UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

POSGRADO EN CIENCIAS BIOLÓGICAS

FACULTAD DE CIENCIAS

EVOLUCIÓN

ESTUDIO PANGENÓMICO DE LOS VIRUS DE DNA DE CADENA SENCILLA

TESIS

QUE PARA OPTAR POR EL GRADO DE:

MAESTRO EN CIENCIAS

PRESENTA:

ABELARDO AGUILAR CÁMARA

TUTOR PRINCIPAL DE TESIS: **DR. ARTURO CARLOS II BECERRA BRACHO**
FACULTAD DE CIENCIAS, UNAM

COMITÉ TUTOR: **DRA. MARÍA COLÍN GARCÍA**
INSTITUTO DE GEOLOGÍA, UNAM

COMITÉ TUTOR: **DR. LEÓN PATRICIO MARTÍNEZ CASTILLA**
CONSEJO NACIONAL DE CIENCIA Y TECNOLOGÍA

CIUDAD UNIVERSITARIA, CD. MX **2022**

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

POSGRADO EN CIENCIAS BIOLÓGICAS

FACULTAD DE CIENCIAS

EVOLUCIÓN

ESTUDIO PANGENÓMICO DE LOS VIRUS DE DNA DE CADENA SENCILLA

TESIS

QUE PARA OPTAR POR EL GRADO DE:

MAESTRO EN CIENCIAS

PRESENTA:

ABELARDO AGUILAR CÁMARA

TUTOR PRINCIPAL DE TESIS: **DR. ARTURO CARLOS II BECERRA BRACHO**

FACULTAD DE CIENCIAS, UNAM

COMITÉ TUTOR: **DRA. MARÍA COLÍN GARCÍA**

INSTITUTO DE GEOLOGÍA, UNAM

COMITÉ TUTOR: **DR. LEÓN PATRICIO MARTÍNEZ CASTILLA**

CONSEJO NACIONAL DE CIENCIA Y TECNOLOGÍA

CIUDAD UNIVERSITARIA, CD. MX **2022**

COOORDINACIÓN DEL POSGRADO EN CIENCIAS BIOLÓGICAS

FACULTAD DE CIENCIAS

DIVISIÓN ACADÉMICA DE INVESTIGACIÓN Y POSGRADO

OFICIO FCIE/DAIP/0448/2022

ASUNTO: Oficio de Jurado

M. en C. Ivonne Ramírez Wence
Directora General de Administración Escolar, UNAM
P r e s e n t e

Me permito informar a usted que en la reunión ordinaria del Comité del Posgrado en Ciencias Biológicas, celebrada el día **13 de junio de 2022** se aprobó el siguiente jurado para el examen de grado de **MAESTRO EN CIENCIAS BIOLÓGICAS** en el campo de conocimiento de **Biología Evolutiva** del estudiante **AGUILAR CÁMARA ABELARDO** con número de cuenta **312223584** con la tesis titulada **"Estudio pangenómico de los virus de DNA de cadena sencilla",** realizada bajo la dirección del **DR. ARTURO CARLOS II BECERRA BRACHO,** quedando integrado de la siguiente manera**:**

Presidente:	**DR. EDGAR ENRIQUE SEVILLA REYES**
Vocal:	**DR. LUIS DAVID ALCARAZ PERAZA**
Vocal:	**DRA. CLAUDIA SELENE ZÁRATE GUERRA**
Vocal:	**DR. ALEJANDRO RODRIGO JÁCOME RAMÍREZ**
Secretario:	**DR. LEÓN PATRICIO MARTÍNEZ CASTILLA**

Sin otro particular, me es grato enviarle un cordial saludo.

A T E N T A M E N T E
"POR MI RAZA HABLARÁ EL ESPÍRITU"
Ciudad Universitaria, Cd. Mx., a 13 de octubre de 2022

COORDINADOR DEL PROGRAMA

POSGRADO EN CIENCIAS BIOLOGICAS
COORDINACIÓN

DR. ADOLFO GERARDO NAVARRO SIGÜENZA

AGRADECIMIENTOS INSTITUCIONALES

Agradezco al Posgrado en Ciencias Biológicas, UNAM por permitirme continuar mi formación académica.

Al Consejo Nacional de Ciencia y Tecnología, pues solo me fue posible realizar este proyecto gracias a la beca de posgrado que recibí, CVU 1034340.

A la Universidad Nacional Autónoma de México, a la Facultad de Ciencias y al Laboratorio de Origen de la Vida, por brindarme el apoyo económico para presentar mi proyecto en el *International Symposium on ssDNA Viruses (IS^{3}DV)*

A mi tutor, el Dr. Arturo C. II Becerra Bracho y a los miembros de mi comité, la Dra. María Colín García y el Dr. León P. Martínez Castilla por su constante orientación y valiosos comentarios

AGRADECIMIENTOS PERSONALES

A mi tutor Arturo Becerra por recibirme en el laboratorio e incitarme a explorar los temas y herramientas utilizados con suficiente libertad académica pero siempre guiándome. Aun más, por su amistad y cercanía más allá del quehacer científico.

