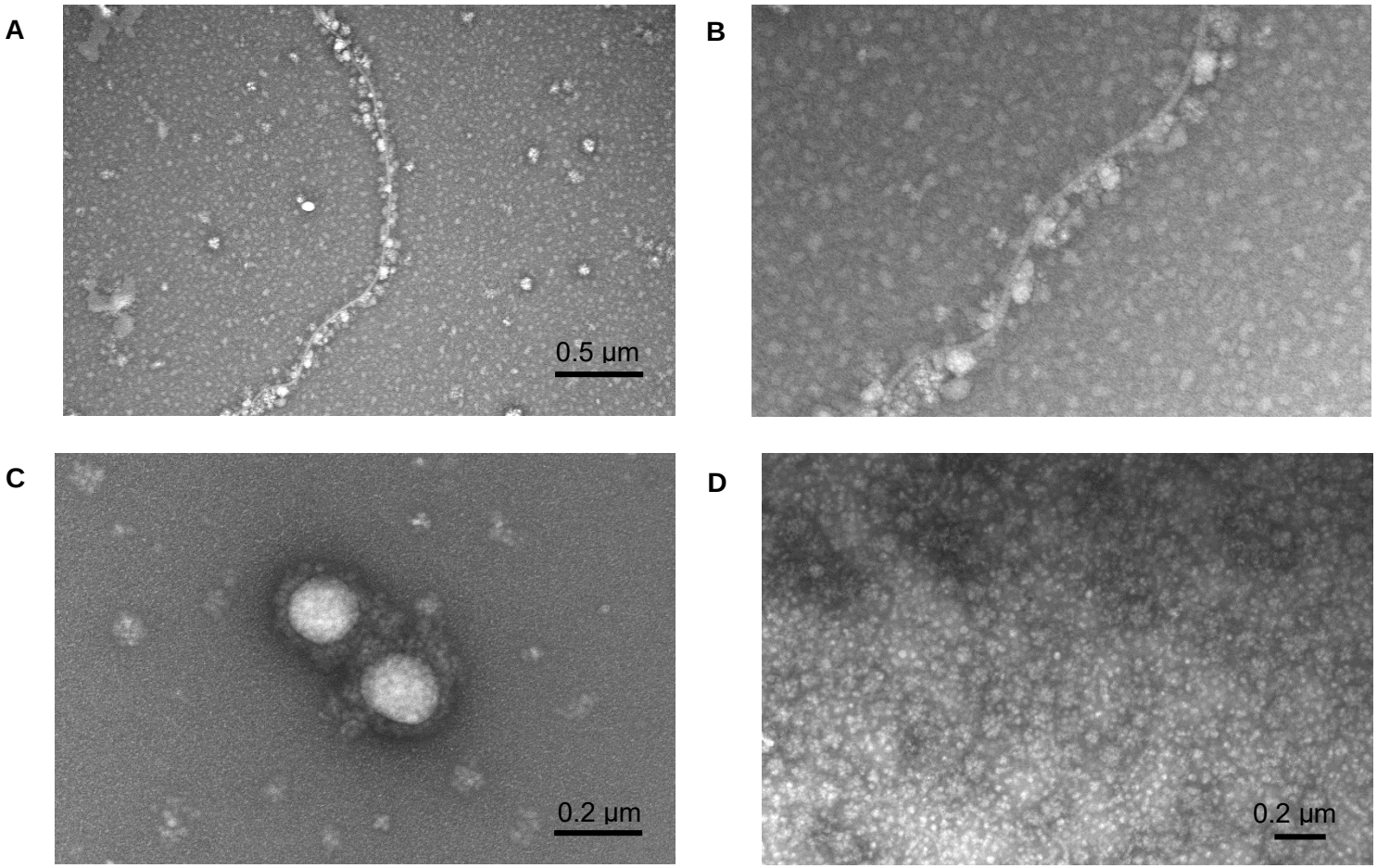

Figura 45. Micrografías de VLPs obtenidas a partir de la muestra P1. (A) Se observó una partícula bastoniforme de ~10 x 2000 nm cubierta por partículas granulares amorfas. (B) Una magnificación del campo de observación anterior donde se aprecia con más detalle la forma irregular de las partículas que se aglomeran sobre la partícula central. (C) Partículas de apariencia esferoide de ~150 nm de diámetro, que se observan cubiertas por material denso (posiblemente las partículas de apariencia granular de fondo). En esta captura el material granular es menos conspicuo en el área circundante a las VLPs respecto a lo que se observa en (A). (D) Se muestra la vista predominante en el campo de observación de la muestra P1, que constó de material granular de tamaño variable; las partículas esferoides o bastoniformes como las mostradas en (A) y (C) se observaron sólo esporádicamente.

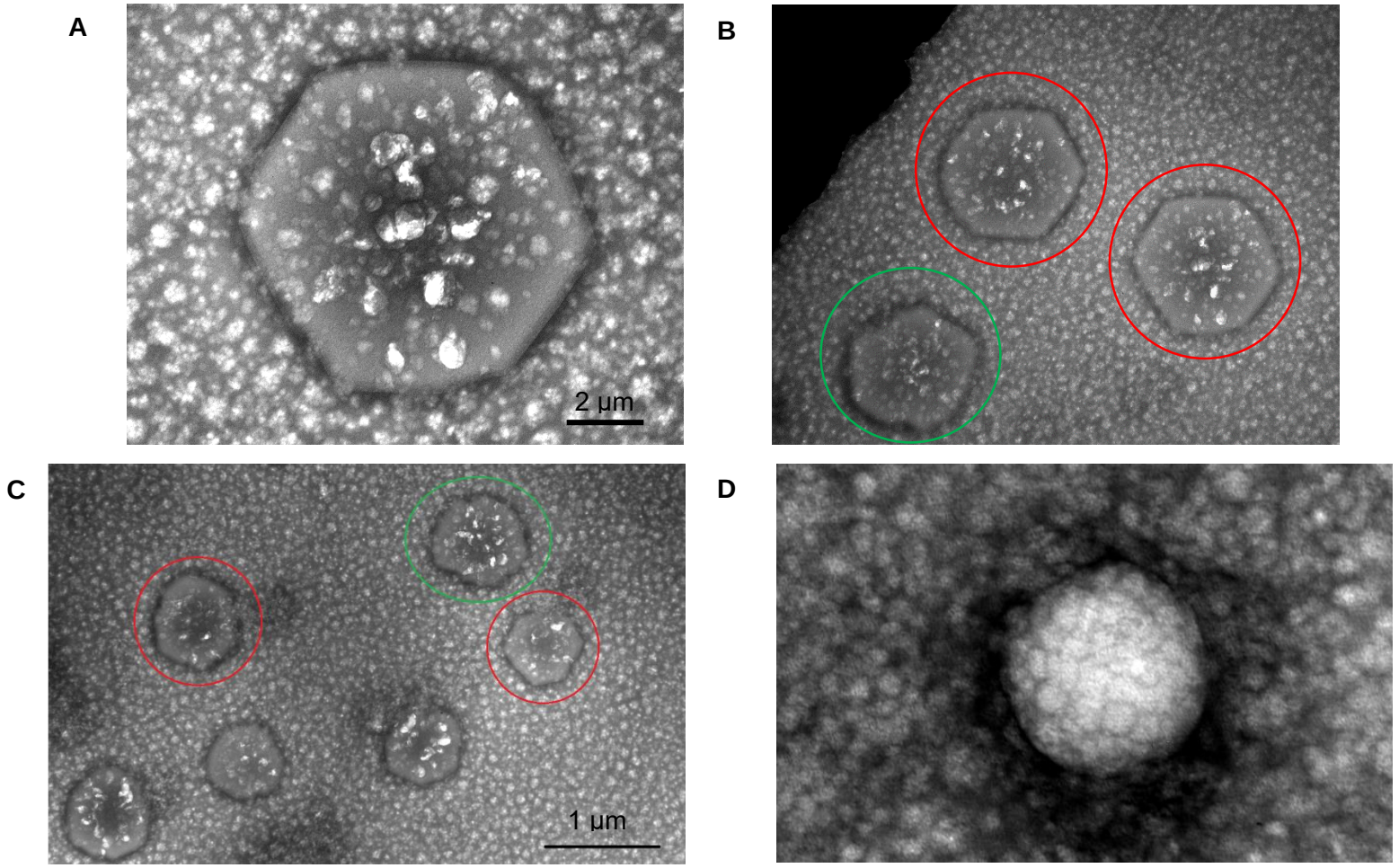

Figura 46. Micrografías de VLPs obtenidas a partir de la muestra P16. (A) Se muestra una de las partículas observadas con contorno hexagonal (una de las más regulares en forma) de ~10 µm entre vértices opuestos. (B) y (C) Algunas partículas presentaban aristas irregulares (circulos verdes) y otras presentaron aristas bien definidas (circulos rojos) de una silueta (casi) hexagonal, además de que presentaron un tamaño menor a la partícula mostrada en (A), con una distancia de < 1 µm entre vértices puestos. (D) Otras partículas observadas en la muestra P16 exhibieron una apariencia “curva” en su superficie que sugiere una simetría esferoide.

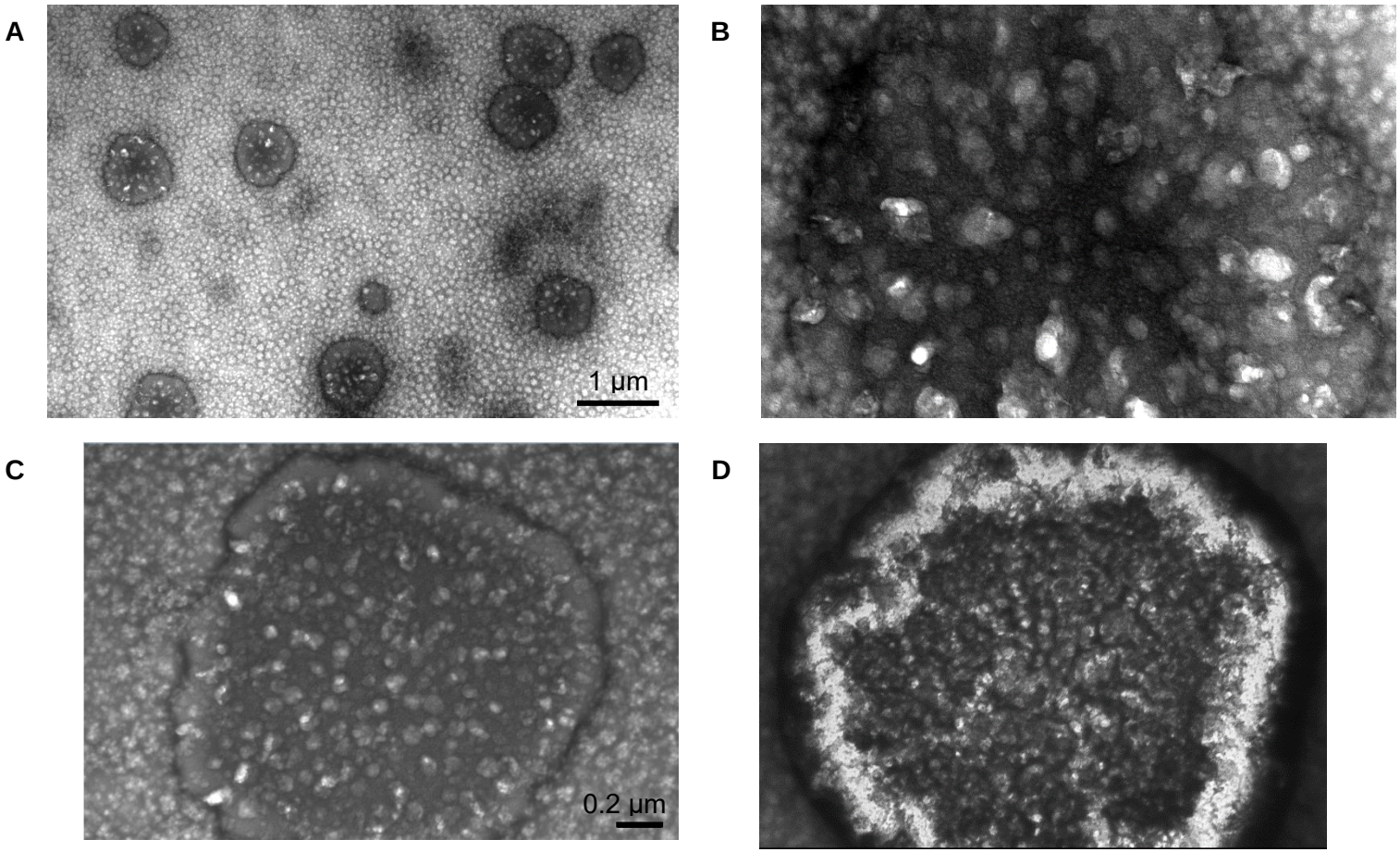

Figura 47. Micrografías de VLPs obtenidas a paritr de las muestras P5 y P16. (A) Se muestran agregados densos dispersos en el campo de observación, delimitados por un contorno circular que representan posibles partículas esferoides de ~500-700 nm. (B) Un acercamiento a los agregados de (A) muestra que el margen oscuro (la "circunferencia") es irregular y se observan partículas granulares sobre el área de mayor densidad. (C) y (D) Magnificación de las partículas observadas en P16 (figura 45) de contorno ovalado, se muestran para su comparación, similares en tamaño y forma a las mostradas en (A) y (B).

Capítulo 4. Discusión

El objetivo central de los estudios de virómica como el que aquí se desarrolla es la exploración de la diversidad viral asociada a una especie de insecto fitófago en particular, lo que representa un primer paso hacia el entendimiento de las interacciones ecológicas y la coevolución que ocurren entre los virus e insectos. Este conocimiento tiene el potencial para derivar, en última instancia, en la identificación y uso de virus en diferentes aplicaciones prácticas. Así, el interés por la identificación de virus con potencial para el biocontrol de plagas de insectos, tanto en el ámbito clínico como agroalimentario, se ha incrementado en años recientes.

En lo concerniente a hemípteros fitófagos que son plaga de cultivos a nivel mundial, los estudios más recientes de virómica se han enfocado principalmente en especies de áfidos (Bonning, 2009; Chen et al., 2012a; Feng et al., 2017; Liu et al., 2016; Yasmin et al., 2020), psílidos (Nouri et al, 2016) o mosquitas blancas (Rosario et al., 2014), mientras que a la fecha, sólo un estudio se ha realizado para elucidar y caracterizar parte del viroma de una especie de piojo harinoso (*Phenacoccus solenopsis*) (Luria et al., 2020). Por lo tanto, un aspecto de principal relevancia del presente trabajo es que representa la primera prospección del viroma de *P. ficus*, una plaga de insecto que es de interés regional y nacional.

La metodología de secuenciación de RNA sin rRNA obtenido a partir de muestras de piojo se seleccionó por varias razones: i) debido a la extensa información reportada sobre la relación estrecha entre diferentes taxa de insectos y los virus de RNA; ii) porque se pueden recuperar secuencias virales putativas de RNA y DNA; iii) es útil para la recuperación de secuencias virales más divergentes y para la reconstrucción de contigs más largos, respecto a lo que se puede obtener con la secuenciación de ácidos nucleicos virales enriquecidos (i.e., secuenciación de RNA de VLPs o de RNAs pequeños).

Así mismo, para maximizar la detección de secuencias virales putativas y para comparar los resultados de diferentes estrategias y herramientas, se utilizó un flujo de análisis bioinformático que incluye algunos de los análisis más importantes para la detección de secuencias virales divergentes, como el análisis de dominios con RPS-BLAST, usado en análisis de metatranscriptomica de invertebrados (Li et al., 2015; Shi et al., 2016). Además, se añadieron algunos otros análisis al flujo de trabajo como el uso de HMMER con las bases de datos de Pfam, VOGs y eggNOG, utilizados en análisis de virómica de insectos.

Cabe mencionar además que la diversidad viral asociada con *P. ficus* presentada aquí es solo un subconjunto de su viroma; se presentan únicamente las secuencias virales putativas que cumplieron los criterios de selección, i.e., aquellas cuya similitud, organización genómica y relaciones filogenéticas indicaron homología con taxa virales que infectan a insectos o a plantas. Además, la prospección del viroma se hizo a nivel regional, por lo que, de acuerdo a estudios de virómica llevados a cabo con diferentes insectos cosmopolitas (Nouri et al., 2016; Ottati et al., 2020) es altamente probable que se añadan miembros nuevos al viroma de *P. ficus* con muestreo a mayor escala geográfica.

Por otra parte, aunque se ha propuesto la inferencia del rango de hospederos con base en la asignación taxonómica de genomas virales putativos, la ausencia de información biológica asociada a las secuencias implica que no se puede garantizar su especificidad por el hospedero de interés, y es posible que procedan de fuentes ambientales como la dieta, contaminantes u otros organismos. Debido a esto, algunos estudios de virómica incorporan dos metodologías independientes (p. ej. secuenciación de RNA de VLPs y RNA total sin rRNA) para una misma muestra, la confirmación experimental de las secuencias por (RT-)PCR y/o el enriquecimiento y observación de VLPs. En este sentido, para reforzar la prospección inicial mostrada aquí, se realizó el diseño de primers para algunas secuencias selectas de RNA (que servirán para futuros análisis de confirmación) y se complementó con el enriquecimiento de VLPs a partir de tejido completo de piojo y la observación al TEM, en un esfuerzo por colectar la mayor cantidad de evidencia posible con aproximaciones dependientes e independientes de secuencias.

4.1. La diversidad viral putativa seleccionada y su posible asociación con *P. ficus*

4.1.1. Diversidad viral putativa de RNA

La mayoría de las secuencias virales putativas de RNA recuperadas están relacionadas con el orden *Picornavirales* y la familia *Reoviridae*, mientras que el resto de las secuencias mostraron mayor similitud con las familias *Rhabdoviridae*, *Tymoviridae* y *Tombusviridae*. Además, la mayoría de las secuencias recuperadas mostró un porcentaje de identidad ≤50% con las secuencias de las bases de datos utilizadas, por lo que pueden representan fragmentos de genomas virales nuevos.

4.1.1.1. Secuencias virales putativas relacionadas con la familia *Dicistroviridae.*

De las 12 secuencias recuperadas que mostraron similitud con los miembros del orden *Picornavirales*, el contig PF_pv_1 tiene una secuencia y organización genómica relacionadas que lo relacionan con los miembros del género *Cripavirus*, de la familia *Dicistroviridae*. Con base en la longitud de secuencia de PF_dv_1 (que representa más del 90% del genoma completo según su asignación taxonómica), se clasifica como "*high quality draft genome*" según los criterios MIUViGs de clasificación de secuencias de virus no cultivados (Roux et al., 2019). Así pues, esta fue una secuencia considerada para el diseño de primers y para la futura confirmación experimental.

La caracterización de las proteínas predichas de PF_dv_1 indicó que el módulo no estructural detectado en la proteína del ORF1 (con el orden HEL-PC3-RdRP) es el mismo que sigue el módulo no estructural canónico de los taxa del orden *Picornavirales* (HEL-VPg-PC3-RdRP), si bien no se detectó una secuencia correspondiente a la proteína VPg de dicistroviridos (Gall et al, 2008). Por su parte, la proteína del ORF2 presentó tres dominios de cápside JRC, lo que corresponde a los 3 de los 4 dominios de cápside presentes en dicistroviridos (Chen et al., 2012a). Tal organización, aunado con la detección de un IRES en la IGR de PF_dv_1, las relaciones filogenéticas y los análisis de similitud a nivel nucleotídico y de aminoácidos, ubican inequívocamente a PF_dv_1 como el genoma de un virus de RNA putativo relacionado con dicistroviridos.