A los miembros del jurado que revisaron esta tesis. Agradezco su compromiso para robustecer sus bases y posibles alcances.

A los miembros del Laboratorio de Origen de la Vida por el trato fraterno y las variadas charlas.

A mi papá y mamá, Ata y Kikos por ser una fuente inagotable de apoyo en todos los sabores posiblles. Por cultivar el interés del niño fanático de los dinosaurios y de dibujar bichitos bajo el microscopio. Por encaminarme en una vida maravillosa. Los amo.

Al solecito con el que amanezco. Por ese amor invisible que todo trasciende. Por los pequeños pasos que damos. Te amo Diana.

A mi segunda mamá, Mita. Por todo el amor que das.

A mi hermano y hermanxs colados. Joel, Ededo, Yoz y Álvaro por perseguirme a la distancia.

A mi amigo Erick, porque te dejas encontrar en donde quiera que te busque.

A los otros dos mosqueteros Erick y Gustavo, por todas las desventuras, malos consejos y exagerados desvelos ~~in~~necesarios.

A mi amigo Toro por tantos ratos divertidos a la vuelta de la esquina y por nuestro humor unísono.

A las becerritas, Hilda e Ingrid. Su hermandad académica escaló a un boleto rápido para bonitos días de trabajo con amigas.

A los supervivientes, Erick, Iván y Sebas por los loops de madrugada.

A los que nunca se irán, Milo, Zubs, Vero, Mitzi.

A quienes admiro a la distancia, Carmina, Sebas, Chicho, ATG

A los de suaves patitas, Centella, Migi y Hakuna. Su existencia me revalora.

ÍNDICE

RESUMEN

Gracias a los métodos modernos de secuenciación de material genético y al muestreo metagenómico en hospederos y ambientes previamente inexplorados, la concepción de los virus de DNA de cadena sencilla (ssDNA) como una arquitectura genómica rara en la virósfera ha cambiado. La adición de nuevas secuencias sigue ampliando el repertorio genético conocido y nuestra capacidad para hacer inferencias sobre su origen y evolución. En este trabajo presentamos una descripción actual de los grupos de genes ortólogos que constituyen los pangenomas de los virus ssDNA, bajo un enfoque funcional y evolutivo. Para ello, se incorporaron y procesaron más de 1.450 genomas con una metodología particular de concatenación y filtrado para superar la segmentación genómica y cumplir con los requisitos del análisis pangenómico. Esto nos permitió acoplar los resultados del análisis pangenómico con la anotación funcional automatizada mediante homología, así como la búsqueda sistemática de homólogos remotos guiada por dominios en la base de datos UniRef50. El estudio pangenómico de las familias ssDNA develó que la diversidad de grupos de ortólogos conocida es incompleta, a excepción de las familias *Geminiviridae*, *Nanoviridae* y *Circoviridae*, para las que es muy poco probable que la inclusión de nuevos genomas cambie el listado de grupos ortólogos. Además, las búsquedas automatizadas permitieron la incorporación de homólogos celulares, particularmente de elementos genéticos móviles de genomas bacterianos, y la construcción de amplias filogenias que contribuyen a esclarecer los múltiples escapes genéticos celulares que dieron lugar a la diversidad actual de virus ssDNA.

ABSTRACT

Thanks to modern sequencing techniques and metagenomic sampling in previously unexplored hosts and environments, the idea of single-stranded DNA (ssDNA) viruses conceived as a comparatively rare genomic architecture in the virosphere has changed. The addition of new sequences continues to expand the known genetic repertoire, and with it, our ability to make inferences about their origin and evolution. In this work we present a current description of the groups of orthologous genes that constitute the pan-genomes of ssDNA viruses at different scales. To do this, more than 1,450 genomes were incorporated and processed with a particular concatenation and filtering methodology to overcome genomic segmentation and meet the requirements of pangenomic analysis. This allowed us to couple pangenomic analysis with automated functional annotation as well as domain-guided systematic remote homology search in the UniRef50 database. We found that the number of known groups of orthologous genes continues to grow, and that the largest proportion of them corresponds to only a few families. In addition, the automated searches allowed the incorporation of cellular homologues, particularly from the bacterial mobilome, and the construction of extensive phylogenies that contribute to the elucidation of the multiple cellular genetic escapes that gave rise to this diverse group.

INTRODUCCIÓN

¿Qué son los virus?

Los virus son entidades biológicas cuyo genoma está compuesto por DNA o RNA de diversas configuraciones, y se encuentra dentro de una envoltura proteica llamada cápside (Flint et al., 2020). La replicación viral depende de mecanismos celulares, característica que los define como parásitos intracelulares obligados (Tello Lacal, 2019). Esta relación parasítica causa enfermedades humanas, así como de plantas y animales de importancia económica que favorecen un fuerte sesgo biomédico en su estudio. No obstante, a nivel organísmico el efecto neto sobre la supervivencia del hospedador puede ser neutro o incluso positivo (Roossinck, 2011), ocupando diversos roles ecológicos. Se distribuyen como agentes infecciosos de los tres dominios de la vida (Nasir et al., 2014) y muy probablemente constituyen a las entidades biológicas más abundantes en el planeta (Breitbart y Rohwer, 2005). Estructuralmente, los virus pueden percibirse más sencillos que las células (Payne, 2017) pues poseen baja diferenciación y número de partes; además, los diferentes tipos de virus carecen de características genéticas compartidas entre ellos y muy probablemente no surgieron de un ancestro común (Moreira y López-García, 2009).

Características genómicas

El material genético de los virus puede ser tanto de DNA o RNA, de cadena doble, sencilla positiva o sencilla negativa. Dichas configuraciones han servido como base de la clasificación más utilizada en la virología, propuesta por David Baltimore, a partir de viriones presentes en animales (Baltimore, 1971). Esta clasificación ha servido como marco conceptual para el desarrollo de la virología durante décadas, aunque otras características genómicas constituyen grandes ejes de variación, como la segmentación y tamaño del genoma (Koonin et al., 2021).

Los virus pueden ser monopartitos o multipartitos, haciendo referencia al número de segmentos en los que existe su material genético (Tello Lacal, 2019). Aunque típicamente una partícula viral o virión contiene todo el complemento genético necesario para producir una nueva generación de virus, en algunos casos, un virus puede conformarse por más de un virión, cada uno transmitido de forma independiente, portando parte del material genético, y finalmente expresándose en conjunto para continuar el ciclo replicativo (Lucía-Sanz y Manrubia, 2017). El tamaño del genoma va desde ~859 bases en virus de DNA de cadena sencilla (*ssDNA*), hasta ~2473 kb en virus de DNA de cadena doble (*ds*DNA) (Campillo-Balderas et al., 2015). El tamaño del genoma está limitado por el tamaño mismo del virión, por lo que existen pocas regiones no codificantes. Por otra parte, los genomas de virus tienen una alta concentración de genes, llegando incluso a tener genes solapados (Chirico et al., 2010).

Características ecológicas

Como se mencionó anteriormente, el rol ecológico más importante de los virus es el de parásitos intracelulares. A nivel microscópico, una célula infectada sufre un déficit de sus recursos, aunque en la escala organísmica se involucran más factores; de hecho, recientemente se han descubierto relaciones ecológicas complejas en los que la presencia de ciertos virus tiene un efecto global positivo sobre su hospedero (Roossinck, 2015). Las interacciones más antiguas entre virus y hospederos han suscitado que la línea entre virus y hospederos sea borrosa, y es posible que las relaciones comensales y mutualistas sean mayoría (Roossinck, 2011; Roossinck y Bazán, 2017). Existen virus que atenúan enfermedades causadas por otros patógenos, virus que afectan a los competidores del hospedero, o incluso permiten habitar ambientes extremos (Roossinck, 2011). Entre las relaciones mutualistas conocidas de mayor relevancia, destacamos los siguientes ejemplos:

- **Virus mutualistas (Edson et al., 1981).** Los polidnavirus son un ejemplo de virus mutualistas de avispas parasitoides. En una relación muy antigua, los genes encargados de la replicación del virus son ahora parte del genoma de la avispa, y los viriones portan genes de la avispa expresados en sus huevos. Cuando no están presentes, los huevos de avispa no se desarrollan.
- **Retrovirus endógenos (Harris, 1991; Chuong, 2013).** La evolución de la placenta pudo ocurrir tras un evento de endogenización de un virus, mediante proteínas de cubierta (*Env*) que inducen la fusión de membranas, un proceso común en la transmisión de virus célula a célula.
- **Simbiosis de tres vías (Márquez et al., 2007).** En el Parque Nacional de Yellowstone, Estados Unidos, existe una hierba (*Dichanthelium lanuginosum*) que tolera temperaturas mayores a 50°C. Esta termotolerancia depende de la presencia de un hongo endófito (*Curvularia protuberata*), que a su vez, requiere de que esté presente un virus (*Curvularia thermal tolerance virus,* CThTV), mediante mecanismos moleculares presumiblemente conservados.

No obstante, pandemias provocadas por virus como el VIH o SARS-CoV-2 han impulsado la investigación biomédica, teniendo como objeto de estudio su rol como patógenos.