Específicamente, los análisis de similitud indicaron que las secuencias predichas de aminoácidos en ambos ORFs de PF_dv_1 son altamente similares a las proteínas hipotéticas correspondientes de LeDV y PhSoV. El primero de estos es un genoma viral putativo relacionado con dicistrovirus, obtenido de un análisis metagenómico de muestras de la planta silvestre *Gerbera anandria* L. (Zhao et al., sin publicar), colectadas en el canal Zhenjiang de China. Salvo su organización genómica (anotada como completa en el GeneBank) y su relación taxonómica con los cripavirus, no hay información biológica adicional de LeDV. Por su parte, PhSoV representa una nueva especie viral putativa relacionada a dicistrovirus, cuyo genoma se obtuvo a patir de muestras *Hibiscus rosa-sinensis*, una planta ornamental, colectada en Israel central. Posteriormente, se recuperó el genoma de PhSoV a partir de tejido del del piojo de algodón *Phenacoccus solenopsis* (un miembro de la familia Pseudoccidae), que es una plaga de *H. rosa-sinensis.* Los ensayos de infectividad con PhSoV resultaron i) positivos en individuos de *Ph. solenopsis*, ii) negativos en hojas de *H. rosa-sinensis*, y iii) negativos en individuos del áfido *Aphis gossypii*, por lo que se arguye que PhSoV es el primer dicistrovirus específico de una especie de piojo harinoso (Luria et al., 2020).

Tanto las proteínas estructurales (codificadas en el ORF2) como las proteínas no estructurales (codificadas en el ORF1) de PF_dv_1, PhSoV y LeDV son altamente similares entre sí en su secuencia, siendo las estructurales más similares entre sí (89-92%) que las no estructurales (79-83%), y en ambos casos, las proteínas correspondientes de PF_dv_1 y LeDV son más similares entre sí que las de estos con las proteínas correspondientes de PhSoV. La elevada similitud observada entre las proteínas del ORF2 de PF_dv_1, PhSoV y LeDV es de particular importancia, ya que un criterio de demarcación establecido por el ICTV para aislados de una especie de dicistrovirido es una identidad >90% en la secuencia de la proteína estructural (Chen et al., 2012a). Sólo la similitud entre las proteínas del ORF2 de PF_dv_1 y LeDV está por arriba de este umbral (id=92%), no así la similitud entre las proteínas del ORF2 de PF_dv_1 y PhSoV (id=89%), aunque es un valor muy cercano.

En contraste con lo anterior, no hay un criterio de demarcación establecido oficialmente en torno a la similitud de proteínas del ORF1 de dicistroviridos. Por otro lado, considerando los valores de identidad pareada, se observa que la similitud de las proteínas del ORF1 de las especies IAPV y KBV oscila alrededor del valor de similitud entre las proteínas del ORF1 de PF_dv_1, LeDV y PhSoV, si bien la similitud de las proteínas del ORF2 de aquellos es menor (id= ~80%). Pese a dicha similitud relativamente elevada, se ha demostrado que IAPV es serológicamente distinto de KBV o ABPV (Bonning, 2009), lo cual también es considerado un criterio de demarcación de especies de dicistroviridos por el ICTV (Chen et al., 2012a). El caso anterior, donde las propiedades biológicas de dos virus pueden ser diferentes pese a la elevada similitud entre proteínas podría extrapolarse en principio al caso de PF_dv_1, LeDV y PhSoV. Sin embargo, la similitud de las proteínas del ORF2 que muestran PF_dv_1, LeDV y PhSoV, reduce la posibilidad de una diferencia fenotípica (como ocurre con IAPV y KBV).

De las regiones no codificantes de dicistrovirus, aquella que presenta una conservación de estructura más alta es la IGR, que contiene un IRES, una estructura secundaria que permite la traducción independiente de cap-5’ de los mRNAs, y cuyas características también son usadas como criterio taxonómico (Jan, 2006). El MSA de la IGR de cripavirus y otros taxa relacionados mostró que PF_dv_1 presenta las SIC para la formación de las estructuras PK1, PK2 y PK3, así como los nucleótidos conservados de los “Bulges”, SL1 y SL2 y “Pseudo-knot”, PK1, PK2 y PK3 de una IRES de tipo I. Como se mencionó anteriormente, para obtener un MSA de las IGRs que permitiera encontrar todas las SIC de los taxa incluidos se seleccionaron diferentes porciones de las IGRs y/o del ORF2 de PF_dv_1, LeDV y PhSoV; esto indica que existen variaciones en la posición de los nucleotidos correspondientes al IRES en los tres taxa y sugiere a su vez que puede haber variaciones en la posición del codón de inicio de sus respectivas proteínas del ORF2. Lo anterior concuerda con lo observado en el caso de otros dicistroviridos y taxa del orden *Picornavirales*, donde la traducción

del ORF2 comienza en un codón posterior a la primera metionina o el inicio predicho *in silico* para dicho ORF (Moon et al., 1998; van Munster et al., 2002; Yasmin et al., 2020).

En este punto es importante destacar que todos los análisis de secuencia realizados con PF_dv_1, tanto a nivel de aminoácidos como de NCRs, se realizaron para verificar su asignación taxonómica con la familia *Dicistroviridae*, y así poder extrapolar algunas propiedades como el rango de hospederos. La mayoría de las especies reconocidas de dicisitrovirus son, a saber, específicos de hospederos como hemípteros (principalmente áfidos) o himenópteros (como las abejas y hormigas), es decir, son ISVs (Bonning, 2009). En particular, PF_dv_1 se asemeja a dos virus identificados a partir de material vegetal, uno de los cuales (PhSoV) se ha demostrado infectivo y específico para una especie de piojo harinoso. Dicha especificidad es crucial, no solo para PF_dv_1 sino para LeDV, ya que para ninguno existe información biológica asociada a los mismos, más allá de las muestras donde se encontraron (i.e., el organismo asociado). Como se ha demostrado para PhSoV o RhPV, es posible recuperar al virus a partir del tejido vegetal dada la ecología de su hospedero (insectos fitófagos), pero esto no implica que el virus se replique en la planta.

En conjunto, las características observadas en PF_dv_1 sugieren que este es un genoma de dicistrovirus que puede replicarse en *P. ficus*. Sin embargo, dado que la similitud entre las proteínas del ORF2 de PF_dv_1, PhSoV y LeDV está cerca del umbral de similitud establecido por el ICTV para distinguir aislados de una misma especie de dicistrovirido, es difícil definir si dichos taxa son en efecto aislados de una misma especie de virus o representan a especies distintas. Por tanto, es necesario demostrar posteriormente propiedades biológicas distintas entre ellos para considerarlos como especies distintas (por ejemplo, diferencias serológicas o de especificidad respecto a las especies de insectos hospederos); en caso contrario, PhSoV y PF_dv_1 representarían aislados de una misma especie de dicistrovirus, con rango de especificidad de hospederos más amplio que sólo *Ph. solenopsis* (i.e., otras especies de piojo harinoso).

En relación con lo anterior, el hecho de que PhSoV y PF_dv_1 fuesen encontrados en especies distintas de piojo harinoso puede ser en principio un indicador de que dichas secuencias correspondan a especies de virus similares, pero diferentes. No obstante, se ha demostrado que algunos dicistroviridos (como ALPV) pueden infectar a varias especies de insectos, incluso de órdenes diferentes. Otro factor a tomar en cuenta es que la región geográfica donde se encontró a PhSoV es distante respecto a la región correspondiente a PF_dv_1, por lo cual, incluso en escenario donde ambas secuencias representen aislados de un mismo virus, cabría considerarlos como altamente divergentes.

Finalmente, se debe considerar que PF_dv_1 es uno de los contigs con mayor número de reads/Kb mapeados, que puede corresponder a una alta representación de su secuencia en el transcriptoma (de reads no procedentes de rRNA) de *P. ficus* respecto a los otros contig virales putativos encontrados. Esto sugiere a su vez que el virus putativo al que corresponde PF_dv_1 puede replicarse activamente en *P. ficus*. Por otra parte, el hecho de que no se encontraran reads que mapearan a su secuencia en las tres bibliotecas sugiere que se trata de un elemento episomal (no integrado de forma constitutiva en el genoma del hospedero).

4.1.1.2. Secuencias virales putativas relacionadas con la familia *Iflaviridae*

Los contigs PF_ifv_1 a PF_ifv_6 se recuperaron con base en su similitud a taxa virales relacionados con la familia *Iflaviridae*, y a la presencia de dominios conservados específicos de picornavirus en su secuencia.

La asignación taxonómica de estos contigs es relevante debido a que algunos de los iflavirus mejor caracterizados son ISVs de himenópteros (van Oers, 2010), mientras que los taxa relacionados y no clasificados se han asociado mayormente con otros insectos (Shi et al., 2016). A diferencia de PF_dv_1, los contigs ifv (al igual que la mayoría de los contigs recuperados en el presente estudio) representan secuencias virales parciales (de un tamaño < 90% del genoma de iflaviridos), i.e., entran en la categoría de "*genome fragment*" de la clasificación MIUViGs (Roux et al., 2019). Al respecto, es importante mencionar que algunos estudios de virómica se enfocan en recuperar únicamente genomas virales putativos o fragmentos casi completos en los análisis bioinformáticos; se descartan las secuencias genómicas parciales (con hit a secuencias virales) con el fin de evitar elementos endógenos o integrados al genoma del hospedero (EVEs). En el trabajo de Ottati et al. (2020), las secuencias parciales que presentaron ORFs que no abarcaron todo el contig y que codifican genes parciales o completos (estructurales o no estructurales), y/o que presentaron múltiples codones de paro a lo largo de la secuencia, se confirmaron como EVEs mediante análisis de PCR. En dicho trabajo también se descartaron secuencias parciales cuyo ORF no contuvieran un dominio RdRP (i.e., un gen viral marcador) o secuencias parciales que se encontraran presentes en todas las bibliotecas.

Considerando lo anterior, debe observarse que sólo los contigs PF_ifv_2 y PF_ifv_5 contienen ORFs que abarcan toda la secuencia (del primer al último nucleótido), mientras que los ORFs de PF_ifv_3 y PF_ifv_6 abarcan una porción >90% de su respectivo contig y los ORFs de PF_ifv_1 y PF_ifv_4 abarcaban <80% de

su respectivo contig. Esto implica que sólo para los ORFs de PF_ifv_2 y PF_ifv_5 no se encuentran codones de inicio o paro dentro del marco de lectura correspondiente, permitiendo así la continuidad del ORF a lo largo de todo el contig. Dado que los productos proteicos de cinco contigs se alinearon a regiones intermedias de una proteína de referencia (excepto la proteína de PF_ifv_1 cuyo extremo N-terminal coincide aproximadamente con el extremo N-terminal de la referencia en el alineamiento), no se esperaría encontrar codones de paro o inicio dentro del marco de lectura correspondiente al ORF (ver ejemplos A y C del **anexo C**), sino que estos fueran continuos sobre el contig como de hecho ocurre con PF_ifv_2 y PF_ifv_5. Esto es porque los iflavirus codifican sólo una poliproteína a lo largo de su genoma y por ende sólo se presenta un codón de inicio (cerca al extremo 5') y término (cerca del extremo 3') para el ORF único del genoma (las proteínas estructurales y no estructurales se delimitan por sitios de escisión post-traduccionales). Por lo tanto, considerando dicha organización genómica, es improbable que las secuencias parciales de un iflavirus hipotético (como se presume con los contigs ifv) presenten codones de paro o inicio para sus respectivos ORFs, cuando estos no correspondan a las regiones codificantes terminales de la proteína hipotética.

Cabe mencionar que en los seis contigs relacionados a iflavirus se encontraron dominios correspondientes a genes virales marcadores de picornavirus. Cada uno de los productos proteicos de los ORFs de PF_ifv_1 y PF_ifv_2 tuvo hit con el dominio Rhv_like (cd00205), mientras que el producto del ORF de PF_ifv_3 presentó un dominio CRPV_capsid. Tanto Rhv_like como CRPV_capsid son dominios conservados de cápside jelly-roll o JRC (por sus siglas en inglés) que forma una estructura de barril beta (con ocho láminas beta y dos alfa-hélices) en cada uno de los capsómeros de cápside de picornavirus. Aunque las proteínas con dominio JRC corresponden a genes virales exclusivos (Koonin et al., 2006) y ampliamente representados en la virósfera (tanto en virus de RNA como de DNA), la secuencia consenso del dominio Rhv_like y del domino CRPV_capsid está construida considerando sólo secuencias de virus picorna-like, por lo que cabría esperarse cierta afinidad taxonómica hacia este grupo por parte de las secuencias que presentan hit. Todo lo anterior sugiere una función estructural de las proteínas parciales predichas en los contigs anteriores como elementos de cápside.

Por su parte, las proteínas codificadas por los contigs Pf_ifv_4, PF_ifv_5 y PF_ifv_6 presentaron dominios no estructurales marcadores de virus de RNA de cadena simple (si bien no específicamente de picornavirus); el contig PF_ifv_4 presentó un dominio de helicasa RNA_helicase (pfam00910), mientras que los dominios de PF_ifv_5 y PF_ifv_6 correspondieron a RdRP_1 (pfam00680; cd01699). La familia de dominios RNA_helicase está construida predominantemente con helicasas de virus de RNA, aunque también considera secuencias procedentes de organismos celulares. Por su parte, el dominio conservado

de RdRP es ubicuo y exclusivo de los genomas virales de RNA y presenta una alta conservación de motivos catalíticos (pese a la baja similitud de secuencia de los genes de polimerasa), lo que permite detectar secuencias virales (parciales y totales) de RNA altamente divergentes. Lo anterior sugiere entonces una función no estructural de las proteínas parciales predichas para los contigs PF_ifv_4, PF_ifv_5 y PF_ifv_6.

En conjunto, se tiene que los ORFs de los contigs ifv codifican proteínas con dominios específicos de virus de RNA que sugieren funcionalidad replicativa o estructural propias de virus de RNA. Sin embargo, ninguno de los seis contigs se puede vincular a ambas funciones, que distinguen a un virus episomal. Además de que la presencia de dominios estructurales o no estructurales, si bien indica homología con secuencias virales, no descarta la posibilidad de que dichos contigs sean EVEs (lo cual se relaciona con la presencia de codones de paro en el mismo marco de lectura que el ORF que codifica la proteína predicha). Es importante poner de manifiesto que si bien se han reportado que los EVEs encontrados en el genoma de diferentes hospederos son neutrales (i.e., contienen genes que no se expresan), se ha demostrado que algunos casos los EVEs sí se expresan, produciendo RNAs mensajeros para la síntesis de proteínas virales (parciales o defectuosas) como parte de los mecanismos de respuesta antiviral (Holmes, 2011). Por lo tanto, es posible que en los análisis de virómica a patir del RNA total del hospedero puedan recuperarse secuencias que representen a los RNAs de EVEs (ya que son homólogas a las secuencias de virus episomales).

Adicionalmente, es importante considerar que los contigs relacionados a iflavirus presentaron una cobertura mucho menor en reads/Kb respecto a otros contigs seleccionados (como PF_dv_1, PF_mfv_1 o PF_mfv_2). Lo anterior puede deberse, en efecto, a que las secuencias biológicas de la que proceden se encuentren en baja abundancia en la muestra y por ende sean escasos los reads que les correspondan. Alternativamente, el rendimiento de la secuenciación puede ser un factor técnico que afecte la cobertura de las secuencias presentes en menor abundancia relativa (comparada, por ejemplo, con aquellas cuyo genoma se obtuvo casi completo), y por tanto impida la reconstrucción de contigs más largos de lo obtenido. Este es un factor a considerar para todos los contigs que representan secuencias parciales, ya que la secuenciación de fracciones del transcriptoma para virómica es una de las aproximaciones que requiere más alto rendimiento de secuenciación (en comparación de otras técnicas como la secuenciación de RNA/DNA de VLPs o sRNAs), ya que la mayoría de los reads son de origen no viral. En este estudio el rendimiento máximo obtenido (en la biblioteca PV1) fue de 30 millones de reads, mientras que en trabajos como el de Nouri et al. (2016), con una aproximación metodológica similar, el rendimiento de reads por biblioteca de RNA total sin rRNA fue de 50-100 millones.

Por último, cabe mencionar que aunque las regiones de los extremos de los contigs son muy disímiles (lo que impide unirlos para formar una secuencia más larga), todos ellos mostraron una relación filogenética más estrecha con WCV1 o el grupo (WCV1 + NaV) en el mismo contexto de virus relacionados a iflavirus. Una explicación a esto es que las secuencias parciales recuperadas puedan representar variantes genotípicamente cercanas (si proceden de virus episomales), o cuando menos a virus putativos con secuencia y organización genómica similares entre sí. Sin embargo, se necesita de los genomas completos putativos correspondientes a las secuencias ifv para confirmar dicha hipótesis.

4.1.1.3. Secuencias virales putativas relacionadas a *Picornavirales* no clasificados

Los contigs PF_pv_1 a PF_pv_5 (contigs pv) se recuperaron en función de su similitud con un grupo de taxa virales picorna-like especialmente divergentes, relacionados con el orden *Picornavirales,* pero con características distintivas que los separa de las familias incluidas en dicho orden. Los hits de las proteínas de los ORFs de los cinco contigs pv corresponden a las proteínas de virus como TaPV1, ApGlV1, CVT o MaPV. Aunque dichos taxa picorna-like no tienen una caracterización completa, se han asociado con plantas y/o animales, entre ellos insectos (Milusheva et al., 2019; Yasmin et al., 2020). Todos ellos presentan genomas dicistrónicos, con una organización genómica "invertida" respecto a la de dicistrovirus, con el ORF1 correspondiendo a la poliproteína estructural y el ORF2 a la poliproteína no estructural. Dichos ORFs se encuentran separados entre sí y de los extremos del genoma por NCRs, las cuales en algunos casos (como se ha confirmado en ApGlV1), contienen IRES en la NCR 5' y la IGR.

Al igual que los contigs ifv, los contigs pv se clasifican como "*genome fragment*" de la clasificación MIUViGs (Roux et al., 2019). Los análisis de BLASTP mostraron que el alineamiento de los productos proteicos de los ORFs de PF_pv_4 y PF_pv_5, así como del ORF1 de PF_pv_2, no coincide con alguno o ambos extremos de las poliproteínas de referencia de TaPV1, MAV, CVT o ApGlV1. Por otro lado, en cada caso se encontraron codones de paro o inicio para el ORF que codifica la proteína de PF_pv_1, PF_pv_5 y PF_pv_2 (lo que impide la continuidad del ORF a lo largo de su respectivo contig). La información anterior es contrastante ya que el extremo N-terminal de las proteínas de PF_pv_4, PF_pv_5 y del ORF1 de PF_pv_2 se alinean con regiones intermedias de las secuencias de referencia, y por lo tanto se esperaría que el extremo 5' de los ORFs correspondientes a dichos *queries* se extendieran hasta el extremo del contig (lo que no ocurre ya que el ORF está acotado por codones de paro dentro del mismo marco de lectura, ver ejemplo A del **anexo C**). Como se mencionó anteriormente, la presencia de codones de paro/inicio en los ORFs de contigs virales putativos que no corresponden a regiones terminales de la secuencia de referencia puede ser un indicador de EVEs. Sin embargo, el extremo C-terminal de las proteínas del ORF1 de PF_pv_2

y el ORF de PF_pv_5 si coincide con la región C-terminal de sus respectivos *subjects*. En este mismo sentido, la extensión de los ORFs de los contigs PF_pv_1 y PF_pv_3 si es congruente con la región del alineamiento entre sus proteínas y los extremos y/o regiones intermedias de las secuencias de referencia.

Resulta interesante observar que los valores de similitud de la proteína del ORF del contig PF_pv_4, tanto con la porción del dominio proteasa (PC3) de TaPV1 como con la poliproteína-4 de *Tetranychus truncatus* (araña roja), fueron casi iguales. Con una inspección manual del alineamiento se verificó que los motivos de PC3 se encontraron tanto en la secuencia viral como en la del insecto. Además, cabe resaltar que el contig PF_pv_4 fue el único de los cinco contigs pv aquí considerados que obtuvo un hit con secuencias celulares, por lo cual en efecto puede representar un EVE de *T. truncatus*. Sin embargo, esto puede deberse sólo a una alta divergencia de la secuencia viral putativa.