Características evolutivas

Son muchos los factores que intervienen en la evolución viral, sin embargo, algunas cualidades tienen la mayor influencia sobre las tendencias generales. Particularmente, sus tasas de mutación, la cuantiosa descendencia en los ciclos replicativos y la velocidad de su replicación, todas estas características los posicionan como las entidades biológicas que cambian más rápidamente (Duffy y Shackelton, 2008; Retel *et al.*, 2019).

La ocurrencia de mutaciones es más alta que en los organismos celulares y depende principalmente de la conformación del material genético. Los virus de RNA y ssDNA son los de mayores tasas de mutación, entre ~10^{-6} y ~10^{-4}, mientras que los dsDNA entre ~10^{-8} y ~10^{-7} (Sanjuán y Domingo-Calap, 2016). Esto se debe a que las polimerasas virales son más propensas a errores (Duffy y Shackelton, 2008), aunado a que los mecanismos de corrección únicamente están presentes en algunas familias, como *Coronaviridae* (Robson *et al.*, 2020), *Poxviridae* (Moss, 2013) y *Herpesviridae* (Liu et al., 2006). La considerable aparición de cambios deletéreos es compensada con el gran número de descendientes del ciclo replicativo.

Por su parte, la recombinación es también un proceso común en virus (Pérez-Losada *et al.*, 2015). Esta puede ser homóloga cuando ocurre en el mismo sitio en ambas hebras parentales, o no-homóloga entre sitios diferentes de los fragmentos involucrados, por ejemplo, entre virus y hospedero. Estos procesos han sido asociados con la expansión del rango de hospederos y vectores, origen de nuevos virus, variación del tropismo, así como aumento de la virulencia y patogénesis (Martin *et al.*, 2011; Simon-Loriere y Holmes, 2011). La prevalencia de la recombinación varía entre linajes, alcanzando valores incluso más altos que la tasa de mutación en ciertos retrovirus y virus ssDNA (Onafuwa-Nuga y Telesnitsky, 2009; Lukashev, 2010; Martin *et al.*, 2011).

La diversidad a lo largo de un genoma viral es asimétrica (Agrelli et al., 2019). Entre sus elementos genéticos, las proteínas estructurales presentan mayor variabilidad, mientras que las proteínas que participan en la interacción con el material genético son menos cambiantes (Wolf et al., 2018; Aylward *et al.*, 2021); por lo que éstas últimas pueden ser utilizadas para comparaciones de amplio rango taxonómico, para trazar relaciones profundas que permitan indagar su origen y evolución temprana.

Origen de los virus

El llamado "árbol de la vida", es un sistema de representación del conjunto de todas las relaciones evolutivas que guardan los seres vivos. En 1990, Carl Woese y colaboradores, realizaron comparaciones de RNA ribosomal que les permitieron proponer una versión no enraizada de ese árbol, donde era posible distinguir tres grandes grupos: Bacteria, Archaea y Eukarya (Woese *et al.*, 1990). Los tres son altamente divergentes, pero tienen rasgos moleculares compartidos, cuya presencia es explicada por su ancestría común a una forma viva, el llamado último ancestro común (LCA, por sus siglas en inglés) (Delaye *et al.*, 2005). La comparación de secuencias compartidas involucradas en procesos de síntesis, degradación y plegamiento del RNA, ha robustecido dicha propuesta (Becerra *et al.*, 2007).

A diferencia de la condición monofilética de los organismos celulares, en los virus, no existen genes que soporten un único origen. Algunos intentos por compararlos y agruparlos a todos, han concluido que los elementos compartidos entre linajes lejanos son producto de convergencia (Barocchi *et al.*, 2005; Olson *et al.*, 2007), o al movimiento de información genética entre genomas independiente a la replicación, es decir, transferencia genética horizontal (Kazlauskas *et al.*, 2019). Estos eventos son poco frecuentes en organismos celulares (Keeling y Palmer, 2008), pero vconstituyen un porcentaje considerablemente mayor en virus (Pérez-Losada *et al.*, 2015). Lo anterior pone en duda la utilidad misma de la topología de árbol para representar sus relaciones evolutivas (Koonin y Dolja, 2014), y descarta la existencia de un ancestro común a todos ellos (Moreira y López-García, 2009).

Lo anterior promueve la diversificación de las metodologías para el estudio del origen de los virus, así como cambiar los paradigmas con que entendemos sus relaciones

de ancestría-descendencia. Para entender el origen y evolución de los virus, han sido planteadas distintas hipótesis independientes (Figura 1), que no necesariamente son excluyentes. Dichas hipótesis serán explicadas a detalle en los apartados siguientes.

Virocentrismo

La refutación de la generación espontánea y la postulación del concepto de evolución Darwiniana, a finales del siglo XIX, impulsaron el estudio del origen de la vida, y así también surgieron las ideas virocéntricas. Los avances en biología molecular de la primera mitad del siglo XX derivaron en dos líneas de pensamiento (Fry, 2006): i) la vida iniciando como un sistema metabólico, en las llamadas ideas "*citoplásmicas*" o, ii) mediante moléculas autorreplicativas, ideas "*nucleocéntricas*" (López-García, 2012).

Al considerarlos las entidades vivas más simples, los virus se entremezclaron en el debate del origen de la vida durante varias décadas, caracterizados como "formas de vida primordial" (D'Herelle, 1926), y por Haldane como un vínculo entre lo no vivo y las primeras células (Moreira y López-García, 2009). Las hipótesis virocéntricas engloban estos planteamientos, en ellas los virus son organismos primordiales que proporcionaron la materia prima para los organismos celulares y que, por lo tanto, forman parte del árbol de la vida (Figura 1). Estas propuestas tuvieron un auge con el DNA-centrismo, tras el descubrimiento del DNA como portador de la información genética (Avery *et al.*, 1994) y la publicación de su estructura y mecanismos de replicación (Watson y Crick, 1953).

En la primera década del siglo XXI, las ideas virocéntricas tuvieron un repunte, al encontrarse similitudes en las cápsides de virus que infectan organismos de diferentes dominios (Moreira y López-García, 2009). Además de que se descubrió la existencia de genes virales, como los de proteínas de cápside *jelly-roll* y la helicasa *S3H*, distribuidos en un amplio rango de virus de RNA y DNA (Koonin *et al.*, 2006; 2015), pero para los que no han sido descritos homólogos celulares (*hallmark genes*).

Regresión celular

La pérdida de genes es un patrón común en la evolución de los genomas de los parásitos (Jackson, 2015), por lo que se ha postulado que los virus, parásitos intracelulares obligados, surgieron a partir de la reducción genómica a partir de organismos celulares parasitarios (Bándea, 1983), conformando un cuarto dominio de la vida (Raoult y Forterre, 2008 ;Abrahão *et al.*, 2017). A diferencia de las teorías virocéntricas, la regresión celular supone que los virus atravesaron un estado celular en su historia evolutiva (Figura 1), desde la cual redujeron su complejidad morfológica y fisiológica, perdiendo drásticamente la información genética propia y tomando partida del genoma del hospedero (Bándea, 1983). Durante muchos años esta teoría fue rechazada argumentando la falta de conocimiento de formas intermedias entre células y virus, así como el hecho de que los parásitos conocidos mantienen sus características celulares (ribosomas, maquinarias de síntesis proteica y producción de ATP), incluso aquellos considerados más simples (Forterre, 2006). Fue hasta

el descubrimiento de los virus nucleocitoplasmáticos de DNA de gran tamaño o "virus gigantes" (Raoult et al., 2004), que la falta de dichas observaciones se puso en duda. Los virus gigantes poseen genomas >1Mb y repertorios de funciones similares a genomas de bacterias parasitarias, incluyendo proteínas implicadas en el metabolismo, conservadas en los tres dominios, como la arginyl-tRNA, methionyl-tRNA y tyrosyl-tRNA sintetasas (Forterre, 2006), lo que llevó a la propuesta del concepto de "virocélula" (Forterre, 2013).

La teoría de la regresión celular supone que los seres vivos deberían clasificarse en dos grandes grupos (Raoult y Forterre, 2008). Por un lado, los ***REOs*** o *ribosome encoding organisms*, a los que pertenecen los tres dominios que expresan ribosomas, Archaea, Bacteria y Eukarya (1). Por el otro, los ***CEOs*** o *capsid encoding organisms*, categoría que engloba a los virus, organismos que expresan cápsides para producir viriones que infecten a los *REOs*.

Una posibilidad planteada por la hipótesis de regresión celular es que los virus surgieron desde células simples con genomas de RNA, en el mundo de RNA (Forterre, 2015), y que sus cápsides evolucionaron a partir de estructuras similares a la capa S de las arqueas (Forterre, 2006). Además, la regresión celular considera que la similitud estructural entre polimerasas virales de DNA, reverso transcriptasas, y transcriptasas de RNA (Hansen *et al.*, 1997), hace plausible que los virus de DNA y los retrovirus surgieran de virus de RNA.

Hipótesis de escape

La hipótesis de escape plantea que los virus son elementos escapados de genomas celulares convertidos en entidades infecciosas autónomas (Figura 1). Esta teoría sostiene que la mayoría de las proteínas codificadas por virus, surgieron en organismos celulares (Moreira y López-García, 2009), y que podemos encontrar homólogos en sus genomas, o incluso en elementos del "mobiloma"[1] (Ilyina y Koonin, 1992; Kazlauskas *et al.*, 2019). La hipótesis de escape asume que los virus son polifiléticos y que sus relaciones filogenéticas asemejan más un "rizoma" que un árbol (Moreira y López-García, 2009). Consecuentemente, la hipótesis de escape no propone una sola antigüedad para los virus, y va desde considerar que los virus de RNA aparecieron como elementos escapados de células con genomas de RNA antes del último ancestro común (Forterre y Krupovic, 2012), hasta orígenes recientes por medio de transferencia horizontal (Moreira, 2000).