Por otro lado, los productos de los ORFs de los contigs pv (excepto PF_pv_4) también presentaron dominios virales que son característicos de picornavirus o de virus de ssRNA y otras estructuras distintivas de taxa picorna-like. La presencia de dominios para la encapsidación y la replicación propios de virus de RNA sin intermediarios de DNA (cápside JRC, helicasa de virus ssRNA y RdRP_1, en los contigs PF_pv_1, PF_pv_3 y PF_pv_5) sugiere un origen viral, aunque no descarta la posibilidad de ser EVEs. Cabe mencionar que el contig PF_pv_2 resulta de particular interés, ya que contiene dos ORFs, uno de los cuales codifica una proteína hipotética con dos dominios de cápside Rhv_like, que se alineó a una porción de la proteína del ORF1 de CVT que contiene también dos dominios de cápside. Dicho alineamiento es tal que la IGR de PF_pv_2 coincide con el IGR de la referencia, mientras que el ORF2 de PF_pv_2 coincide en posición con el ORF2 de la referencia, si bien la proteína predicha del ORF2 de PF_pv_2 no obtuvo hits en los análisis de BLASTP. En los MSA de las IGRs se observaron las SIC y otras secuencias nucleotídicas conservadas de las IRES tipo 1 en la IGR de PF_pv_2. En conjunto, la presencia de dos ORFs separados por una IGR (con potencial IRES) y la presencia de dominios de función estructural para encapsidación sugieren un origen viral (no de EVEs) para el contig PF_pv_2. Debido a dichas características de secuencia (y a que fue la secuencia de mayor longitud de los contigs pv) se consideró a PF_pv_2 para el diseño de primers y la confirmación experimental.

Finalmente, se debe considerar que los contigs relacionados con virus picorna-like mostraron similitud a un mismo grupo de taxa, por lo cual posiblemente se asocian a genomas virales putativos genotípicamente similares (i.e., aislados de una misma especie). Al igual que con el resto de secuencias parciales, es necesario contar con un contig completo y de confirmación experimental de la secuencia para descartar posibles EVEs, que permita a su vez esclarecer las similitudes entre las secuencias. No obstante, los cinco

contigs presentaron una baja cobertura en reads/Kb respecto a los contigs presumiblemente más completos (PF_dv_1, PF_mfv_1 y PF_mfv_2). Esto es indicador de una baja abundancia de la secuencia biológica de origen en la muestra, lo que a su vez puede limitar la construcción de un contig más completo.

4.1.1.4. Secuencias virales putativas relacionadas a la familia *Reoviridae*

Los contigs PF_rv_1 a PF_rv_8 se recuperaron en función de su similitud con diferentes taxa de la familia *Reoviridae*. Específicamente, los productos proteicos de los ORFs de los contigs rv tuvieron hit con diferentes secuencias de la base de datos de RefSeq asignadas a un mismo txid; las proteínas codificadas en los segmentos del genoma de OpBRV. Dada la naturaleza multipartita de los genomas de reovirus, una estrategia similar se ha utilizado exitosamente en otros análisis de virómica para recuperar fragmentos o genomas completos de nuevos reovirus. Por ejemplo, Nouri et al. (2016) recuperaron la mayoría de los segmentos del genoma de DcRV en función de su similitud con los segmentos del genoma de Nilaparvata lugens reovirus, mientras que Ottati et al. (2020) recuperaron 10 de los 11 segmentos Scaphoideus titanus reo-like virus 1 en función de su similitud con los segmentos de Homalodisca vitripennis reo-like virus.

Por otra parte, la estrategia inicial de búsqueda por similitud debe complementarse examinando las características particulares de los contigs, tanto para determinar si son de origen viral, como para establecer una relación entre ellos como segmentos de un mismo genoma de reovirus putativo, considerando a su vez el número de segmentos que representan del genoma y la completitud de dichos segmentos.

Respecto a sus características individuales, se observó que los ORFs de cinco de los ocho contigs (PF_rv_1-PF_rv_5) recuperados están completos (tiene codon de inicio y término). Por otra parte, en todos los casos los ORFs abarcaron una porción ≥80% de su respectivo contig. Esta característica (i.e., segmentos monocistrónicos cuyos ORFs abarcan una porción ≥90% de su secuencia) también se observan en los segmentos de los genomas de OpBRV, SBV o BatRV (sus mejores hits), lo que sugiere en principio que los ocho contigs se hayan casi completos en su porción codificante.

Adicionalmente, se observó que cinco de los contigs recuperados (y sus respectivas proteínas) tienen un tamaño de secuencia equiparable a su correspondiente secuencia de referencia; solo en el caso de las proteínas predichas para PF_rv_2, PF_rv_3 y PF_rv_8, los análisis de similitud mostraron que el tamaño

del *query* es de aprox. la mitad del tamaño de la correspondiente secuencia de referencia. No obstante, la comparación anterior se puede considerar como un referente de tamaño más que como un criterio de completitud de los contigs. Esto se debe a que, en algunos casos, se ha observado que las proteínas putativas homólogas codificadas en diferentes genomas de reovirus pueden variar en tamaño (y por ende también los tamaños de los segmentos codificantes), sobre todo si son altamente divergentes entre sí; lo cual de hecho ocurre entre los *queries* con sus respectivas secuencias de referencia, ya que la similitud obtenida fue siempre limitada (≥45%), con una cobertura del *query* variable.

Tomando en cuenta la variación de tamaño entre las proteínas *queries* y sus respectivos hits, la anotación funcional de dichas proteínas pudiera considerarse para deducir la completitud de los contigs rv (bajo el supuesto de que, si las proteínas *queries* y *subjects* son homólogas, entonces deberían contener dominios estrucurales o no estructurales similares y a su vez, se podría verificar posteriormente si el dominio está completo o truncado en el *query*). Sin embargo, solo el producto proteico del ORF de PF_rv_1 presentó un dominio de RdRP completo y es de una longitud similar a la RdRP de OpBR, SRV, ECV, BatRV y HDV21. Ningúna otra proteína de los contigs recuperados presentó dominios virales conservados, y las correspondientes secuencias con las que tuvieron hit tampoco se encuentran anotadas; sólo algunas funciones se han sugerido para las proteínas de OpBRV con base en su similitud a secuencias de idnoreovirus u otros géneros de la subfamilia *Spinareovirinae*. Por lo tanto, la anotación funcional no aporta evidencia sustancial que de referencia sobre la completitud de los contigs rv. A propósito de la anotación con RdRP y el tamaño del contig PF_rv_1 (el más grande de los contigs rv), se seleccionó su secuencia para la confirmación experimental posterior.

Siguiendo con la completitud de los contigs, dado que no se observaron las mismas secuencias terminales en sus extremos (una característica altamente conservada en reovirus), es altamente probable que no representen segmentos virales putativos completos (independientemente de la completitud de la secuencia codificante) (Attoui et al., 2011). En principio, la incompletitud puede ser una limitación impuesta porque la secuenciación de extremos de genomas es particularmente complicada. Sin embargo, algunos segmentos de reovirus, como los de BatRV, se han obtenido completos directamente de la secuenciación masiva de metatranscriptomas. Por otra parte, dado que los ocho contigs relacionados a reovirus obtuvieron una baja cobertura en Reads/Kb (en comparación a otras secuencias putativas virales recuperadas como PF_dv_1 o los contigs mfv), cabe la posibilidad que se haya secuenciado y reconstruido solo una porción de las secuencias biológicas de origen.

En conjunto, es importante mencionar que la ausencia de secuencias terminales conservadas en los contigs rv impide confirmar o descartar su relación como segmentos de un mismo genoma viral putativo (y además verificar su completitud), lo cual, aunado a la ausencia de dominios virales marcadores en las secuencias, impide confirmar un origen viral para las mismas.

Respecto al número de contigs virales putativos encontrados, se debe tomar en cuenta que no siempre es posible llevar a cabo la identificación de todos los segmentos de un reovirus basándose unicamente en análisis de similitud, especialmente cuando se trata de secuencias muy divergentes. Por ello, la completitud de los genomas de reovirusn en cuanto al número de segmentos que lo conforman, a menudo se corrobora con base en evidencia independiente de secuencia como los electroferotipos. Este criterio es utilizado por el ICTV para la caracterización de genomas de reovirus, y en el caso de los genomas de DcRV, OpBRV, ObCPV19 u OpCPV18, la caracterización por electroferotipos y la determinación del tamaño de cada uno de los segmentos del genoma se realizó incluso antes de su secuenciación (Attoui et al., 2011; Graham et al., 2006). Sin embargo, se debe tomar en cuenta que para obtener los electroferotipos, frecuentemente es necesario contar con material viral enriquecido (VLPs), lo cual a su vez se relaciona con un titulo viral alto y/o una replicación activa (Belloncik y Mori, 1999); propiedades biológicas que, en el caso de los virus encubiertos, son difíciles de detectar.

Finalmente, en el escenario donde los ocho contigs recuperados son, en efecto, segmentos de un mismo genoma de reovirus episomal, la asignación taxonómica de la RdRP de PF_rv_1 permite hipotetizar acerca del potencial rango de hospederos para el virus putativo. En la filogenia reconstruida con la RdRP de miembros representativos de la subfamilia *Spinareovirinae* y de PF_rv_1, esta última se agrupó en un clado compartido con SRV, ECV, BatRV y HDV21. A su vez, dicho clado se aparta significativamente de los clados que agrupan 7 de los 8 géneros, relacionandose más estrechamente con clado del género *Idnoreovirus*, cuyo miembro es OpBRV. El rango de hospederos de los taxa más estrechamente relacionados con PF_rv_1 es diverso, ya que se han asociado a diferentes artrópodos como dípteros (ECV y HDV21), garrapatas (SBV), o moscas ematófagas (BatRV) (Shi et al., 2016; Medd et al., 2018; Harvey et al., 2019; Temmam et al., 2019). Por su parte, OpBRV se ha encontrado en tejidos de lepidópteros y de avispas parasitoides. Lo anterior indica que el rango de hospederos potenciales para PF_rv_1 (y el genoma putativo que representa) es predominantemente de insectos. No obstante, considerando la filogenia reconstruida, PF_rv_1 es filogenéticamente distante de los géneros de la subfamilia *Spinareovirinae* que agrupan a ISVs

(*Cypovirus*, *Idnoreovirus* y *Dinovernavirus*), pero también de los géneros conocidos de reovirus que infectan insectos y plantas (*Fijivirus* u *Oryzavirus*), o insectos y vertebrados (*Coltivirus*) (Attoui et al., 2011).

4.1.1.5. Secuencias relacionadas a la familia *Rhabdoviridae*

Los análisis de similitud indicaron homología tentativa entre los contigs rbv y HdV2, WIV7 y TbRV1, que están incluidos en el supergrupo dimarhabdovirus, de la familia *Rhabdoviridae*. Específicamente, los contigs rbv tienen ORFs que codifican proteínas con homología a alguna de las proteínas canónicas de rhabdoviridos (P, M, G y L). Respecto a la longitud del genoma monopartita de rhabdoviridos, los contgis rbv se incluyen en la categoría "*genome fragment*" de los criterios MIUViGs (Roux et al., 2019). A partir de estos resultados, se pueden considerar algunos aspectos de relevancia biológica para los contigs rbv.

El supergrupo *Dhimarhabdovirus* fue descrito por Li et al. en 2015 con base en la **monofilia** encontrada entre varios géneros de la familia *Rhabdoviridae* y otros virus de -ssRNA no clasificados que comparten características de secuencia con dichos géneros; el supergrupo contiene (en su mayoría) a virus con insectos hematófagos vectores (mosquitos, moscas, garrapatas, etc.) que infectan vertebrados (mamíferos, aves, reptiles y peces), los cuales divergen ampliamente respecto a otros virus de la familia *Rhabdoviridae* que infectan a vertebrados (*Lyssavirus*) o plantas mediante insectos vectores (*Nucleorhabdovirus*, *Cytorhabdovirus* y *Varicosairus*). La información anterior permite ubicar a las secuencias de los contigs rbv entre taxa cuya característica más frecuente como grupo es que están asociados a insectos pero no a plantas. Además, para el caso específico de PF_rbv_4, la filogenia resultante con secuencias parciales de RdRP lo muestra anidado en las ramas más internas del supergrupo, lo cual sugiere que no es divergente dentro el mismo y que puede compartir algunas de las características más típicas. Por otro lado, aunque los dimarhabdovirus frecuentemente se relacionan a insectos hematófagos como vectores (pero no necesariamente como hospederos), los hospederos más probables de HdV2, WIV7 y TbRV1 (aquellos con los que se relaciona más estrechamente los contigs rbv) son insectos no hematófagos; en el caso de WIV7, los posibles hospederos son el áfido *Hyalopterus pruni* o avispas del género *Aphelinus* (Li et al., 2015), mientras que para HdV2 los hospederos más probables son odonatos (Shi et al., 2016).

Ahora bien, es importante mencionar que, aunque el análisis de similitud permite inferir el origen viral de los contigs rbv (lo que a su vez permite en algunos casos extrapolar ciertas propiedades biológicas), existen análisis previos donde se demuestra que no siempre es posible inferir la presencia de un virus episomal a partir de fragmentos genómicos. Li et al. (2015) detectaron diversos EVEs en los genomas de invertebrados

(entre ellos algunos hemípteros) que presentaron homología con algunas secuencias de virus incluidos en el supergrupo dimarhabdovirus. Estos EVEs presentan la característica distintiva de contener secuencias codificantes (ORFs) parciales o completas de uno o más genes de proteínas canónicas de rabdoviridos y otros virus de -ssRNA, pero sin contener una secuencia completa, i.e., como la de sus homólogos de secuencia que si contienen todos los genes constitutivos de virus episomales. Adicionalmente, se demostró que las regiones flanqueantes de dichas secuencias eran características de elementos transponibles.

Al considerar los alineamientos se observó que el producto del ORF de PF_rbv_1 y la proteína N de WIV7 coinciden en su extremo C-terminal, mientras que en el extremo N-terminal del *query* coincide con una región intermedia de dicho *subject* (lo cual es congruente con un ORF sin codón de inicio en el contig; ver **anexo C**, ejemplo D). Por su parte, en el alineamiento de los productos de los ORF1 y 2 de PF_rbv_2 se observa que coinciden, respectivamente, con el término y comienzo de las proteínas M y G de WIV7. No obstante, debe notarse que los ORFs de dicho *query* están completos (tienen codones de inicio y término) y ninguo de sus productos es de tamaño similar a su respectivo *subject*. Respecto al producto proteico del ORF de PF_rbv_3, en el alineamiento se observa que el extremo C-terminal del *query* está cerca del término de la proteína G de WIV7 (coincidiendo así la términación de los ORFs de *query* y *subject*), mientras que el extremo N-terminal del *query* se alinea con una región intermedia de dicho *subject* (lo cual contrasta con un ORF que está acotado en la secuencia de PF_rbv_3 por codones de paro dentro del mismo marco de lectura); se esperaría que el ORF se extendiera de tal manera que coincidiera el inicio de los ORFs de *query* y *subject*). Por otra parte, río abajo del ORF de PF_rbv_3 no hay un segundo ORF que coincida con el inicio del ORF de la proteína L de WIV7. Sólo en el caso de la secuencia PF_rbv_4, el ORF es contínuo en el contig y se alinea con una región intermedia de la proteína L de WIV7.

En conjunto, los análisis de secuencia indican que los contigs PF_rbv_2 y PF_rbv_3, que contienen únicamente genes parciales (i.e., codificados en ORFs terminados) de virus de RNA, pueden ser EVEs que derivaron de secuencias virales exógenas (de ahí su homología con rhabdoviridos). Con base en esto, es importante considerar el análisis futuro las regiones flanqueantes y de EVEs usando una estrategia similar a la usada por Li et al. (2015), donde se alinearon las secuencias de proteínas con tBLASTN a genomas de referencia de artrópodos de la base de datos de referencia del NCBI y la base de datos *Whole Genome Shotgun Database*, así como una inspección de las regiones flanqueantes. En la búsqueda de secuencias con HMMER usando la base de datos GyDB, ninguno de los productos proteicos de los contigs rbv obtuvo hit con secuencias de elementos móviles. Sin embargo, cabe considerar que dicha base de datos sólo permite detectar homología con secuencias de proteína involucradas en el ciclo de replicación de los

elementos transponibles (Llorens et al., 2011), por lo que detectar un *query* con homología a secuencias virales (aunque esté codificada/integrada en el genoma del hospedero) queda fuera del alcance de la base de datos.

Cabe mencionar que dadas las características de secuencia de PF_rbv_4, así como a la presencia de una región con bases indeterminadas, se contempló su secuencia para el diseño de primers que permitan la amplificación de la región que incluye dichas bases (**anexo D**).

4.1.1.6. Secuencias relacionadas con la familia *Tombusviridae*

La secuencia completa de PF_tbv_1 (2720 nt) representa más del 50% de la longitud de los taxa de referencia WSV9 (cuyo genoma es de 4059 nt) y Sanxia water strider virus 14 (SWSV14; con genoma de 4175 nt), por lo cual se clasifica en la categoría "*genome fragment*" de los criterios MIUViGs (Roux et al., 2019). Aunque representa una secuencia viral putativa parcial, sus ORFs no están truncados. Considerando que el contig PF_tbv_1 fue el único con homología a la taxa no clasificados y relacionados con familia *Tombusviridae*, así como la completitud de este repecto a las secuencias de referencia, se consideró su secuencia para la confirmación experimental.

Los análisis de similitud con la proteína predicha para el ORF1 de PF_tbv_1 indicaron que su relación más estrecha es con taxa de virus asociados a insectos (como StV5) u otros artrópodos (como WSV9 o SWSV14), pese a que todos estos se relacionan a su vez con la familia *Tombusviridae*, cuyo rango de hospederos abarca únicamente plantas (Rochon et al., 2012). En correspondencia, la filogenia reconstruida con RdRP muestra que PF_tbv_1 y los taxa más estrechamente relacionados se apartan de los clados que representan los diferentes miembros de los géneros de tombusviridos incluídos. El agrupamiento obtenido en la filogenia anterior es similar, aunque no equivalente, al que presenta el clado "Tombus-Noda" recontruido por Shi et al. (2016), donde los taxa de tombusviridos se agrupan en clados más derivados, que se apartan de los taxa relacionados, no clasificados, agrupados en clados más basales. Por otra parte, se observó que la proteína estructural predicha para PF_tbv_1 (del ORF2) presentó homología tentativa con las proteínas de virus como ApGlV2 o HpV4, ambos relacionados con la familia *Permutotetraviridae*. Estos últimos (incluyendo a ApGlV2 y HpV4) se caracterizan por tener una RdRP con el orden de los motivos catalíticos permutados (del orden canónico de A-B-C a C-A-B), por lo que representan un grupo especialmente divergente de virus de RNA (Gorbalenya et al., 2002; Liu et al., 2016). Sin embargo, sus respectivas

proteínas de cápside muestran similitud con aquellas de tombusviridos (Liu et al., 2016). La RdRP predicha para PF_tbv_1 no presentó motivos permutados, por lo que no cumple con la característica típica de los permutotetraviridos, lo cual sugiere que la asignación con taxa de virus relacionada a tombusviridos es la más adecuada. Además, se ha argumentado que algunos virus como ApGlV2 pueden codificar una proteína no estructural con características de secuencia similares a las de virus que infectan animales (como los permutotetraviridos), mientras que su proteína estructural puede relacionarse más estrechamente con virus que infectan plantas (como los tombusviridos), debido a posibles eventos de recombinación (Liu et al., 2016).

En conjunto, se tiene que PF_tbv_1 puede representar un virus de insecto (dada su similitud con virus no clasificados asociados a insectos), pero con características de secuencia similares a taxa de virus que infectan plantas. Los análisis de mapeo indicaron que la cobertura (en reads/Kb) de la secuencia de PF_tbv_1 es del mismo orden de magnitud que para otras secuencias como los contigs pv, ifv, rbv y rv, lo que al final indicaría una baja abundancia de secuencias de origen.

4.1.1.7. Secuencias relacionadas con la familia *Tymoviridae*

PF_mfv_1 y PF_mfv_2 son contigs cuya secuencia y organización genómica está relacionada con los miembros del género *Marafivirus* de la familia *Tymoviridae*. Las características de secuencia de PF_mfv_1 lo identifican como un aislado de GSyV-1, con una longitud de aprox. 90% del genoma de referencia. Por su parte, PF_mfv_2 representa un aislado, con secuencia genomica parcial de ~64%, de GRVFV. Lo anterior ubica a PF_mfv_1 en la categoría de "*high quality draft genome*" y a PF_mfv_2 en la categoría de "*genome fragment*", de los criterios de clasificación MIUViGs (Roux et al., 2019).