La hipótesis de escape contempla que los escapes de elementos genéticos que participan en los mecanismos de transcripción/traducción causan las mayores transiciones evolutivas (Forterre, 2006). Los principales argumentos en favor de esta teoría son: (1) la alta incidencia de eventos de transferencia genética horizontal que ocurren en los virus, principalmente en dirección célula a virus (Moreira, 2008) y (2) homólogos celulares de los elementos genéticos virales involucrados en la interacción con material genético (Kazlauskas *et al.*, 2019) y metabolismo (Mann *et al.*, 2003).

1.- El mobiloma se refiere a todos los elementos genéticos capaces de desplazarse, como los transposones de eucariontes y plásmidos bacterianos.

A pesar de que no son pocos los estudios que apoyan esta hipótesis, muchos de ellos cuentan historias evolutivas recientes; es decir, están enfocados más en la diversificación que en el origen de los grandes grupos, faltando aún explicar los mecanismos más antiguos de esta teoría (Forterre y Krupovic, 2012).

Finalmente, es necesario resaltar que hasta el momento no existe consenso sobre cuál hipótesis explica mejor los datos conocidos, y es posible ver reflejados sus supuestos en ciertas partes de las filogenias virales.

Virus de DNA de cadena sencilla

Los virus de DNA de cadena sencilla (ssDNA) conforman el grupo II de la clasificación de Baltimore, su genoma consiste en una hebra de DNA(+) que puede pasar por un estado de doble hélice desde el cual se sintetiza el RNA mensajero (Baltimore, 1971). Algunos estudios han mostrado que los virus de ssDNA están presentes en gran número de hábitats, desde sistemas hidrotermales hasta el intestino de animales (Krupovic y Forterre, 2015). Están ampliamente distribuidos e incluyen patógenos económica, médica y ecológicamente importantes. Estos virus se distinguen por tener una estructura genómica mayoritariamente circular, con tamaños del genoma entre 1.7 y 25 kb, sus hospederos son bacterias, arqueas y la mayoría de grupos de eucariontes (Koonin *et al.*, 2021). Los virus ssDNA incluyen algunos de los genomas más pequeños, entre 2-6 kb de longitud, así como tasas de mutación alrededor de 10-6 s/n/c (sustituciones por nucleótido por célula) mucho mayores a los virus dsDNA (Sanjuán y Pilar, 2016). A pesar de que hay familias que pueden codificar solo una proteína estructural y una implicada en la replicación del DNA, también existen casos complejos, como la familia *Nanoviridae*, en la que el genoma completo está compuesto por 6-8 segmentos de ssDNA de ~1-kb (Carroll y Rein, 2016). Actualmente el comité internacional de taxonomía de virus clasifica a los virus ssDNA en 15 diferentes familias (Hulo *et al.*, 2020).

Virus CRESS DNA

El término “CRESS DNA” (*Circular Rep Encoding Single Stranded DNA*) se refiere al grupo de virus de DNA de cadena sencilla (ssDNA), al que pertenecen la mayoría de las familias (Zhao *et al.*, 2019). Estos virus codifican la proteína Rep, una helicasa que inicia la replicación vía *rolling circle*. Anteriormente se pensaba que tenían un ancestro común; sin embargo, Kazlauskas y colaboradores (2019) realizaron una red filogenética que sugiere por lo menos tres eventos de origen. Los virus CRESS DNA infectan eucariontes, arqueas y bacterias (Krupovic, 2013). La mayoría de las familias de virus de cadena sencilla (definidas por ICTV) tienen genomas circulares, exceptuando a las familias *Parvoviridae* y *Bidnaviridae*. Entre las familias de virus con DNA de cadena sencilla con genomas circulares, siete infectan organismos eucariontes, siendo *Anelloviridae* la única familia que no codifica proteínas Rep homólogas. Las seis familias de virus CRESS DNA de eucariontes son *Bacilladnaviridae, Circoviridae, Geminiviridae, Genomoviridae, Nanoviridae y Smacoviridae* (Zhao *et al.*, 2019).