A diferencia del resto de las secuencias virales de RNA putativas recuperadas, y dada la elevada similitud de secuencia que presentaron los contigs mfv con sus respectivas referencias, los análisis realizados estuvieron encaminados a una evaluación de los contigs mfv en el contexto de los aislados de GSyV-1 y GRVFV ya reportados, y no a inferir un origen viral y propiedades biológicas putativas a partir de las secuencias de referencia. Para esto, se utilizaron algunas de las secuencias completas reportadas de GRVFV y GSyV-1 de distintas regiones del mundo, a las que se han realizado caracterización de secuencia.

Los análisis de identidad pareada mostraron que en el contexto de los aislados de GSyV-1, las secuencias de PF_mfv_1 y BS3 son especialmente divergentes, con diferencias de hasta 17% y 7% del resto de los aislados. Estos resultados son consistentes con los análisis previos a nivel de nt de 8 aislados (2 de USA, 2 de Eslovaquia, 2 de República Checa y 2 de Brasil), donde se demostró que el aislado brasileño BS3 era el más divergente de todos, con diferencias de hasta 16-18% con el resto de los aislados (Sabanadzovic et al., 2017). A su vez, respecto a los aislados europeos (RC1, ES1 y ES2) y de EUA (EU1 y EU3) de GSyV-1 analizados, los resultados concuerdan con lo obtenido previamente por Glasa et al. (2015), donde se observó que dichos aislados europeos difieren entre sí hasta 7%, mientras que los aislados de EUA divergían entre sí en apenas 1.2%. Lo anterior es relevante dado que al momento del estudio de Glasa et al., los asilados europeos representaban las secuencias completas disponibles más divergentes. Otro aspecto notable respecto a la información disponible al momento de aislados de GSyV-1 es la ausencia de una correlación clara entre su divergencia y la región geográfica de origen. Además de los valores de identidad mencionados, los análisis filogenéticos de Glasa et al. realizados con secuencias parciales y completas de nt no muestran una agrupación de los aislados de regiones diferentes en clados monofiléticos (salvo en el caso donde se reconstruye una filogenia con una porción parcial codificante de metil-transferasa, pero la monofilia solo se observa para los aislados de EUA). Por otra parte, dichos autores también encontraron que una de las regiones más variables del genoma de GSyV-1 es la región proximal al extremo 5', lo cual es relevante considerar con la secuencia de PF_mfv_1, que presentó una extensión de la secuencia del ORF1 y por ende de la proteína predicha, lo que no se observa en ningún otro aislado.

Respecto al análisis de similitud de PF_mfv_2 con el resto de los aislados de GRVFV, se obtuvieron resultados similares a los reportados recientemente por Wu et al. (2020). Estos autores presentaron el mayor compendio de secuencias de GRVFV hasta el momento, con 12 aislados (8 del occidente y 4 del sur de Australia) de secuencias completa o casi completa, algunos de los cuales son altamente divergentes entre sí y respecto a otros aislados previamente reportados. Los resultados obtenidos en el presente trabajo mostraron que en el contexto de los aislados de GRVFV, PF_mfv_2 es muy similar (con identidades de 82%-85% en genoma y 95-96% en proteína) a secuencias de Australia, Suiza, Francia y Nueva Zelanda (S1, A1-A6, F1 y NZ1) previamente reportadas, con la particularidad de que no se observó alguna relación con la proximidad geográfica entre ellos. Así, de acuerdo con el análisis robusto de Wu et al., otros aislados de Australia, Nueva Zelanda y EUA (A7, A8, NZ2 y EU1), continúan siendo al momento los más divergentes de todos los aislados, con similitudes de máximo 79% en genoma y 90% en proteína respecto al resto de los aislados previamente reportados o los encontrados por dichos autores. Al respecto, es importante observar que PF_mfv_2 fue la única secuencia incluida en el presente análisis que mostró valores

ligeramente más elevados de similitud con A7, A8, NZ2 y EU1, respecto al resto de los aislados, con valores de hasta 80% de similitud en genoma y 91% en proteína. Lo anterior sugiere que PF_mfv_2 puede representar una secuencia "intermedia" en las secuencias analizadas. Sin embargo, sigue siendo más similar a los otros aislados (i.e., aquellos de los cuales A7, A8, NZ2 y EU1 divergen ampliamente). Por otra parte, es importante considerar que la secuencia de PF_mfv_2 sólo es parcial respecto al resto de los aislados, por lo que no es posible determinar una divergencia concluyente.

Además del análisis de secuencia, es importante destacar algunos aspectos de relevancia biológica respecto a la detección de aislados de GSyV-1 y GRVFV. De acuerdo con la revisión de Sabanadzovic et al. (2017) de virus asociados a la vid, hasta recientemente Brasil y Chile eran los únicos países de la región de América latina donde se había documentado la presencia de GRVFV y/o GsyV-1. Por otra parte, en EUA se habían registrado aislados de GRVFV en California y de GSyV-1 California, Mississippi, Washington, Carolina del Norte o Nueva York. Pese a la cercanía geográfica con esta región, actualmente no hay secuencias registradas en el Genebank de aislados encontrados en México. Si bien las infecciones de GRVFV no se ha asociado con efectos patogénicos severos y GSyV-1 no se ha asociado contundentemente con alguna enfermedad, el diagnóstico de estos virus se ha realizado en diferentes regiones del mundo como parte del monitoreo de agentes asociados a distintas enfermedades de la vid. Los aislados de GRVFV y GsyV-1 se han identificado a parir de material vegetativo de la planta con distintos síntomas de enfermedad como clorosis de las venas de las hojas, aunque frecuentemente se recuperan sólo secuencias parciales del genoma (Glasa et al., 2015; Wu et al., 2020; Xiao y Meng, 2016).

Considerando lo anterior, la detección de aislados de GRVFV y GsyV-1 en las bibliotecas de RNA de *P. ficus* analizadas no solo representa el primer registro de secuencias de estos virus para México y Baja California, sino la primera detección realizada a partir de tejido de insectos por análisis no dirigidos. Esto último es de particular importancia ya que no se conocen insectos vectores asociados a estos virus. Si bien con anterioridad se han detectado secuencias de GsyV-1 en el saltamontes *Erythroneura variabilis* mediante RT-PCR (Rwahnih et al., 2009), no se ha demostrado la transmisión en especies de insectos (tampoco para GRVFV) (Sabanadzovic et al., 2017). La generación de una secuencia casi completa de GsyV-1 y una secuencia relativamente larga de GRVFV (en relación con la mayoría de las secuencias registradas en el Genebank), la alta abundancia de reads mapeados a dichos contigs (que obtuvieron la mayor cobertura de todas las secuencias recuperadas), así como el número de contigs ensamblados con hit a GsyV-1 y GRVFV (19 y 65, respectivamente), sugieren fuertemente que dichas secuencias no proceden de una fuente ambiental, esto es, indican una relación directa entre estos virus y *P. ficus*.

Lo anterior es consistente con las caracteisticas ecológicas tanto del grupo de virus GFkV-like (que incluye a GRVFV y GsyV-1) como de *P. ficus*. Por ejemplo, autores como Sabanadzovic et al. (2017) argumentan que, dado el nicho ecológico común entre los virus GfkV-like y otros virus de vid como ampelovirus o vitivirus (tubos cribosos del floema), es probable que ocurra una transmisión de todos ellos por los vectores de estos últimos, i.e., diversos géneros de piojo harinoso y otros insectos que se alimentan del floema de la planta (Maliogka et al., 2015). Además, se sabe que algunos marafivirus son transmitidos de forma específica (persistente-propagativa) por cicadélidos y membrácidos (King et al., 2015). Así pues, el floema de la planta representa una región de encuentro entre virus e insectos fitófagos, por lo que es plausible que estos últimos pueden efectuar la transmisión de los virus, incluso sin que ocurran interacciones específicas.

Finalmente, es importante mencionar que la recuperación y análisis de los contigs mfv (los que presentaron la longitud de secuencia más larga de todos los contigs con similitud a GsyV-1 y GRVFV) se realizó como parte de la prospección del viroma de *P. ficus*. Los análisis realizados aquí se enfocaron únicamente a la evaluación *in silico* de sus secuencias. Sin embargo, el análisis conjunto de todos los contigs recuperados con similitud a GsyV-1 y GRVFV es un aspecto importante de la caracterización de aislados de virus de planta, pero quedan fuera de los objetivos de este trabajo. Sin embargo, dada la alta representatividad de los contigs mfv y sus características de secuencia, se utilizó su secuencia para el diseño de primers y confirmación posterior por RT-PCR.

Considerando lo anterior, como parte de futuros análisis de los aislados de GsyV-1 y GRVFV en la región de Baja California, debe contemplarse un análisis más robusto, que permita evaluar todos los contigs recuperados con similitud elevada a dichos virus, así como aspectos del diagnóstico de estos virus en muestras de planta, la evaluación de síntomatología de alguna enfermedad asociada al complejo fleck (al que se asocian los virus GFkV-like) en los individuos donde se realiza el diagnóstico, y establecer una región geográfica donde se detecten dichos aislados.

4.1.2. Diversidad viral putativa de DNA

La búsqueda de secuencias de virus de DNA putativos de los grupos comúnmente asociados a insectos, o caracterizados como ISVs (todos ellos con genoma de dsDNA), se realizó de forma complementaria a la búsqueda de secuencias virales de RNA. Al utilizar una secuenciación basada en RNA (usando RNA total

del insecto), la identificación de secuencias virales de dsDNA se realizó bajo la premisa de buscar únicamente transcritos de dichos genomas, esto es, de fragmentos representativos (puesto que no todo el genoma es transcrito a RNA). Las secuencias seleccionadas, aquellas con similitud a secuencias de virus de dsDNA (tabla 13), son representativas de un pequeño conjunto de ~200 secuencias que no presentaron similitud con secuencias de organismos celulares.

Es importante mencionar que el grupo de familias de virus de dsDNA en las que se concentró la búsqueda, se definió considerando que el rango de hospderos incluyera a insectos y/o ISVs. No obstante, sólo se ha reportado un estudio de iridoviridos encontrados en hemípteros (Williams, 2008). Por su parte, Williams et al. (2017) reportan que se han confirmado infecciones de baculoviridos en hemípteros, aunque el perfil taxonómico (más reciente) de dicho grupo no incluye a hemípteros en el rango de hospederos (Harrison et al., 2018). El grupo de polydnaviridos se incluyó considerando la aplicación de control biológico con la avispa parasitoide *Anagyrus pseudococci* en las zonas de muestro de *P. ficus* (comunicación personal del Ing. P. López, CESVBC, 2020), aunque esta no pertenence a las familias de parasitoides previamente reportadas como simbiogénicas con polydnaviridos (Renault, 2012). Por otra parte, hasta el momento no se han reportado ascoviridos o poxviridos previamente encontrados en hemípteros (Asgari et al., 2017; Williams et al., 2017). Al considerar la asociación biológica de dichos virus se puede explicar en parte la escasa información rescatada respecto a dichos taxa.

Inicialmente, durante la inspección se realizó una búsqueda enfocada en identificar secuencias con homología remota a genes conservados de cada uno de los taxa de dsDNA de interés usando una lista de OGs marcadores propuesta por Roux et al. (2019) a partir de la base de datos VOGDB. Esta lista considera algunos de los genes esenciales (conservados) en diferentes familias de virus de RNA y DNA, que se expresan durante los ciclos de replicación. Sin embargo, se observaron escasos hits a dicho subgrupo de OGs de dsDNA, lo que sugiere una baja representación en la muestra de secuencias que permitan inferir colectivamente la presencia/replicación de algún virus de dsDNA putativo de interés.

Por otra parte, la comparación de los resultados de similitud indicó que de la totalidad de secuencias con hit a OGs de dsDNA y/o secuencias de virus de dsDNA de RefSeq, la gran mayoría mostró homología tentativa con secuencias no virales de la base de datos NR, con identidad más alta y en porciones más largas de los *queries*. Este resultado fue de particular importancia puesto que permitió evaluar más ampliamente todas las secuencias que en principio pudiesen tomarse como virales putativas, considerando sólo sus hits con las bases de datos especializadas en virus. Esto pone de manifiesto que la selección de secuencias virales de dsDNA debe ser altamente específica (enfocándose en genes

marcadores/conservados, como se hizo en un principio), puesto que muchas secuencias con homología tentativa a secuencias virales también lo pueden ser a secuencias de organismos celulares (Hughes y Friedman, 2003; Hughes y Friedman, 2005). De hecho, se ha documentado la adquisición independiente, mediante transferencia horizontal, de genes del hospedero por parte de diferentes virus de dsDNA (como los de baculoviridos o poxviridos) (Koonin et al., 2015; Thézé et al., 2015).

En correspondencia con lo anterior, se debe destacar que el análisis de dominios fue menos efectivo en el caso de la selección de secuencias virales putativas de dsDNA respecto a la selección de aquellas relacionadas a virus de RNA. Al inspeccionar la descripción de algunos dominios con los que las secuencias *queries* tuvieron hit (con palabras clave como “baculovirus” o “poxvirus”), se verificó que dichos dominios se construyeron incluyendo secuencias de organismos celulares, esto es, con secuencias homólogas a los virus de dsDNA. Esto no ocurre en la mayoría de los dominios conservados que se encontraron en la selección de virus de RNA, los cuales (como los dominios de RdRP, cápside, helicasa, etc.) están construidos únicamente con secuencias virales. De acuerdo con Koonin et al. (2015), lo anterior es consecuencia del reducido número de genes, altamente conservados (como los módulos no estructurales y estructurales de virus picorna-like) y de origen independiente respecto a los genes de organismos celulares que presentan los virus de RNA. Esto contrasta con el origen común de distintos genes o genomas de dsDNA y elementos del genoma de organismos celulares (Krupovic y Koonin, 2015; Koonin et al., 2020).

4.1.3. Enriquecimiento y observación de VLPs

El enriquecimento de VLPs y las observaciones por microscopía electrónica son análisis de importancia fundamental en la confirmación y/o caracterización completa de especies virales episomales putativas. Los enriquecimientos de VLPs no sólo proporcionan material útil para las observaciones de partículas (que complementan el análisis de secuenciación masiva), sino que ayudan al enriquecimiento de ácidos nucleicos de una amplia gama de virus (Conceição-Neto et al., 2015; Roossnick, 2015). Por ejemplo, en el trabajo de Luria et al. (2020), donde el análisis bioinformático permitió identificar a PhSoV (el dicistrovirido encontrado en el piojo harinoso *Phenacoccus solenopsis*), se realizó adicionalmente el enriquecimiento y observación de VLPs. Con esto, se logró encontrar partículas de simetría esférica de ~30nm, lo que coincide en forma y tamaño con los viriones de virus picorna-like. Más aún, la extracción de RNA de los concentrados de VLPs y su electroforesis demostraron que el material más abundante era una molécula

de ~10 Kb, lo que confirmó la naturaleza viral de las VLPs, y específicamente, como partículas de virus picorna-like.

De forma análoga a los análisis anteriores, en el presente trabajo se intentó relacionar la información obtenida de VLPs con la información extraida de trabajo bioinformático. Las observaciones de VLPs permitieron detectar en su mayoría a partículas de tamaño y forma similar a las estructuras derivadas de algunos virus específicos de insectos, particularmente de aquellos que producen diferentes tipos de polihedrosis en el hospedero (Belloncik y Mori, 1998, Attoui et al., 2011, Harrison et al., 2018). Sin embargo, dicha información no se pudo vincular contundentemente con aquellos grupos de virus de RNA que se encontraron como mejor representados en el análisis de secuencias (los virus putativos de la familia *Dicistroviridae* y *Tymoviridae*). Para poner lo anterior de manifiesto, deben considerse algunas de sus características biológicas relevantes.

Los dicistroviridos y tymoviridos son virus de +ssRNA que forman viriones de simetría icosahédrica (aunque de apariencia esférica) de ~30 nm de diámetro. Los virus de dichos grupos no forman macroestructuras cristalinas como las que se ha reportado para virus de dsDNA (Baculo, Irido, Pox), además de que sus viriones son de hasta un orden de magnitud más pequeños. Por otra parte, considerando específicamente a los miembros del género *Marafivirus* (cuyas secuencias se encontraron mejor representadas por los reads utilizados en el análisis bioinformáticos), no se conoce vector específico ni se ha demostrado que se puedan replicar en algún insecto (a excepción maize rayado fino virus, oat blue dwarf virus y Bermuda grass etched-line virus, que se replican de forma persistente propagativa en hemípteros) (King et al., 2011). Lo anterior es importante ya que entre menos específica sea la interacción entre un virus de planta con su potencial vector, el tiempo de asociación entre virus e insecto (y por tanto el título viral en los tejidos del vector) se reduce (Whitfield y Rotenberg, 2015). Así pues, la recuperación de viriones de virus de planta como GRVFV y GSyV-1, o incluso de otros virus de planta, a patir de tejidos de insectos fitófagos como *P. ficus* puede verse limitada si no ocurren interacciones persistentes (en perdiodos largos) y/o propagativas (donde el virus se replica en el insecto). Más aún, dado que GRVFV y GSyV-1 no causan patologías severas en la planta o no se han asociado con alguna (Martelli, 2017), el título viral en la planta (y por tanto el título en que los viriones son adquiridos por el insecto) también puede ser bajo. En contraste con lo anterior, aunque se ha reportado que los dicistroviridos sí se replican en sus respectivos insectos hospederos, en algunas ocasiones las infecciones que producen no son abiertas y por tanto los títulos de partículas virales no son elevados (Bonning, 2009).

Tomando en cuenta lo anterior, la ausencia de VLPs con características similares a los viriones de virus de ssRNA puede deberse a múltiples causas de índole biológica, aunque también a razones técnicas como el considerar sólo tres sitios en el enriquecimiento de VLPs (mientras en la construcción de bibliotecas se utilizaron muestras de 14 sitios distintos). Por lo tanto, a continuación se discuten algunas observaciones de VLPs más notables obtenidas y se contrastan con observaciones previas de microscopía electrónica. Cuando es posible, la información de secuencias se coteja con las observaciones de VLPs.

4.1.3.1. Observación de VLPs al TEM

Algunas partículas observadas al TEM son semejantes en tamaño y forma a los OBs producidos por virus cuyas infecciones se distinguen por causar polihedrosis citoplasmática (*Cypovirus*) o nuclear (*Baculoviridae*) (Belloncik y Mori, 1998; Cory y Myers, 2003; Attoui et al., 2011; Harrison et al., 2018). Por ejemplo, los cuerpos polihedricos producidos por Bombix mori nucleopolihedrovirus (BmNPV) en el hospedero *B. mori* reportados por Hong et al. (2000) presentan una forma icosahédrica en observaciones con microscopia electrónica de barrido (i.e., se observa una imagen tridimensional, a diferencia de las observaciones al TEM que proveen una imagen en dos dimensiones). En algunos casos, las aristas de dichas partículas se observan desde una vista superior como un cuerpo hexagonal (aunque en una perspectiva planar se observa con mayor frecuencia contornos octagonales incluso irregulares) de ~1-3µm de longitud. Este tipo de simetría es similar, aunque de menor tamaño, a la observada en las VLPs de las muestras de P16 (**figura 45**, A, B y C).

Adicionalmente, las observaciones de microscopía de barrido obtenidas por Horta et al. (2018) muestran que los OBs producidos durante la infección de cypovirus 14 en el hospedero *Thyrinteina arnobia* (Lepidoptera: Geometridae) son de forma icosahédrica (no de la forma cúbica típica de los OBs producidos por cypovirus). Dichos autores también obtuvieron observaciones al TEM de secciones tranversales de los OBs icosahedricos, donde se aprecia en un plano el contorno hexagonal de los OBs que contienen a los viriones en su interior. Al respecto, es interesante notar que las VLPs observadas en la muestra P16 asemejan a los OBs como los de cypovirus 14 en su simetría planar (hexagonal) y tamaño (1-2 µm). No obstante, estos últimos se observan como cuerpos densos oscuros (después de la tinción con acetato de uranilo) mientras que las VLPs de P16 presentan una apariencia traslucida (pese a que fueron teñidas con el mismo reactivo). Además, no es posible observar si las VLPs de P16 obtenidas presentan algún contenido ya que la imagen sólo es planar de superficie y no una sección transversal.

Por otra parte, las observaciones al TEM de OBs producidos por algunos cypovirus reportadas por Bellonzik y Mori (1998) presentan una forma esferoide de ~3µm. En principio, las VLPs encontradas en la muestra P5 presentan una simetría similar, aunque de menor tamaño a los OBs "circulares" mencionados (< 1 µm) (**figura 46**). Sin embargo, al igual que con las VLPs de P16, no se puede saber si dichas VLPs presentan algún contenido.

El conjunto de información previa de VLPs se trató de corresponder con el análisis bioinformático, enfocándose en las secuencias relacionadas con los grupos de virus que se sabe producen OBs. Sin embargo, como se describo anteriormente, las secuencias recuperadas relacionadas a baculoviridos fueron escasas, lo que impide inferir la presencia de algún virus putativo relacionado a dicho grupo. Por otra parte, los contigs rv representan un grupo de secuencias que sugieren la presencia de un virus putativo relacionado a reoviridos que infectan insectos. En relación a esto, es importante mencionar que anteriormente se ha demostrado (mediante vectores de expresión) que el segmento S4 de OpBRV (el mejor hit de la proteína hipotética de PF_rv_4) no es una polihedrína (Graham et al., 2008). Por otra parte, las relaciones filogenéticas con la proteína codificada por PF_rv_1 y el resto de las RdRPs de reoviridos indican que PF_rv_1 se aparta considerablemente del género *Cypovirus* (cuyos taxa son los únicos miembros de la familia *Roviridae* que produce polihedrina) (**figura 31**).

En conjunto, pese a que PF_rv_4 y PF_rv_1 se relacionan más estrechamente con idnoreovirus y otros taxa no clasificados, hace falta trabajo experimental para verificar un vínculo entre dichas secuencias (como segmentos de un mismo genoma), así como verificar la secuencia completa de PF_rv_4 y las propiedades de la proteína (saber si presenta características de función y/o estrucutra similares a las de otras polihedrinas). Sin embargo, las características de las partículas encontradas en el presente análisis de VLPs incentivan a realizar experimentos poseriores que permitan obtener una mayor abundancia de partículas con semejanza a OBs.

Capítulo 5. Conclusiones

La secuenciación masiva a partir del RNA de *P. ficus* y los análisis bioinformáticos permitieron encontrar secuencias cuyas características indican la presencia de posibles virus de RNA asociados al insecto o a la vid. Algunos de los grupos taxonómicos a los que fueron asignados las secuencias seleccionadas se distinguen por agrupar ISVs.

Se recuperó el genoma casi completo (PF_dv_1) de un virus de RNA putativo, relacionado con la familia *Dicistroviridae*. De acuerdo con los hallazgos recientes respecto al virus Phenacoccus solenopsis virus (PhSoV), y a la elevada similitud entre PhSov y PF_dv_1, este último podría representar un virus de RNA que infecta a *P. ficus*, aunque no específico para esta especie. Sin embargo, se necesitan análisis ulteriores de confirmación y caracterización física y molecular sobre la interacción entre el virus putativo y las especies de piojo harinoso asociadas.

Por otra parte, se recuperaron dos secuencias correspondientes a virus de vid; el genoma casi completo (PF_mfv_1) de un aislado de Grapevine Syrah virus 1 (GSyV-1) y el genoma parcial (PF_mfv_2) de un aislado de Grapevine rupestris vein feathering virus (GRVFV). Hasta el momento, estas secuencias representan los primeros aislados de dichos virus encontrados en México y los primeros en ser recuperados mediante un análisis no dirigido a partir de tejido de insecto. Los análisis de similitud de estos aislados y su alta abundancia en las bibliotecas de *P. ficus* sugieren, respectivamente, una alta divergencia entre PF_mfv_1 respecto a otros aislados reportados previamente y una relación directa entre ambos aislados encontrados y el insecto. La identificación de secuencias adicionales con alta similitud a GSyV-1 y GRVFV sugiere la necesidad de realizar estudios posteriores de carcterización de aislados de dichos virus encontrados en Baja California.

El resto de las secuencias recuperadas que representan fragmentos con homología a genomas de virus de RNA o DNA putativos deben ser evaluadas mediante análisis experimentales para verificar su procedencia, ya que comparten algunas características de secuencia con EVEs encontrados en estudios previos de virómica de insectos. Sin embargo, algunas de ellas, pese a que representan secuencias parciales, pueden de hecho representar fragmentos de genomas de virus episomales. En el caso de las secuencias relacionadas a segmentos de genomas multipartitas (como las secuencias relacionadas con la familia *Reoviridae)*, la integridad de todos los segmentos como parte de un mismo genoma viral debe ser validada verificando la conservación de extremos y el análisis de electroferotipos.

La cantidad de secuencias recuperadas con relación a virus de dsDNA fue muy reducida debido en parte a que la mayoría de las secuencias mostraron una mayor similitud con secuencias de organismos celulares. Por otro lado, en el rango de hospederos conocido para la mayoría de los virus de dsDNA de los grupos taxonómicos considerados, los hemípteros no están incluidos. Por lo tanto, en el presente trabajo no se excluía la posibilidad de no encontrar abundancia en las secuencias distintivas de dichos virus.

Finalmente, las observaciones de VLPs indicaron la presencia de ciertos agregados con semejanza a las partículas oclusivas de virus de DNA o RNA. Sin embargo, deben considerarse futuros análisis para tratar de incrementar la abundancia de VLPs y así confirmar o descartar la presencia de virus ocluidos, puesto que el análisis de secuencias no reveló evidencia contundente que soporte la presencia o alta representatividad de secuencias correspondientes a los grupos de virus de dsDNA o RNA que forman OBs.

Literatura citada

Aguiar, E. R. G. R., Olmo, R. P., & Marques, J. T. (2016). Virus-derived small RNAs: Molecular footprints of host-pathogen interactions. *Wiley Interdisciplinary Reviews. RNA*, *7*(6), 824–837. https://doi.org/10.1002/wrna.1361

Al Rwahnih, M., Daubert, S., Golino, D., & Rowhani, A. (2009). Deep sequencing analysis of RNAs from a grapevine showing Syrah decline symptoms reveals a multiple virus infection that includes a novel virus. *Virology*, *387*(2), 395–401. https://doi.org/10.1016/j.virol.2009.02.028

Almeida, R., Daane, K., Bell, V., Blaisdell, G. K., Cooper, M., Herrbach, E., & Pietersen, G. (2013). Ecology and management of grapevine leafroll disease. *Frontiers in Microbiology*, *4*. https://doi.org/10.3389/fmicb.2013.00094

Anderson, M. W., & Schrijver, I. (2010). Next Generation DNA Sequencing and the Future of Genomic Medicine. *Genes*, *1*(1), 38–69. https://doi.org/10.3390/genes1010038

Asgari, S., Bideshi, D. K., Bigot, Y., Federici, B. A., & Cheng, X.-W. (2017). ICTV Virus Taxonomy Profile: Ascoviridae. *The Journal of General Virology*, *98*(1), 4–5. https://doi.org/10.1099/jgv.0.000677

Attoui, H., Mertens, P., Becnel, Belaganahalli, M., Bergoin, Brussaard, C., Chappell, Ciarlet, Del, V., Dermody, Dormitzer, Duncan, Fang, Graham, G., Guglielmi, Harding, Hillman, B., Makkay, A., Marzachì, & Zhou, H. (2011). Family Reoviridae. En *Virus Taxonomy: Ninth Report of the International Committee on Taxonomy of Viruses*.vol. 9 Elsevier.

Batson, J., Dudas, G., Haas-Stapleton, E., Kistler, A. L., Li, L. M., Logan, P., Ratnasiri, K., & Retallack, H. (2020). Single mosquito metatranscriptomics recovers mosquito species, blood meal sources, and microbial cargo, including viral dark matter. *BioRxiv*, 2020.02.10.942854. https://doi.org/10.1101/2020.02.10.942854

Becerra, V., González, M., Herrera, M. E., & Miano, J. L. (2006). Dinámica poblacional de *Planococcus ficus* sign.(hemiptera-pseudococcidae) en vinedos. Mendoza (argentina). Revista de la Facultad de Ciencias Agrarias, 38(1), 1-6.

Belloncik, S., & Mori, H. (1998). Cypoviruses. En L. K. Miller & L. A. Ball (Eds.), *The Insect Viruses* (pp. 337–369). Springer US. https://doi.org/10.1007/978-1-4615-5341-0_11

Beltrà, A., Addison, P., Ávalos, J. A., Crochard, D., Garcia-Marí, F., Guerrieri, E., Giliomee, J. H., Malausa, T., Navarro-Campos, C., Palero, F., & Soto, A. (2015). Guiding Classical Biological Control of an Invasive Mealybug Using Integrative Taxonomy. *PLOS ONE*, *10*(6), e0128685. https://doi.org/10.1371/journal.pone.0128685

Bleidorn, C. (2017). Phylogenomics. Cham: Springer International Publishing. doi: https://doi.org/10.1007/978-3-319-54064-1

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Bolling, B. G., Weaver, S. C., Tesh, R. B., & Vasilakis, N. (2015). Insect-Specific Virus Discovery: Significance for the Arbovirus Community. *Viruses*, *7*(9), 4911–4928. https://doi.org/10.3390/v7092851

Bonning, B. C. (2009). The Dicistroviridae: An emerging family of invertebrate viruses. *Virologica Sinica*, *24*(5), 415. https://doi.org/10.1007/s12250-009-3044-1

Bonning, B. C., & Nusawardani, T. (2007). Introduction to the Use of Baculoviruses as Biological Insecticides. En D. W. Murhammer (Ed.), *Baculovirus and Insect Cell Expression Protocols* (pp. 359–366). Humana Press. https://doi.org/10.1007/978-1-59745-457-5_18

Bonning, B. C., Pal, N., Liu, S., Wang, Z., Sivakumar, S., Dixon, P. M., King, G. F., & Miller, W. A. (2014). Toxin delivery by the coat protein of an aphid-vectored plant virus provides plant resistance to aphids. *Nature Biotechnology*, *32*(1), 102–105. https://doi.org/10.1038/nbt.2753

Brown, K., Olendraite, I., Valles, S. M., Firth, A. E., Chen, Y., Guérin, D. M. A., Hashimoto, Y., Herrero, S., de Miranda, J. R., Ryabov, E., & ICTV Report Consortium. (2019). ICTV Virus Taxonomy Profile: Solinviviridae. *Journal of General Virology, 100*(5), 736–737. https://doi.org/10.1099/jgv.0.001242

Caballero, P., & Williams, T. (2008). Virus entomopatógenos. In Control biológico de plagas agrícolas (pp. 121-135). Phytoma.

Camacho, C., Madden, T., Ma, N., Tao, T., Agarwala, R., & Morgulis, A. (2008). BLAST command line applications user manual, BLAST® help [Internet]. National Center for Biotechnology Information (US), Bethesda, MD USA.

Capella-Gutiérrez, S., Silla-Martínez, J. M., & Gabaldón, T. (2009). trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, *25*(15), 1972–1973. https://doi.org/10.1093/bioinformatics/btp348

Caragata, E. P., Dutra, H. L. C., & Moreira, L. A. (2016). Exploiting Intimate Relationships: Controlling Mosquito-Transmitted Disease with Wolbachia. *Trends in Parasitology*, *32*(3), 207–218. https://doi.org/10.1016/j.pt.2015.10.011

Castillo, A. A. F., & Del Real, A. A. (2009). Guía para el control de piojo harinoso de la vid. Folleto Técnico, (38).

Castillo, A. A. F., Blanco, J. L. M., Acosta, G. O., & Carrillo, J. L. M. (2004). Control quimico de piojo harinoso *Planococcus ficus* Signoret (Homoptera: Pseudococcidae) en vid de mesa. Agricultura técnica en Mexico, 30(1), 101-105.

Chen, Y. P., & Siede, R. (2007). Honey Bee Viruses. En *Advances in Virus Research* (Vol. 70, pp. 33–80). Academic Press. https://doi.org/10.1016/S0065-3527(07)70002-7

Chen, Y. P., Pettis, J. S., Corona, M., Chen, W. P., Li, C. J., Spivak, M., Visscher, P. K., DeGrandi-Hoffman, G., Boncristiani, H., Zhao, Y., vanEngelsdorp, D., Delaplane, K., Solter, L., Drummond, F., Kramer, M., Lipkin, W. I., Palacios, G., Hamilton, M. C., Smith, B., … Evans, J. D. (2014). Israeli Acute Paralysis Virus: Epidemiology, Pathogenesis and Implications for Honey Bee Health. *PLOS Pathogens*, *10*(7), e1004261. https://doi.org/10.1371/journal.ppat.1004261

Chen, Y., Nakashima, N., Christian, P., Bakonyi, T., Bonning, B., Valles, S., & Lightner, D. (2012a). Family—Dicistroviridae. *Virus Taxonomy: Ninth Report of the International Committee on Taxonomy of Viruses*, vol. 9, 840–845. Elsevier. https://doi.org/10.1016/B978-0-12-384684-6.00071-9

Chen, Y., Nakashima, N., Christian, P., Bakonyi, T., Bonning, B., Valles, S., & Lightner, D. (2012b). Family—Iflaviridae. *Virus Taxonomy: Ninth Report of the International Committee on Taxonomy of Viruses*, vol. 9, 846–849.Elsevier. https://doi.org/10.1016/B978-0-12-384684-6.00072-0

Colson, P., De Lamballerie, X., Yutin, N., Asgari, S., Bigot, Y., Bideshi, D. K., & La Scola, B. (2013). "Megavirales", a proposed new order for eukaryotic nucleocytoplasmic large DNA viruses. Archives of virology, 158(12), 2517-2521.

Comité Estatal de Sanidad Vegetal de Baja California, CESVBC. (2018). Programa de trabajo de la campaña contra piojo harinoso de la vid., a operar con recursos del programa de sanidad e inocuidad agroalimentaria 2018, componente sanidad federalizado, subcomponente de sanidad vegetal, en el estado de Baja Balifornia. Consultado en 2018 de: https://www.cesvbc.org/

Conceição-Neto, N., Zeller, M., Lefrère, H., De Bruyn, P., Beller, L., Deboutte, W., Yinda, C. K., Lavigne, R., Maes, P., Ranst, M. V., Heylen, E., & Matthijnssens, J. (2015). Modular approach to customise sample preparation procedures for viral metagenomics: A reproducible protocol for virome analysis. *Scientific Reports*, *5*, 16532. https://doi.org/10.1038/srep16532

Cory, J. S., & Myers, J. H. (2003). The Ecology and Evolution of Insect Baculoviruses. *Annual Review of Ecology, Evolution, and Systematics*, *34*(1), 239–272. https://doi.org/10.1146/annurev.ecolsys.34.011802.132402

Cox, J. M. (1989). The mealybug genus *Planococcus* (Homoptera: Pseudococcidae). *Bulletin of the British Museum (Natural History), Entomology*, *58*(1), 1–78.

D'Arcy, C. J., Burnett, P. A., & Hewings, A. D. (1981). Detection, biological effects, and transmission of a virus of the aphid Rhopalosiphum padi. *Virology*, *114*(1), 268–272. https://doi.org/10.1016/0042-6822(81)90275-0

Daane, K. M., Almeida, R. P. P., Bell, V. A., Walker, J. T. S., Botton, M., Fallahzadeh, M., Mani, M., Miano, J. L., Sforza, R., Walton, V. M., & Zaviezo, T. (2012). Biology and Management of Mealybugs in Vineyards. En N. J. Bostanian, C. Vincent, & R. Isaacs (Eds.), *Arthropod Management in Vineyards: Pests, Approaches, and Future Directions* (pp. 271–307). Springer Netherlands. https://doi.org/10.1007/978-94-007-4032-7_12

Daane, K. M., Middleton, M. C., Sforza, R. F. H., Kamps-Hughes, N., Watson, G. W., Almeida, R. P. P., Correa, M. C. G., Downie, D. A., & Walton, V. M. (2018). Determining the geographic origin of invasive populations of the mealybug *Planococcus ficus* based on molecular genetic analysis. *PLOS ONE*, *13*(3), e0193852. https://doi.org/10.1371/journal.pone.0193852

Daane, K. M., Middleton, M. C., Sforza, R., Cooper, M. L., Walton, V. M., Walsh, D. B., Zaviezo, T., & Almeida, R. P. P. (2011). Development of a Multiplex Pcr for Identification of Vineyard Mealybugs. *Environmental Entomology*, *40*(6), 1595–1603. https://doi.org/10.1603/EN11075

Daane, K., Cooper, M., Triapitsyn, S., Walton, V., Yokota, G., Haviland, D., Bentley, W., Godfrey, K., & Wunderlich, L. (2008). Vineyard managers and researchers seek sustainable solutions for mealybugs, a changing pest complex. *California Agriculture*, *62*(4), 167–176.

Daane, K., Weber, E., & Bentley, W. (2004). Formidable pest spreading through California vineyards. Practical Winery and Vineyard Magazine. Cosultado http://cenapa.ucdavis.edu/files/52580.pdf

de Farias, S. T., dos Santos Junior, A. P., Rêgo, T. G., & José, M. V. (2017). Origin and Evolution of RNA-Dependent RNA Polymerase. *Frontiers in Genetics*, *8*. https://doi.org/10.3389/fgene.2017.00125

de Miranda, J. R., & Genersch, E. (2010). Deformed wing virus. *Journal of Invertebrate Pathology*, *103*, S48–S61. https://doi.org/10.1016/j.jip.2009.06.012

Deitz, L. L., Alvarez, P. A., Bartlett, C. R., Cryan, J. R., Dietrich, C. H., and Rakitov R. A. (2008). Suborder Auchenorryncha. Consultado el 20 de octubre de 2020 de: https://www.lib.ncsu.edu/specialcollections/digital/metcalf/auchenorrhyncha.html

Depledge, D. P., Mohr, I., & Wilson, A. C. (2019). Going the distance: optimizing RNA-Seq strategies for transcriptomic analysis of complex viral genomes. *Journal of Virology*, *93*(1). https://doi.org/10.1128/JVI.01342-18

Douglas, A. E. (1998). Nutritional Interactions in Insect-Microbial Symbioses: Aphids and Their Symbiotic Bacteria Buchnera. *Annual Review of Entomology*, *43*(1), 17–37. https://doi.org/10.1146/annurev.ento.43.1.17

Doyle, J. J., & Doyle, J. L. (1990). Isolation ofplant DNA from fresh tissue. Focus, 12(13), 39-40.

Eddy, S. R. (2011). Accelerated Profile HMM Searches. *PLoS Computational Biology*, *7*(10). https://doi.org/10.1371/journal.pcbi.1002195

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., Qureshi, M., Richardson, L. J., Salazar, G. A., Smart, A., Sonnhammer, E. L. L., Hirsh, L., Paladin, L., Piovesan, D., Tosatto, S. C. E., & Finn, R. D. (2019). The Pfam protein families database in 2019. *Nucleic Acids Research*, *47*(D1), D427–D432. https://doi.org/10.1093/nar/gky995

Estopà Consuegra, L. (2015). Control biológico de la cochinilla algodonosa de la vid *Planococcus ficus* (Signoret) (Hemiptera: Pseudococcidae) en uva de mesa en el Valle del Vinalopó. Influencia y manejo de las hormigas. Tesis de Máster en producción vegetal y ecosistemas agroforestales. Universitat Politècnica de Valencia. 43 pp.

Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, *32*(19), 3047–3048. https://doi.org/10.1093/bioinformatics/btw354

Fassler, J., & Cooper, P. (2011). BLAST Glossary. En *BLAST® Help [Internet]*. National Center for Biotechnology Information (US). Consultado el 20 de octubre de 2020 de: https://www.ncbi.nlm.nih.gov/books/NBK62051/

Feng, Y., Krueger, E. N., Liu, S., Dorman, K., Bonning, B. C., & Miller, W. A. (2017). Discovery of Known and Novel Viral Genomes in Soybean Aphid by Deep Sequencing. *Phytobiomes Journal*, *1*(1), 36–45. https://doi.org/10.1094/PBIOMES-11-16-0013-R

Flint, S. J., Racaniello, V. R., Rall, G. F., Skalka, A. M., & Enquist, L. W. (2015). Principles of virology, Volume 1: The Science of Virology and the Molecular Biology of Viruses. (4th ed). Whasington. D.C., USA: American Society for Microbiology.