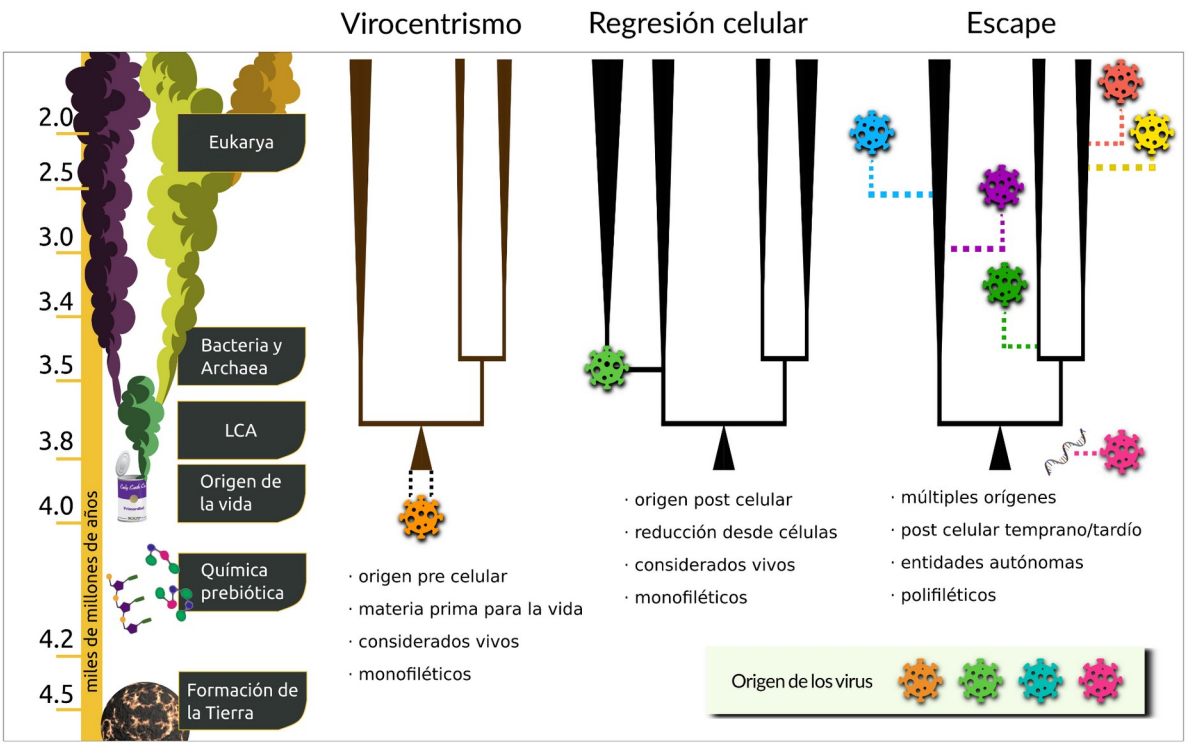


Figura 1. Hipótesis del origen de los virus. Se muestran las principales características del virocentrismo, la hipótesis de regresión celular y la hipótesis de escape, así como sus eventos en contexto del tiempo geológico, el origen y evolución temprana de la vida y la divergencia de los tres dominios. Línea de tiempo modificada de Becerra *et al.* 2007.

Los virus ssDNA y el estudio del origen de la vida

Los virus ssDNA podrían ser considerados antiguos en consonancia con las ideas virocéntricas, debido a que poseen genomas pequeños y simples, que cambian rápidamente (Sanjuán y Pilar, 2016). En contraste, existen indicios recientes que proponen múltiples orígenes tardíos a partir de escapes genómicos de la proteína Rep (Kazlauskas *et al.*, 2019). Sin embargo, no todos los virus ssDNA pertenecen a los CRESS DNA, por lo que dichas conclusiones podrían no extenderse a todas las familias.

Estrategias metodológicas para estudiar el origen de los virus

La principal herramienta para estudiar el origen de cualquier grupo biológico es la comparación, actualmente centrada en la genómica comparada, gracias a la creciente disponibilidad de datos genéticos. Si bien no debemos pensar que el material genético de un organismo nos brinda toda su información, el genoma es un gran recuento de su historia evolutiva y, dependiendo del enfoque, nos permite indagar desde procesos recientes a nivel de poblaciones, así como explorar su origen y evolución temprana (Anisimova, 2019).

Lo más común para el estudio evolutivo de un grupo es acotar la comparación a lo que llamamos un marcador, es decir, un elemento genético común y que deseablemente varía acorde a la profundidad temporal de las relaciones que se pretenden describir (Chenuil, 2006). Esto ha permitido, por ejemplo, la concepción de los tres dominios que

engloban a los seres vivos (Woese *et al.*, 1990), y es aquí donde radica la principal diferencia metodológica cuando los virus son el objeto de estudio (Moreira y López-García, 2009). A diferencia de los organismos celulares, los virus no comparten un marcador evolutivo universal.

Los virus son un grupo "politético", es decir, definido por elementos compartidos sólo entre una parte de los representantes (Van Regenmortel, 2018), pero ninguno universal. Esta condición resalta la historia polifilética de los virus, e implica que no debemos realizar la comparación de un solo elemento viral, sino dividir el problema en múltiples comparaciones.

La primera sección del presente trabajo se enfoca en definir conjuntos de elementos genéticos comparables entre los virus de DNA de cadena sencilla, a nivel de sus familias. Dichos marcadores servirán como punto de partida para buscar homólogos en otros linajes, tanto virales como celulares. Los fundamentos y propósitos generales de las herramientas utilizadas serán explicadas a continuación.