Freeman, S., & Herron, J. C. (2007). Evolutionary analysis Upper Saddle River, NJ: Pearson Prentice Hall.

Gall, O. L., Christian, P., Fauquet, C. M., King, A. M. Q., Knowles, N. J., Nakashima, N., Stanway, G., & Gorbalenya, A. E. (2008). Picornavirales, a proposed order of positive-sense single-stranded RNA viruses with a pseudo-T = 3 virion architecture. *Archives of Virology*, *153*(4), 715–727. https://doi.org/10.1007/s00705-008-0041-x

García Morales M, Denno BD, Miller DR, Miller GL, Ben-Dov Y, Hardy NB. 2016. ScaleNet: A literature-based model of scale insect biology and systematics. Database. doi: 10.1093/database/bav118. http://scalenet.info.

Gilbert, K. B., Holcomb, E. E., Allscheid, R. L., & Carrington, J. C. (2019). Hiding in plain sight: New virus genomes discovered via a systematic analysis of fungal public transcriptomes. *PLOS ONE*, *14*(7), e0219207. https://doi.org/10.1371/journal.pone.0219207

Glasa, M., Predajňa, L., Šoltys, K., Sabanadzovic, S., & Olmos, A. (2015). Detection and molecular characterisation of Grapevine Syrah virus-1 isolates from Central Europe. *Virus Genes*, *51*(1), 112–121. https://doi.org/10.1007/s11262-015-1201-1

Gorbalenya, A. E., Pringle, F. M., Zeddam, J.-L., Luke, B. T., Cameron, C. E., Kalmakoff, J., Hanzlik, T. N., Gordon, K. H. J., & Ward, V. K. (2002). The Palm Subdomain-based Active Site is Internally Permuted in Viral RNA-dependent RNA Polymerases of an Ancient Lineage. *Journal of Molecular Biology*, *324*(1), 47–62. https://doi.org/10.1016/S0022-2836(02)01033-1

Graham, R. I., Rao, S., Possee, R. D., Sait, S. M., Mertens, P. P. C., & Hails, R. S. (2006). Detection and characterisation of three novel species of reovirus (Reoviridae), isolated from geographically separate populations of the winter moth Operophtera brumata (Lepidoptera: Geometridae) on Orkney. *Journal of Invertebrate Pathology*, *91*(2), 79–87. https://doi.org/10.1016/j.jip.2005.11.003

Graham, R. I., Rao, S., Sait, S. M., Attoui, H., Mertens, P. P. C., Hails, R. S., & Possee, R. D. (2008). Sequence analysis of a reovirus isolated from the winter moth Operophtera brumata (Lepidoptera: Geometridae) and its parasitoid wasp Phobocampe tempestiva (Hymenoptera: Ichneumonidae). *Virus Research*, *135*(1), 42–47. https://doi.org/10.1016/j.virusres.2008.02.005

Gregory R. Warnes, Ben Bolker, Lodewijk Bonebakker, Robert Gentleman, Wolfgang Huber Andy Liaw, Thomas Lumley, Martin Maechler, Arni Magnusson, Steffen Moeller, Marc Schwartz and Bill Venables (2016). gplots: Various R Programming Tools for Plotting Data. R package version 3.0.1. Consultado el 20 de octubre de 2020 de: https://CRAN.R-project.org/package=gplots

Gullan, P. J., & Martin, J. H. (2009). Chapter 244 - Sternorrhyncha: (Jumping Plant-Lice, Whiteflies, Aphids, and Scale Insects). En V. H. Resh & R. T. Cardé (Eds.), *Encyclopedia of Insects (Second Edition)* (pp. 957–967). Academic Press. https://doi.org/10.1016/B978-0-12-374144-8.00253-8

Haase, S., Sciocco-Cap, A., & Romanowski, V. (2015). Baculovirus Insecticides in Latin America: Historical Overview, Current Status and Future Perspectives. *Viruses*, *7*(5), 2230–2267. https://doi.org/10.3390/v7052230

Hall, R. J., Wang, J., Todd, A. K., Bissielo, A. B., Yen, S., Strydom, H., Moore, N. E., Ren, X., Huang, Q. S., Carter, P. E., & Peacey, M. (2014). Evaluation of rapid and simple techniques for the enrichment of viruses prior to metagenomic virus discovery. *Journal of Virological Methods*, *195*, 194–204. https://doi.org/10.1016/j.jviromet.2013.08.035

Harrison, R. L., Herniou, E. A., Jehle, J. A., Theilmann, D. A., Burand, J. P., Becnel, J. J., Krell, P. J., van Oers, M. M., Mowery, J. D., Bauchan, G. R., & Ictv Report Consortium, null. (2018). ICTV Virus Taxonomy Profile: Baculoviridae. *The Journal of General Virology*, *99*(9), 1185–1186. https://doi.org/10.1099/jgv.0.001107

Harvey, E., Rose, K., Eden, J.-S., Lo, N., Abeyasuriya, T., Shi, M., Doggett, S. L., & Holmes, E. C. (2019). Extensive Diversity of RNA Viruses in Australian Ticks. *Journal of Virology*, *93*(3). https://doi.org/10.1128/JVI.01358-18

Heather, J. M., & Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. *Genomics*, *107*(1), 1–8. https://doi.org/10.1016/j.ygeno.2015.11.003

Herrbach, E., Alliaume, A., Prator, C. A., Daane, K. M., Cooper, M. L., & Almeida, R. P. P. (2017). Vector Transmission of Grapevine Leafroll-Associated Viruses. En B. Meng, G. P. Martelli, D. A. Golino, & M. Fuchs (Eds.), *Grapevine Viruses: Molecular Biology, Diagnostics and Management* (pp. 483–503). Springer International Publishing. https://doi.org/10.1007/978-3-319-57706-7_24

Hertz, M. I., & Thompson, S. R. (2011). Mechanism of translation initiation by Dicistroviridae IGR IRESs. *Virology*, *411*(2), 355–361. https://doi.org/10.1016/j.virol.2011.01.005

Hily, J.-M., Candresse, T., Garcia, S., Vigne, E., Tannière, M., Komar, V., Barnabé, G., Alliaume, A., Gilg, S., Hommay, G., Beuve, M., Marais, A., & Lemaire, O. (2018). High-Throughput Sequencing and the Viromic Study of Grapevine Leaves: From the Detection of Grapevine-Infecting Viruses to the Description of a New Environmental Tymovirales Member. *Frontiers in Microbiology*, *9*. https://doi.org/10.3389/fmicb.2018.01782

Holmes, E. C. (2011). The evolution of endogenous viral elements. *Cell host & microbe*, *10*(4), 368-377. https://doi.org/10.1016/j.chom.2011.09.002

Hong, H. K., Woo, S. D., Choi, J. Y., Lee, H. K., Kim, M. H., Je, Y. H., & Kang, S. K. (2000). Characterization of four isolates of Bombyx mori nucleopolyhedrovirus. Archives of virology, 145(11), 2351-2361.

Horta, A. B., Ardisson-Araujo, D. M. P., da Silva, L. A., de Melo, F. L., da Silva Morgado, F., Franco Lemos, M. V., Ribeiro, Z. A., Boiça, A. L., Wilcken, C. F., & Ribeiro, B. M. (2018). Genomic analysis of a cypovirus isolated from the eucalyptus brown looper, Thyrinteina arnobia (Stoll, 1782)

(Lepidoptera: Geometridae). *Virus Research*, *253*, 62–67. https://doi.org/10.1016/j.virusres.2018.05.026

Huang, X., & Madan, A. (1999). CAP3: A DNA Sequence Assembly Program. *Genome Research*, *9*(9), 868–877. https://doi.org/10.1101/gr.9.9.868

Huerta-Cepas, J., Szklarczyk, D., Forslund, K., Cook, H., Heller, D., Walter, M. C., Rattei, T., Mende, D. R., Sunagawa, S., Kuhn, M., Jensen, L. J., von Mering, C., & Bork, P. (2016). eggNOG 4.5: A hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Research*, *44*(D1), D286–D293. https://doi.org/10.1093/nar/gkv1248

Hughes, A. L., & Friedman, R. (2003). Genome-Wide Survey for Genes Horizontally Transferred from Cellular Organisms to Baculoviruses. *Molecular Biology and Evolution*, *20*(6), 979–987. https://doi.org/10.1093/molbev/msg107

Hughes, A. L., & Friedman, R. (2005). Poxvirus genome evolution by gene gain and loss. *Molecular Phylogenetics and Evolution*, *35*(1), 186–195. https://doi.org/10.1016/j.ympev.2004.12.008

Iasur-Kruh, L., Taha-Salaime, L., Robinson, W. E., Sharon, R., Droby, S., Perlman, S. J., & Zchori-Fein, E. (2015). Microbial Associates of the Vine Mealybug *Planococcus ficus* (Hemiptera: Pseudococcidae) under Different Rearing Conditions. *Microbial Ecology*, *69*(1), 204–214. https://doi.org/10.1007/s00248-014-0478-2

Ikeda, M., Hamajima, R., & Kobayashi, M. (2015). Baculoviruses: Diversity, evolution and manipulation of insects. *Entomological Science*, *18*(1), 1–20. https://doi.org/10.1111/ens.12105

Illumina, I. (2015). An introduction to next-generation sequencing technology. Consultado el 20 de octubre de 2020 de:https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&ved=2ahUKEwja8K-fnOrsAhUeJzQIHfMlD24QFjAAegQIARAC&url=https%3A%2F%2Fwww.illumina.com%2Fcontent%2Fdam%2Fillumina-marketing%2Fdocuments%2Fproducts%2Fillumina_sequencing_introduction.pdf&usg=AOvVaw1EmZyrpFkQHu7D9dMbtGBf

International Committee on Taxonomy of Viruses Executive Committee. (2020a). The new scope of virus taxonomy: partitioning the virosphere into 15 hierarchical ranks. Nature Microbiology, 5(5), 668.

International Committee on Taxonomy of Viruses Executive Committee. (2020b). How to write virus, species, and other taxa names. Consultado el 20 de octubre de 2020 de:https://talk.ictvonline.org/files/ictv_documents/m/gen_info/7004/download

Jan, E. (2006). Divergent IRES elements in invertebrates. *Virus Research*, *119*(1), 16–28. https://doi.org/10.1016/j.virusres.2005.10.011

Junglen, S., & Drosten, C. (2013). Virus discovery and recent insights into virus diversity in arthropods. *Current Opinion in Microbiology*, *16*(4), 507–513. https://doi.org/10.1016/j.mib.2013.06.005

Katoh, K., Misawa, K., Kuma, K., & Miyata, T. (2002). MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, *30*(14), 3059–3066. https://doi.org/10.1093/nar/gkf436

King, A. M., Lefkowitz, E., Adams, M. J., & Carstens, E. B. (2011). *Virus Taxonomy: Ninth Report of the International Committee on Taxonomy of Viruses*. Vol. 9. Elsevier.

King, L. A., Wilkinson, N., Miller, D. P., & Marlow, S. A. (1998). Entomopoxviruses. En L. K. Miller & L. A. Ball (Eds.), *The Insect Viruses* (pp. 1–29). Springer US. https://doi.org/10.1007/978-1-4615-5341-0_1

Kleiner, M., Hooper, L. V., & Duerkop, B. A. (2015). Evaluation of methods to purify virus-like particles for metagenomic sequencing of intestinal viromes. *BMC Genomics*, *16*(1), 7. https://doi.org/10.1186/s12864-014-1207-4

Koonin, E. V., & Dolja, V. V. (2013). A virocentric perspective on the evolution of life. *Current Opinion in Virology*, *3*(5), 546–557. https://doi.org/10.1016/j.coviro.2013.06.008

Koonin, E. V., & Yutin, N. (2010). Origin and Evolution of Eukaryotic Large Nucleo-Cytoplasmic DNA Viruses. *Intervirology*, *53*(5), 284–292. https://doi.org/10.1159/000312913

Koonin, E. V., Dolja, V. V., & Krupovic, M. (2015). Origins and evolution of viruses of eukaryotes: The ultimate modularity. *Virology*, *479–480*, 2–25. https://doi.org/10.1016/j.virol.2015.02.039

Koonin, E. V., Dolja, V. V., Krupovic, M., Varsani, A., Wolf, Y. I., Yutin, N., Zerbini, F. M., & Kuhn, J. H. (2020). Global Organization and Proposed Megataxonomy of the Virus World. *Microbiology and Molecular Biology Reviews*, *84*(2). https://doi.org/10.1128/MMBR.00061-19

Koonin, E. V., Senkevich, T. G., & Dolja, V. V. (2006). The ancient Virus World and evolution of cells. *Biology Direct*, *1*(1), 29. https://doi.org/10.1186/1745-6150-1-29

Krupovic, M., & Koonin, E. V. (2015). Polintons: A hotbed of eukaryotic virus, transposon and plasmid evolution. *Nature Reviews Microbiology*, *13*(2), 105–115. https://doi.org/10.1038/nrmicro3389

Lacey, L. A., Grzywacz, D., Shapiro-Ilan, D. I., Frutos, R., Brownbridge, M., & Goettel, M. S. (2015). Insect pathogens as biological control agents: Back to the future. *Journal of Invertebrate Pathology*, *132*, 1–41. https://doi.org/10.1016/j.jip.2015.07.009

Laubscher, J. M., & von Wechmar, M. B. (1992). Influence of aphid lethal paralysis virus and Rhopalosiphum padi virus on aphid biology at different temperatures. *Journal of Invertebrate Pathology*, *60*(2), 134–140. https://doi.org/10.1016/0022-2011(92)90086-J

Lefeuvre, P., Martin, D. P., Elena, S. F., Shepherd, D. N., Roumagnac, P., & Varsani, A. (2019). Evolution and ecology of plant viruses. *Nature Reviews Microbiology*, *17*(10), 632–644. https://doi.org/10.1038/s41579-019-0232-3

Letunic, I., & Bork, P. (2007). Interactive Tree Of Life (iTOL): An online tool for phylogenetic tree display and annotation. *Bioinformatics*, *23*(1), 127–128. https://doi.org/10.1093/bioinformatics/btl529

Li, C.-X., Shi, M., Tian, J.-H., Lin, X.-D., Kang, Y.-J., Chen, L.-J., Qin, X.-C., Xu, J., Holmes, E. C., & Zhang, Y.-Z. (2015a). Unprecedented genomic diversity of RNA viruses in arthropods reveals the ancestry of negative-sense RNA viruses. *eLife*, *4*, e05378. https://doi.org/10.7554/eLife.05378

Li, D., Liu, C.-M., Luo, R., Sadakane, K., & Lam, T.-W. (2015b). MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*, *31*(10), 1674–1676. https://doi.org/10.1093/bioinformatics/btv033

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li H., Handsaker, Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G. (2009). Durbin R, and 1000 Genome Project Data Processing Subgroup. The Sequence alignment/map (SAM) format and SAMtools, Bioinformatics, 25(16), 2078-2079.

Lin, D., Zhang, L., Shao, W., Li, X., Liu, X., Wu, H., & Rao, Q. (2019). Phylogenetic analyses and characteristics of the microbiomes from five mealybugs (Hemiptera: Pseudococcidae). *Ecology and Evolution*, *9*(4), 1972–1984. https://doi.org/10.1002/ece3.4889

Liu, S., Vijayendran, D., & Bonning, B. C. (2011). Next Generation Sequencing Technologies for Insect Virus Discovery. *Viruses*, *3*(10), 1849–1869. https://doi.org/10.3390/v3101849

Liu, S., Vijayendran, D., Chen, Y., & Bonning, B. C. (2016). Aphis Glycines Virus 2, a Novel Insect Virus with a Unique Genome Structure. *Viruses*, *8*(11), 315. https://doi.org/10.3390/v8110315

Llorens, C., Futami, R., Covelli, L., Domínguez-Escribá, L., Viu, J. M., Tamarit, D., Aguilar-Rodríguez, J., Vicente-Ripolles, M., Fuster, G., Bernet, G. P., Maumus, F., Munoz-Pomer, A., Sempere, J. M., Latorre, A., & Moya, A. (2011). The Gypsy Database (GyDB) of mobile genetic elements: Release 2.0. *Nucleic Acids Research*, *39*(suppl_1), D70–D74. https://doi.org/10.1093/nar/gkq1061

Luria, N., Smith, E., Lachman, O., Laskar, O., Sela, N., & Dombrovsky, A. (2020). Isolation and characterization of a novel cripavirus, the first Dicistroviridae family member infecting the cotton mealybug Phenacoccus solenopsis. *Archives of Virology*, *165*(9), 1987–1994. https://doi.org/10.1007/s00705-020-04702-7

Maliogka, V. I., Martelli, G. P., Fuchs, M., & Katis, N. I. (2015). Chapter Six—Control of Viruses Infecting Grapevine. En G. Loebenstein & N. I. Katis (Eds.), *Advances in Virus Research* (Vol. 91, pp. 175–227). Academic Press. https://doi.org/10.1016/bs.aivir.2014.11.002

Mani, M., & Amala, U. (2016). Fruit Crops: Grapevine. En M. Mani & C. Shivaraju (Eds.), *Mealybugs and their Management in Agricultural and Horticultural crops* (pp. 329–351). Springer India. https://doi.org/10.1007/978-81-322-2677-2_38

Mani, M., & Shivaraju, C. (2016a). Ant Association. En M. Mani & C. Shivaraju (Eds.), *Mealybugs and their Management in Agricultural and Horticultural crops* (pp. 199–208). Springer India. https://doi.org/10.1007/978-81-322-2677-2_15

Mani, M., & Shivaraju, C. (2016b). Damage. En M. Mani & C. Shivaraju (Eds.), *Mealybugs and their Management in Agricultural and Horticultural crops* (pp. 117–122). Springer India. https://doi.org/10.1007/978-81-322-2677-2_9

Mani, M., & Shivaraju, C. (2016c). Introduction. En M. Mani & C. Shivaraju (Eds.), *Mealybugs and their Management in Agricultural and Horticultural crops* (pp. 1–3). Springer India. https://doi.org/10.1007/978-81-322-2677-2_1

Manousis, T., & Moore, N. F. (1987). Cricket Paralysis Virus, a Potential Control Agent for the Olive Fruit Fly, Dacus oleae Gmel. *Applied and Environmental Microbiology*, *53*(1), 142–148.

Martelli, G. P. (2017). An Overview on Grapevine Viruses, Viroids, and the Diseases They Cause. En B. Meng, G. P. Martelli, D. A. Golino, & M. Fuchs (Eds.), *Grapevine Viruses: Molecular Biology, Diagnostics and Management* (pp. 31–46). Springer International Publishing. https://doi.org/10.1007/978-3-319-57706-7_2

Martelli, G. P., Sabanadzovic, S., Sabanadzovic, N. A.-G., Edwards, M. C., & Dreher, T. (2002). The family Tymoviridae. *Archives of Virology*, *147*(9), 1837–1846. https://doi.org/10.1007/s007050200045

Martin, S. J., Highfield, A. C., Brettell, L., Villalobos, E. M., Budge, G. E., Powell, M., Nikaido, S., & Schroeder, D. C. (2012). Global Honey Bee Viral Landscape Altered by a Parasitic Mite. *Science*, *336*(6086), 1304–1306. https://doi.org/10.1126/science.1220941

McMenamin, A. J., & Genersch, E. (2015). Honey bee colony losses and associated viruses. *Current Opinion in Insect Science*, *8*, 121–129. https://doi.org/10.1016/j.cois.2015.01.015

McWilliam, H., Li, W., Uludag, M., Squizzato, S., Park, Y. M., Buso, N., Cowley, A. P., & Lopez, R. (2013). Analysis Tool Web Services from the EMBL-EBI. *Nucleic Acids Research*, *41*(W1), W597–W600. https://doi.org/10.1093/nar/gkt376

Medd, N. C., Fellous, S., Waldron, F. M., Xuéreb, A., Nakai, M., Cross, J. V., & Obbard, D. J. (2018). The virome of Drosophila suzukii, an invasive pest of soft fruit. *Virus Evolution*, *4*(1). https://doi.org/10.1093/ve/vey009

Metzker, M. L. (2010). Sequencing technologies—The next generation. *Nature Reviews Genetics*, *11*(1), 31–46. https://doi.org/10.1038/nrg2626

Miller, M. A., Pfeiffer, W., & Schwartz, T. (2012). The CIPRES science gateway: Enabling high-impact science for phylogenetics researchers with limited resources. *Proceedings of the 1st Conference of the Extreme Science and Engineering Discovery Environment: Bridging from the eXtreme to the campus and beyond*, 1–8. https://doi.org/10.1145/2335755.2335836

Milusheva, S., Phelan, J., Piperkova, N., Nikolova, V., Gozmanova, M., & James, D. (2019). Molecular analysis of the complete genome of an unusual virus detected in sweet cherry (Prunus avium) in Bulgaria. *European Journal of Plant Pathology*, *153*(1), 197–207. https://doi.org/10.1007/s10658-018-1555-z

Mokili, J. L., Rohwer, F., & Dutilh, B. E. (2012). Metagenomics and future perspectives in virus discovery. *Current Opinion in Virology*, *2*(1), 63–77. https://doi.org/10.1016/j.coviro.2011.12.004

Monroy Corral, K. M. (2019). Presencia de virus transmitidos por piojo harinoso (*Planococcus ficus*) en viñedos de Baja California. Tesis de Ingeniería bioquímica. Tecnológico Nacional de México/Instituto Tecnológico de Tepic. 51 pp.

Moon, J. S., Domier, L. L., McCoppin, N. K., D'Arcy, C. J., & Jin, H. (1998). Nucleotide Sequence Analysis Shows thatRhopalosiphum padiVirus Is a Member of a Novel Group of Insect-Infecting RNA Viruses. *Virology*, *243*(1), 54–65. https://doi.org/10.1006/viro.1998.9043

Nooij, S., Schmitz, D., Vennema, H., Kroneman, A., & Koopmans, M. P. G. (2018). Overview of Virus Metagenomic Classification Methods and Their Biological Applications. *Frontiers in Microbiology*, *9*. https://doi.org/10.3389/fmicb.2018.00749

Nouri, S., Matsumura, E. E., Kuo, Y.-W., & Falk, B. W. (2018). Insect-specific viruses: From discovery to potential translational applications. *Current Opinion in Virology*, *33*, 33–41. https://doi.org/10.1016/j.coviro.2018.07.006

Nouri, S., Salem, N., Nigg, J. C., & Falk, B. W. (2016). Diverse Array of New Viral Sequences Identified in Worldwide Populations of the Asian Citrus Psyllid (Diaphorina citri) Using Viral Metagenomics. *Journal of Virology*, *90*(5), 2434–2445. https://doi.org/10.1128/JVI.02793-15

Nurk, S., Bankevich, A., Antipov, D., Gurevich, A., Korobeynikov, A., Lapidus, A., ... & Stepanauskas, R. (2013). Assembling genomes and mini-metagenomes from highly chimeric reads. In *Annual International Conference on Research in Computational Molecular Biology* (pp. 158-170). Springer, Berlin, Heidelberg.

Öhlund, P., Lundén, H., & Blomström, A.-L. (2019). Insect-specific virus evolution and potential effects on vector competence. *Virus Genes*, *55*(2), 127–137. https://doi.org/10.1007/s11262-018-01629-9

Oliver, K. M., Degnan, P. H., Hunter, M. S., & Moran, N. A. (2009). Bacteriophages Encode Factors Required for Protection in a Symbiotic Mutualism. *Science*, *325*(5943), 992–994. https://doi.org/10.1126/science.1174463

Oliver, K. M., Russell, J. A., Moran, N. A., & Hunter, M. S. (2003). Facultative bacterial symbionts in aphids confer resistance to parasitic wasps. *Proceedings of the National Academy of Sciences*, *100*(4), 1803–1807. https://doi.org/10.1073/pnas.0335320100

Ondov, B. D., Bergman, N. H., & Phillippy, A. M. (2011). Interactive metagenomic visualization in a Web browser. *BMC Bioinformatics*, *12*(1), 385. https://doi.org/10.1186/1471-2105-12-385

Ottati, S., Chiapello, M., Galetto, L., Bosco, D., Marzachì, C., & Abbà, S. (2020). New Viral Sequences Identified in the Flavescence Dorée Phytoplasma Vector Scaphoideus titanus. *Viruses*, *12*(3), 287. https://doi.org/10.3390/v12030287

Pearson, W. R. (2013). An Introduction to Sequence Similarity ("Homology") Searching. *Current Protocols in Bioinformatics*, *42*(1), 3.1.1-3.1.8. https://doi.org/10.1002/0471250953.bi0301s42

Pecman, A., Kutnjak, D., Gutiérrez-Aguirre, I., Adams, I., Fox, A., Boonham, N., & Ravnikar, M. (2017). Next Generation Sequencing for Detection and Discovery of Plant Viruses and Viroids: Comparison of Two Approaches. *Frontiers in Microbiology*, *8*. https://doi.org/10.3389/fmicb.2017.01998

Pevsner J. (2005). Bioinformatics and functional genomics. (3rd ed). Oxford, UK: Willey Blackwell.

Prosser, W. A., & Douglas, A. E. (1991). The aposymbiotic aphid: An analysis of chlortetracycline-treated pea aphid, Acyrthosiphon pisum. *Journal of Insect Physiology*, *37*(10), 713–719. https://doi.org/10.1016/0022-1910(91)90104-8

Radford, A. D., Chapman, D., Dixon, L., Chantrey, J., Darby, A. C., & Hall, N. (2012). Application of next-generation sequencing technologies in virology. *The Journal of General Virology*, *93*(Pt 9), 1853–1868. https://doi.org/10.1099/vir.0.043182-0

Renault, S. (2012). Chapter 16—RNA Viruses in Parasitoid Wasps. En N. E. Beckage & J.-M. Drezen (Eds.), *Parasitoid Viruses* (pp. 193–201). Academic Press. https://doi.org/10.1016/B978-0-12-384858-1.00016-3

Renault, S., Petit, A., Bénédet, F., Bigot, S., & Bigot, Y. (2002). Effects of the Diadromus pulchellus ascovirus, DpAV-4, on the hemocytic encapsulation response and capsule melanization of the leek-moth pupa, Acrolepiopsis assectella. *Journal of Insect Physiology*, *48*(3), 297–302. https://doi.org/10.1016/S0022-1910(01)00174-3

Rochon, D., Lommel, S., Martelli, G. P., Rubino, L., & Russo, M. (2012). Family Tombusviridae. *Virus Taxonomy: Ninth Report of the International Committee on the Taxonomy of Viruses*, vol. 9, 1111–1138. Elsevier

Roossinck, M. J. (2011). The good viruses: Viral mutualistic symbioses. *Nature Reviews Microbiology*, *9*(2), 99–108. https://doi.org/10.1038/nrmicro2491

Roossinck, M. J., Martin, D. P., & Roumagnac, P. (2015). Plant Virus Metagenomics: Advances in Virus Discovery. *Phytopathology*, *105*(6), 716–727. https://doi.org/10.1094/PHYTO-12-14-0356-RVW

Rosario, K., & Breitbart, M. (2011). Exploring the viral world through metagenomics. *Current Opinion in Virology*, *1*(4), 289–297. https://doi.org/10.1016/j.coviro.2011.06.004

Rosario K., Capobianco H., Ng T.F.F., Breitbart M., Polston J.E. (2014). RNA Viral Metagenome of Whiteflies Leads to the Discovery and Characterization of a Whitefly-Transmitted Carlavirus in North America. PLoS ONE 9(1): e86748. doi:10.1371/journal.pone.0086748

Roundy, C. M., Azar, S. R., Rossi, S. L., Weaver, S. C., & Vasilakis, N. (2017). Chapter Four - Insect-Specific Viruses: A Historical Overview and Recent Developments. En M. Kielian, T. C. Mettenleiter, & M. J. Roossinck (Eds.), *Advances in Virus Research* (Vol. 98, pp. 119–146). Academic Press. https://doi.org/10.1016/bs.aivir.2016.10.001

Roux, S., Adriaenssens, E. M., Dutilh, B. E., Koonin, E. V., Kropinski, A. M., Krupovic, M., Kuhn, J. H., Lavigne, R., Brister, J. R., Varsani, A., Amid, C., Aziz, R. K., Bordenstein, S. R., Bork, P., Breitbart, M., Cochrane, G. R., Daly, R. A., Desnues, C., Duhaime, M. B., … Eloe-Fadrosh, E. A. (2019). Minimum Information about an Uncultivated Virus Genome (MIUViG). *Nature Biotechnology*, *37*(1), 29–37. https://doi.org/10.1038/nbt.4306

Ryabov, E. V. (2017). Invertebrate RNA virus diversity from a taxonomic point of view. *Journal of Invertebrate Pathology*, *147*, 37–50. https://doi.org/10.1016/j.jip.2016.10.002

Ryabov, E. V., Keane, G., Naish, N., Evered, C., & Winstanley, D. (2009). Densovirus induces winged morphs in asexual clones of the rosy apple aphid, Dysaphis plantaginea. *Proceedings of the National Academy of Sciences*, *106*(21), 8465–8470. https://doi.org/10.1073/pnas.0901389106

Rzedowski J., y Calderón de Rzedowski, G. (2005). Fascículo 131: Vitaceae. En: Flora del Bajío y de regiones adyacentes. Instituto de Ecología/Consejo Nacional de Ciencia y Tecnología/Comisión Nacional para el Conocimiento y Uso de la Biodiversidad. Consultado el 20 de octubre de 2020 de: http://inecolbajio.inecol.mx/floradelbajio/documentos/fasciculos/ordinarios/Vitaceae%20131.pdf

Sabanadzovic, S., Aboughanem-Sabanadzovic, N., & Martelli, G. P. (2017). Grapevine fleck and similar viruses. En B. Meng, G. P. Martelli, D. A. Golino, & M. Fuchs (Eds.), *Grapevine Viruses: Molecular Biology, Diagnostics and Management* (pp. 331–349). Springer International Publishing. https://doi.org/10.1007/978-3-319-57706-7_16

Sadeghi, M., Altan, E., Deng, X., Barker, C. M., Fang, Y., Coffey, L. L., & Delwart, E. (2018). Virome of > 12 thousand Culex mosquitoes from throughout California. *Virology*, *523*, 74–88. https://doi.org/10.1016/j.virol.2018.07.029

Sánchez, R., Serra, F., Tárraga, J., Medina, I., Carbonell, J., Pulido, L., de María, A., Capella-Gutíerrez, S., Huerta-Cepas, J., Gabaldón, T., Dopazo, J., & Dopazo, H. (2011). Phylemon 2.0: A suite of web-tools for molecular evolution, phylogenetics, phylogenomics and hypotheses testing. *Nucleic Acids Research*, *39*(Web Server issue), W470–W474. https://doi.org/10.1093/nar/gkr408

Sayers, E. W., Barrett, T., Benson, D. A., Bolton, E., Bryant, S. H., Canese, K., Chetvernin, V., Church, D. M., DiCuccio, M., Federhen, S., Feolo, M., Fingerman, I. M., Geer, L. Y., Helmberg, W., Kapustin, Y., Landsman, D., Lipman, D. J., Lu, Z., Madden, T. L., … Ye, J. (2011). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, *39*(suppl_1), D38–D51. https://doi.org/10.1093/nar/gkq1172

Shi, M., Lin, X.-D., Tian, J.-H., Chen, L.-J., Chen, X., Li, C.-X., Qin, X.-C., Li, J., Cao, J.-P., Eden, J.-S., Buchmann, J., Wang, W., Xu, J., Holmes, E. C., & Zhang, Y.-Z. (2016). Redefining the invertebrate RNA virosphere. *Nature*, *540*(7634), 539–543. https://doi.org/10.1038/nature20167

Shi, M., Zhang, Y.-Z., & Holmes, E. C. (2018). Meta-transcriptomics and the evolutionary biology of RNA viruses. *Virus Research*, *243*, 83–90. https://doi.org/10.1016/j.virusres.2017.10.016

Signoret, V., 1875. Essay sur les cochenilles ou gallinsectes (Homopteres: Coccides), xv. Annis Soc. ent. Fr. (5)5:305-352.

Simmonds, P., Adams, M. J., Benkő, M., Breitbart, M., Brister, J. R., Carstens, E. B., Davison, A. J., Delwart, E., Gorbalenya, A. E., Harrach, B., Hull, R., King, A. M. Q., Koonin, E. V., Krupovic, M., Kuhn, J. H., Lefkowitz, E. J., Nibert, M. L., Orton, R., Roossinck, M. J., … Zerbini, F. M. (2017). Virus taxonomy in the age of metagenomics. *Nature Reviews Microbiology*, *15*(3), 161–168. https://doi.org/10.1038/nrmicro.2016.177

Skewes-Cox, P., Sharpton, T. J., Pollard, K. S., & DeRisi, J. L. (2014). Profile Hidden Markov Models for the Detection of Viruses within Metagenomic Sequence Data. *PLoS ONE*, *9*(8). https://doi.org/10.1371/journal.pone.0105067

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*(9), 1312-1313. https://doi.org/10.1093/bioinformatics/btu033

Stark, R., Grzelak, M., & Hadfield, J. (2019). RNA sequencing: The teenage years. *Nature Reviews Genetics*, *20*(11), 631–656. https://doi.org/10.1038/s41576-019-0150-2

Teixeira, L., Ferreira, Á., & Ashburner, M. (2008). The Bacterial Symbiont Wolbachia Induces Resistance to RNA Viral Infections in Drosophila melanogaster. *PLOS Biology*, *6*(12), e1000002. https://doi.org/10.1371/journal.pbio.1000002

Temmam, S., Vongphayloth, K., Hertz, J. C., Sutherland, I., Douangboubpha, B., Grandadam, M., Bigot, T., Brey, P. T., & Eloit, M. (2019). Six Nearly Complete Genome Segments of a Novel Reovirus Identified in Laotian Batflies. *Microbiology Resource Announcements*, *8*(46). https://doi.org/10.1128/MRA.00733-19

Terral, J.-F., Tabard, E., Bouby, L., Ivorra, S., Pastor, T., Figueiral, I., Picq, S., Chevance, J.-B., Jung, C., Fabre, L., Tardy, C., Compan, M., Bacilieri, R., Lacombe, T., & This, P. (2010). Evolution and history of grapevine (Vitis vinifera) under domestication: New morphometric perspectives to understand seed domestication syndrome and reveal origins of ancient European cultivars. *Annals of Botany*, *105*(3), 443–455. https://doi.org/10.1093/aob/mcp298

Thermo Scientific T042-TECHNICAL BULLETIN. Wilmington, Delaware. Consultado 20 de octubre de 2020 de:https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&ved=2ahUKEwjt3IPo6rsAhVkIDQIHU91Cw4QFjACegQICxAC&url=https%3A%2F%2Fwww.researchgate.net%2Fprofile%2FZx_Chong%2Fpost%2FLow_260_230_ratio%2Fattachment%2F5c74a938cfe4a781a5834be0%2FAS%253A730387542179840%2540155114936814 6%2Fdownload%2Fnanodrop.pdf&usg=AOvVaw0B4Vj3HCnv0-jMB2FDokNw

Thézé, J., Takatsuka, J., Nakai, M., Arif, B., & Herniou, E. A. (2015). Gene Acquisition Convergence between Entomopoxviruses and Baculoviruses. *Viruses*, *7*(4), 1960–1974. https://doi.org/10.3390/v7041960

Thiem, S. M. (1999). Insect virus diversity: The Insect Viruses, edited by L.K. Miller and L.A. Ball. *Trends in Microbiology*, *7*(11), 463. https://doi.org/10.1016/S0966-842X(99)01597-8

Thurber, R. V., Haynes, M., Breitbart, M., Wegley, L., & Rohwer, F. (2009). Laboratory procedures to generate viral metagenomes. *Nature Protocols*, *4*(4), 470–483. https://doi.org/10.1038/nprot.2009.10

Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B. C., Remm, M., & Rozen, S. G. (2012). Primer3—New capabilities and interfaces. *Nucleic Acids Research*, *40*(15), e115–e115. https://doi.org/10.1093/nar/gks596

Valles, S. M., & Hashimoto, Y. (2009). Isolation and characterization of Solenopsis invicta virus 3, a new positive-strand RNA virus infecting the red imported fire ant, Solenopsis invicta. *Virology*, *388*(2), 354–361. https://doi.org/10.1016/j.virol.2009.03.028

Valles, S. M., & Rivers, A. R. (2019). Nine new RNA viruses associated with the fire ant Solenopsis invicta from its native range. *Virus Genes*, *55*(3), 368–380. https://doi.org/10.1007/s11262-019-01652-4

Valles, S. M., Porter, S. D., & Firth, A. E. (2014). Solenopsis invicta virus 3: Pathogenesis and stage specificity in red imported fire ants. *Virology*, *460–461*, 66–71. https://doi.org/10.1016/j.virol.2014.04.026

Valles, S. M., Porter, S. D., Choi, M.-Y., & Oi, D. H. (2013). Successful transmission of Solenopsis invicta virus 3 to Solenopsis invicta fire ant colonies in oil, sugar, and cricket bait formulations. *Journal of Invertebrate Pathology*, *113*(3), 198–204. https://doi.org/10.1016/j.jip.2013.04.003

van Munster, M., Dullemans, A. M., Verbeek, M., van den Heuvel, J. F. J. M., Clérivet, A., & van der Wilk, F. (2002). Sequence analysis and genomic organization of Aphid lethal paralysis virus: A new member of the family Dicistroviridae. *Journal of General Virology, 83*(12), 3131–3138. https://doi.org/10.1099/0022-1317-83-12-3131

van Oers, M. M. (2010). Genomics and biology of Iflaviruses. Insect virology, 231-250.

Vasilakis, N., & Tesh, R. B. (2015). Insect-specific viruses and their potential impact on arbovirus transmission. *Current Opinion in Virology*, *15*, 69–74. https://doi.org/10.1016/j.coviro.2015.08.007

Venkataraman, S., Prasad, B. V. L. S., & Selvarajan, R. (2018). RNA Dependent RNA Polymerases: Insights from Structure, Function and Evolution. *Viruses*, *10*(2), 76. https://doi.org/10.3390/v10020076

Viljakainen, L., & Jurvansuu, J. (2020). Discovery and Analysis of RNA Viruses in Insects. En F. Sandrelli & G. Tettamanti (Eds.), *Immunity in Insects* (pp. 191–200). Springer US. https://doi.org/10.1007/978-1-0716-0259-1_12

Vincent, C., Isaacs, R., Bostanian, N. J., & Lasnier, J. (2012). Principles of Arthropod Pest Management in Vineyards. En N. J. Bostanian, C. Vincent, & R. Isaacs (Eds.), *Arthropod Management in Vineyards: Pests, Approaches, and Future Directions* (pp. 1–16). Springer Netherlands. https://doi.org/10.1007/978-94-007-4032-7_1

Walton, V. M., & Pringle, K. L. (2004). Vine mealybug, *Planococcus ficus* (Signoret) (Hemiptera: Pseudococcidae), a Key Pest in South African vineyards. A Review. *South African Journal of Enology and Viticulture*, *25*(2), 54–62. https://doi.org/10.21548/25-2-2140

Werren, J. H., Baldo, L., & Clark, M. E. (2008). *Wolbachia:* master manipulators of invertebrate biology. *Nature Reviews Microbiology*, *6*(10), 741–751. https://doi.org/10.1038/nrmicro1969

Whitfield, A. E., & Rotenberg, D. (2015). Disruption of insect transmission of plant viruses. *Current Opinion in Insect Science*, *8*, 79–87. https://doi.org/10.1016/j.cois.2015.01.009

Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H (2019). Welcome to the Tidyverse. Journal of Open Source Software, 4(43), 1686. doi: 10.21105/joss.01686.

Williams, T. (2008). Natural invertebrate hosts of iridoviruses (Iridoviridae). *Neotropical Entomology*, *37*(6), 615–632. https://doi.org/10.1590/S1519-566X2008000600001

Williams, T., Bergoin, M., & van Oers, M. M. (2017). Diversity of large DNA viruses of invertebrates. *Journal of Invertebrate Pathology*, *147*, 4–22. https://doi.org/10.1016/j.jip.2016.08.001

Wooley, J. C., Godzik, A., & Friedberg, I. (2010). A Primer on Metagenomics. *PLOS Computational Biology*, *6*(2), e1000667. https://doi.org/10.1371/journal.pcbi.1000667

Wu, Qi, Kehoe, M., Kinoti, W. M., Wang, C., Rinaldo, A., Tyerman, S., Habili, N., & Constable, F. E. (2020). First report of grapevine rupestris vein feathering virus in grapevine in Australia. *Plant Disease*. https://doi.org/10.1094/PDIS-06-20-1240-PDN

Wu, Qingfa, Ding, S.-W., Zhang, Y., & Zhu, S. (2015). Identification of Viruses and Viroids by Next-Generation Sequencing and Homology-Dependent and Homology-Independent Algorithms. *Annual Review of Phytopathology*, *53*(1), 425–444. https://doi.org/10.1146/annurev-phyto-080614-120030

Xiao, H., & Meng, B. (2016). First Report of Grapevine asteroid mosaic-associated virus and Grapevine rupestris vein feathering virus in Grapevines in Canada. *Plant Disease*, *100*(10), 2175–2175. https://doi.org/10.1094/PDIS-03-16-0413-PDN

Yasmin, T., Thekke-Veetil, T., Hobbs, H. A., Nelson, B. D., McCoppin, N. K., Lagos-Kutz, D., Hartman, G. L., Lambert, K. N., Walker, D. R., & Domier, L. L. (2020). Aphis glycines virus 1, a new bicistronic virus with two functional internal ribosome entry sites, is related to a group of unclassified viruses in the Picornavirales. *Journal of General Virology, 101*(1), 105–111. https://doi.org/10.1099/jgv.0.001355

Ye, J., Coulouris, G., Zaretskaya, I., Cutcutache, I., Rozen, S., & Madden, T. L. (2012). Primer-BLAST: A tool to design target-specific primers for polymerase chain reaction. *BMC Bioinformatics*, *13*(1), 134. https://doi.org/10.1186/1471-2105-13-134

Zhang, Y.-Z., Chen, Y.-M., Wang, W., Qin, X.-C., & Holmes, E. C. (2019). Expanding the RNA Virosphere by Unbiased Metagenomics. *Annual Review of Virology*, *6*(1), 119–139. https://doi.org/10.1146/annurev-virology-092818-015851

Zhang, Y.-Z., Shi, M., & Holmes, E. C. (2018). Using Metagenomics to Characterize an Expanding Virosphere. *Cell*, *172*(6), 1168–1172. https://doi.org/10.1016/j.cell.2018.02.043

Glosario

Base de datos: Compendio o colección de secuencias (u otro término de búsqueda), generalmente accesibles en repositorios de acceso público, y usada como referencia o acervo (Pevsner, 2005).

Carácter (estado de c.): Cualquier rasgo o atributo de un organismo/sistema sujeto a variación es un carácter, y las diferentes variantes de ese carácter se definen como estados de carácter. Ejemplos de caracteres: genes, proteínas, estructuras anatómicas, etc. Los caracteres homólogos y/o sus estados se usan para guiar las reconstrucciones filogenéticas (Freeman y Herron, 2007). Véase "Homología".

Cobertura de secuenciación (nivel de c.): El número promedio de bases secuenciadas que se alinean a cada base de la secuencia de referencia de DNA. Por ejemplo, un genoma secuenciado a una cobertura de 30X significa que, en promedio, cada base del genoma se secuencio 30 veces (Illumina, 2015).

Cobertura del *query* (qcov): Porción de la secuencia *query* (de nucleótidos o aminoácidos) que se incluye en un alineamiento local (por ejemplo de BLAST) con la secuencia *subject*. Nótese que es un concepto distinto a la "cobertura de secuenciación" (Pevsner, 2005). Véase "secuencia problema" y "secuencia de referencia".

Dominio: Una porción discreta de una proteína que presuntivamente se pliega de forma independiente al resto de la cadena de aminoácidos y posee su propia función. Así, los dominios son estructuras terciarias independientes de las proteínas, distinguidas estructural y funcionalmente, que pueden existir, evolucionar, y funcionar independientemente. Diferentes combinaciones de dominios dan origen a los diferentes tipos de proteínas encontradas en la naturaleza (Fassler y Cooper, 2011; El-Gebali et al., 2019; https://www.ebi.ac.uk/training/online/glossary/domain).

Genes ortólogos: Genes que divergieron al ocurrir un evento de especiación; describe la relación entre genes homólogos encontrados en diferentes especies (Freeman y Herron, 2007). Comparese con "genes parálogos".

Genes parálogos: Genes duplicados que se encuentran en el mismo genoma; describe la relación entre miembros de una misma familia de genes (Freeman y Herron, 2007).

Grupo monofilético (monofilia): Grupo de taxa (organismos/entidades) en el que se encuentran todos los descendientes de un mismo ancestro común más reciente (Freeman y Herron, 2007).

Grupo polifilético (polifilia): Grupo de taxa (organismos/entidades) en el que se encuentra a descendientes que no comparten al ancestro común más reciente (Freeman y Herron, 2007).

Homología: Similitud debido a ancestría/ascendencia común. Los rasgos/caracteres biológicos con esta cualidad se denominan homólogos. En sistemática molecular, los genes homólogos pueden ser ortólogos o parálogos (Freeman y Herron, 2007; Fassler y Cooper, 2011). Véase "Carácter" y compárese con "Homoplasia".

Homoplasia: Similitud debida a convergencia, es decir, a orígenes a partir de linajes independientes (Freeman y Herron, 2011).

Identidad: La medida en que dos secuencias de nucleótidos o aminoácidos tienen los mismos residuos en la misma posición de un alineamiento, frecuentemente expresada como porcentaje (Fassler y Cooper, 2011).

Marco de lectura abierta (ORF): Secuencia de nucleótidos que al traducirse a aminoácidos no contiene codones de paro. La lectura de una hebra de nucleótidos ocurre en sentido 5' →3' y en tripletes o codones (p. ej. AAA, UUU o AUG). Por lo tanto, la lectura en tripletes se lleva a cabo en hasta tres posibles marcos de referencia (denominados marcos de lectura), dependiendo de la posición nucleotídica donde comience a considerarse el primer codón. El marco de lectura abierta corresponde al único de los tres posibles que usa el ribosoma para la traducción (Alberts et al., 2014).

Secuencia problema (*query/queries*): La secuencia problema (u otro término de búsqueda) con la cual se comparan todas las entradas (p. ej. secuencias de referencia) de una base de datos (Pevsner, 2005; Fassler y Cooper, 2011).

Secuencia de referencia (*subject/subjects*): Todas las secuencias (o términos de búsqueda) que se encuentran depositadas en una base de datos (Pevsner, 2005; Fassler y Cooper, 2011).

Similitud: La medida en la cual dos secuencias de nucleótidos o aminoácidos están relacionadas. La similitudentre dos secuencias puede ser expresada como porcentaje de identidad y/o porcentaje de sustituciones positivas (Pevsner, 2005; Fassler y Cooper, 2011).

Valor de expectación *e*/E-value: Representa el número de alineamientos diferentes con mejor puntaje o equivalente al obtenido con una secuencia *subject* determinada, que se espera ocurran por azar en una búsqueda con una secuencia *query*, en una base de datos de tamaño determinado. Entre menor sea el valor de e, es más significativo el puntaje del alineamiento (Pevsner, 2005; Fassler y Cooper, 2011).

Anexo A

Tabla 14. Lista de organismos y genomas/secuencias (con ID de Genebank) utilizados en el filtrado de reads no virales. Se enlistan según su fuente de procedencia.

Secuencias de hospedero	
Paracoccus marginatus	GCA_900065295.1
Pseudococcus longispinus	GCA_900064475.1
Microbioma de *V. vinifera*	
Aureobasidium pullulans	GCF_000721785.1
Lasiodiplodia theobromae	GCA_002111425.1
Rhizopus stolonifer	GCA_000697035.1
Aspergillus niger	GCF_000002855.3
Botrytis cinerea	GCF_000143535.2
Penicillium expansum	GCF_000769745.1
Epicocum nigrium	GCA_002116315.1
Erwinia persicina	GCF_001571305.1
Ustilago maydis	GCF_000328475.2
Fusarium solani	GCA_002215905.1
Bacillus thuringensis	GCF_000008505.1
Pseudomonas fluorescens	GCF_000237065.1
Pantoea agglomerans	GCF_001709315.1
Lactococcus raffinolactis	GCF_001591765.1
Escherichia coli	GCF_000005845.2
Ewingella americana	GCF_000735345.1

Contaminación de laboratorio	
Octopus vulgaris	GCA_003957725.1
Crassostrea gigas	GCA_000297895.2
Alternaria alternata	NW_017306190.1
Microbioma de hospedero	
Candidatus *Morenella endobia* PCIT (complete genome)	NC_015735.1
Candidatus *Morenella endobia* PCVAL (complete genome)	NC_021057.1
Metschnikowia aff. *pulcherrima* strain APC 1.2 chromosome I	CP034456.1
Secondary endosymbiont of *Planococcus citri* 16S ribosomal RNA and 23S ribosomal RNA genes	AF476107.1
Secondary endosymbiont of *Planococcus ficus* 16S ribosomal RNA and 23S ribosomal RNA genes, partial sequence	AF476108.1
Candidatus *Tremblaya princeps* PCVAL, complete genome	CP002918.1
Secondary symbiont of *Cinara* sp. 16S ribosomal RNA gene, partial sequence	FJ655515.1
Secondary symbiont of *Stomaphis cupressi* from Spain 16S ribosomal RNA gene, partial sequence	EU348326.1
Secondary symbiont of *Stomaphis quercus* 16S ribosomal RNA gene, partial sequence	FJ655516.1
Uncultured *Gilliamella* sp. clone G1 16S ribosomal RNA gene, partial sequence	KF709652.1
Hongos epibiontes asociados a *P. ficus*	
Meyerozyma guilliermondi	GCF_000149425.1
Sarocladium oryzae	GCA_001972265.1
Purpureocillium lilacinum	GCF_001653265.1

Anexo B

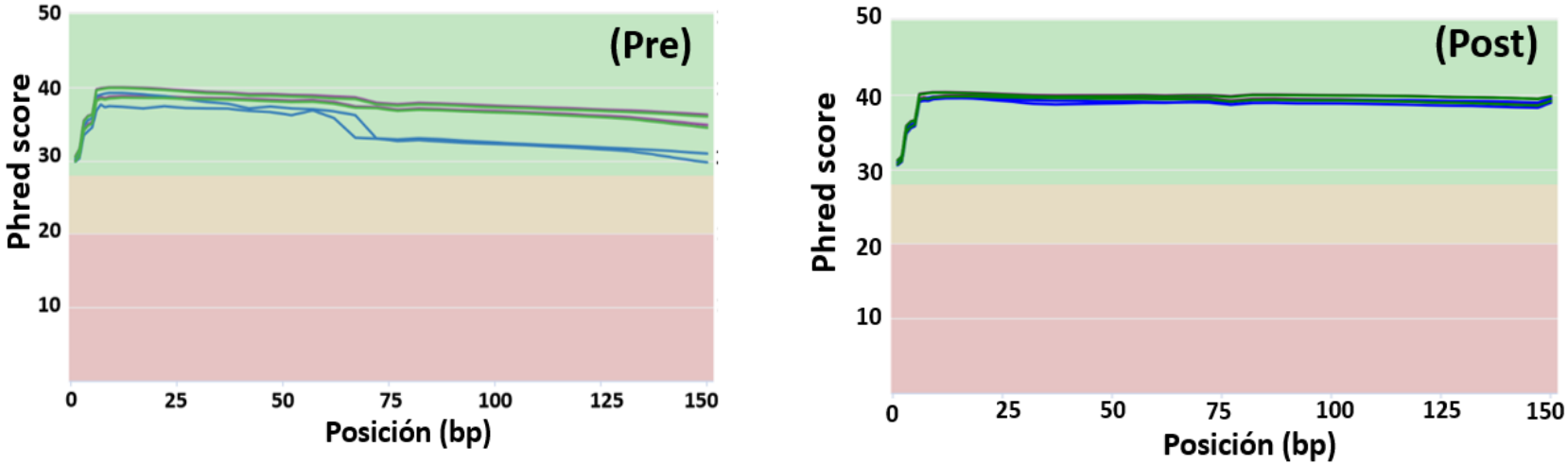

Figura 48. Histogramas de valor medio de calidad para los reads de cada biblioteca. Los gráficos muestran el valor medio de puntaje de calidad *Phred score* (eje vertical) que se asigna para las secuencias de los reads crudos 1P y 2P en una posición determinada (eje horizontal) antes (Pre) y después (Post) del corte de calidad. Las líneas del histograma se muestran en morado para la biblioteca PV1, en azul para la PV2 y en verde para la PV3. Los valores de Phred score que son considerados de calidad baja, media y óptima se denotan en regiones de color pastel rosa, beige y verde, respectivamente, en el fondo de la gráfica.

Anexo C

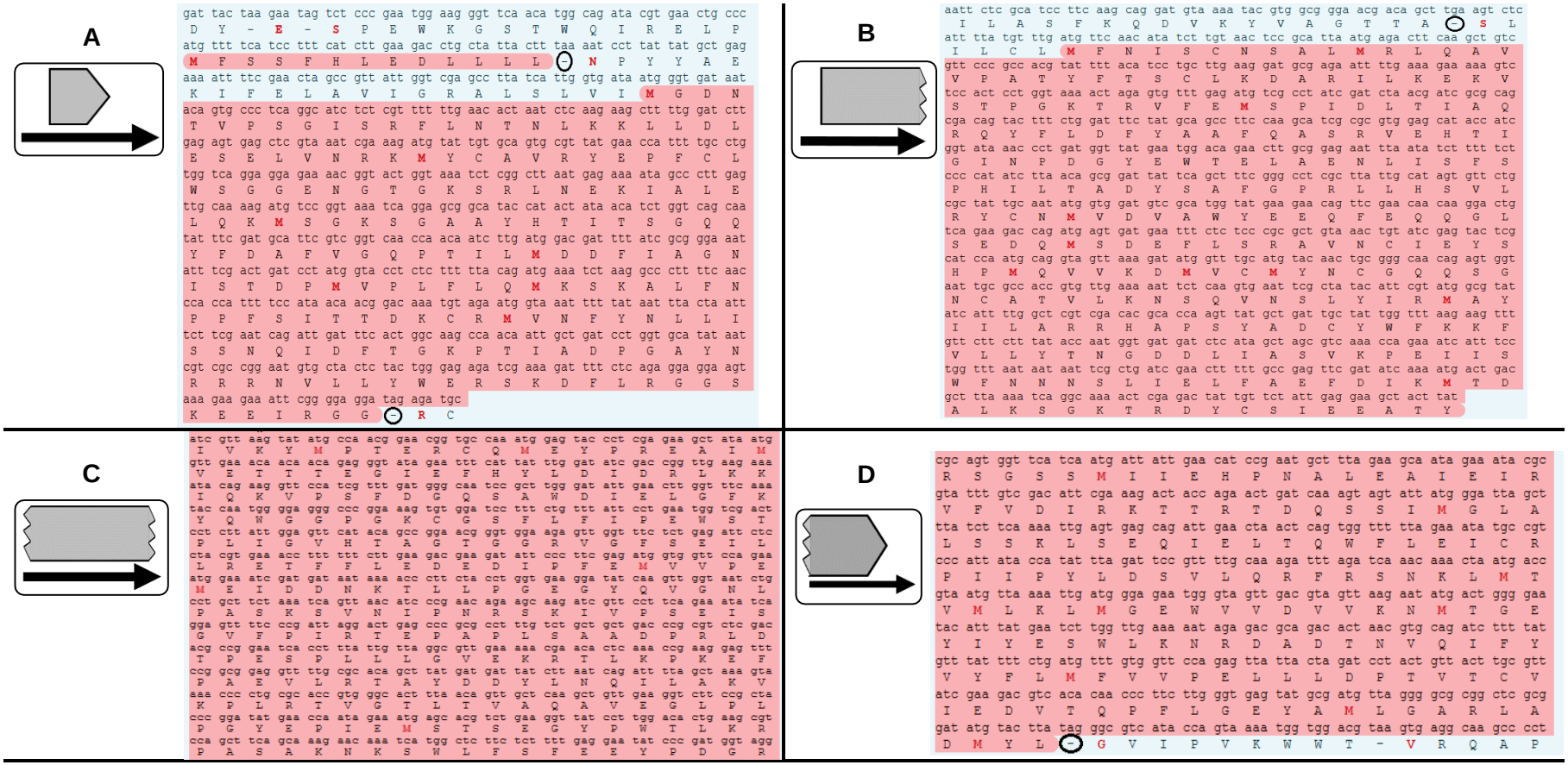

Figura 49. Extensión de los ORFs en los contigs recuperados. Se presentan ejemplos de las 4 diferentes posibles disposiciones de los ORFs predichos en los diferentes contigs recuperados. Se muestran los marcos de lectura (lecturas en tripletes de nt) correspondientes a los ORFs (área rosada) predichos en los contigs (todas las secuencias en dirección 5'→3'), así como la traducción a aminoácidos. En (A) el contig contiene un ORF (considérese sólo el más extenso) delimitado por un codón de inicio (Metionina o M en este caso), y un codón de paro (círculo negro). Nótese que en el mismo marco de lectura del ORF hay un codón de paro (triplete TAA) rio arriba del ORF (i.e., la M abre el marco de lectura). En (B) el inicio del ORF es similar que en (A), pero el ORF se extiende hasta el extremo del contig (no hay codón de paro). En (C) el ORF se extiende sobre todo el contig (no hay codones de inicio ni término); sólo en este caso el ORF es contínuo en el contig. En (D) el ORF se extiende hasta el extremo 5' del contig y encuentra un codón de paro antes de llegar al extremo 3'.

Anexo D

Tabla 15. Lista de primers diseñados para ocho secuencias seleccionadas.

Contig	No. de par	Orientación	Secuencia (5 → 3)	Longitud (nt)	%GC	Región blanco	Tamaño del producto de PCR (pb)
PF_dv_1 (9902 nt)	1	F	CGCATCAACTGTCAACGGGG	20	60	4478-4497	847
		R	GCATGGGTCCACACGCAAAA	20	55	5306-5325	
	2	F	AGTGGAAGTTTCGGCTGGAC	20	55	8690-8709	867
		R	GGCTACTTGGTAGGGAGGGA	20	60	9538-9557	
PF_rv_1 (3620 nt)	1	F	AGTGGGACAGCGTATGAACCT	21	52	1607-1627	866
		R	TCTGCACACATCGTCGTTCCT	21	52.38	2453-2473	
PF_pv_2 (3401 nt)	1	F	TTACCGTCCCCTTTGTGTCC	20	55	1055-1074	573
		R	CGAAGAGATCCTGTCCACCC	20	60	1609-1628	
PF_rbv_4 (2529)	1	F	AATAGCCCTGCCACCAACTG	20	55	650-669	661
		R	TCTGCGTCATTTCGGTATGCT	21	47.6	1291-1311	
PF_tbv_1 (2720)	1	F	GGGCAATAGGGAGTCGTGTA	20	55	482-501	334
		R	CTGGGACCACTTTTCGTCAT	20	50	796-815	
PF_mfv_1 (6423)	1	F	ACCAAATCACAGCCAAGGAC	20	50	4541-4560	788
		R	AGAGAGGGTTTTCCGAGAGC	20	55	5310-5329	
PF_mfv_2 (4305)	1	F	TCTCCTGTTTGAAGCCCACT	20	50	2320-2339	996
		R	AAAGACGGCAGATTGGTACG	20	50	3297-3316	
PF_ifv_2 (1236)	1	F	CTGTAGAGGGCAACGACCAT	20	55	359-378	666
		R	AAATACCAGGCAACGCAAAG	20	45	1006-1025	