Pangenómica

En esencia, los análisis pangenómicos son procesos de agrupación o *clustering*. A partir de un conjunto de genomas, se identifican elementos presuntamente homólogos, partiendo de sus valores de similitud. Dichos valores de similitud se estiman mediante alineamientos pareados de secuencias de nucleótidos y aminoácidos.

El principal aporte metodológico de un análisis pangenómico, consiste en estructurar los alineamientos entre los elementos de un conjunto de genomas de tal manera que este no constituya un problema de alta complejidad computacional en el que se exploren todas las comparaciones posibles, imposible de realizar en la escala de genomas completos. Los algoritmos pangenómicos son de esta forma, mecanismos heurísticos para realizar las comparaciones pertinentes que permitan tomar en consideración todos los elementos genéticos de un conjunto de genomas, estos son: i) **BDBH**, toma en consideración un genoma de referencia para determinar los grupos de ortólogos y añadir uno a uno los elementos genéticos de la muestra. Requiere al menos de dos genomas a comparar ii) **COG Triangles**, efectúa comparaciones pareadas, fusionando los elementos con los mejores valores de alineamiento cuando estos coinciden en sistemas triangulares simétricos, por ello requiere de un mínimo de tres genomas a comparar. iii) **OMCL**, construye y agrupa una red mediante el algoritmo *Markov Cluster Algorithm*, en la que los nodos son los elementos genéticos de un genoma y los ejes sus valores de alineamiento.

El pangenoma es entonces, el conjunto de grupos de homólogos identificados, y se define como el repertorio genético de los individuos de un linaje (Vernikos *et al.*, 2015). Entre los principales objetivos del análisis pangenómico se encuentran: i) formación de grupos de secuencias de nucleótidos y proteínas homólogas a partir de algoritmos de alineamiento, ii) identificación de regiones intergénicas homólogas flanqueadas por

marcos abiertos de lectura ortólogos, y iii) cálculo de conjuntos superpuestos de proteínas. El último de estos objetivos es también el más común, y consiste en la clasificación de los grupos de homólogos de acuerdo con su prevalencia (Contreras-Moreira y Vinuesa, 2013). Los algoritmos que permiten este cálculo están basados en la teoría de conjuntos y a partir del conteo de los integrantes de un conjunto de ortólogos, nos permite responder, ¿qué porcentaje del total de genomas individuales aportaron un representante al grupo?. La respuesta por cada grupo de ortólogos nos permiten categorizarlos en:

- ***Core* o núcleo*:*** Genes presentes en todos los miembros del linaje, prevalencia del 100%
- ***Soft-core* o núcleo laxo*:*** Genes presentes en casi todos los miembros del linaje.
- ***Shell* o cubierta*:*** Genes ampliamente distribuidos, esenciales para los miembros de un subgrupo.
- ***Cloud* o nube*:*** Genes accesorios, específicos a un grupo de miembros del linaje

Además, la robustez de esta categorización pangenómica se sostiene en distintos supuestos, surgidos en el campo de la microbiología, por lo que deben ser adaptados tomando en cuenta las particularidades genéticas y evolutivas de los virus. En primer lugar, el análisis pangenómico considera que todos los genomas evaluados constituyen el repertorio genético total de una especie. Si bien este objetivo es más sencillo de cumplir con genomas celulares, entre los virus ssDNA existen familias con genomas segmentados, que requieren un manejo especial de los archivos genómicos. Por otra parte se asume que los elementos genéticos comunes mantienen similitudes a nivel de su secuencia de nucleótidos y/o aminoácidos. Sin embargo, los genomas de los virus cambian a un ritmo mucho mayor que los de las bacterias, particularmente cierto para los virus de RNA y ssDNA (Sanjuán y Pilar, 2016), por ello los parámetros de agrupamiento deben ser menos estrictos. Finalmente, no existen elementos genéticos compartidos por todos los virus ssDNA, por lo que el núcleo pangenómico sólo es aplicable a rangos taxonómicos debajo del nivel de familia. Debido a los reducidos tamaños de los virus ssDNA no es plausible la existencia de elementos genéticos “accesorios”, por lo que las definiciones de los grupos por prevalencia no deben extrapolarse literalmente a los virus (Figura 2).

Redes filogenéticas

En 1859, Charles Darwin desarrolló la metáfora del árbol de la vida, ilustrando la idea de que la diversidad biológica es producto de la descendencia con modificación y supervivencia diferencial. Las filogenias o árboles filogenéticos son, desde hace mucho tiempo, herramientas muy importantes para ilustrar las relaciones evolutivas entre las entidades biológicas, dibujando esquemas de nodos conectados por ejes que retroceden en el tiempo hasta ancestros comunes; de la misma manera que las ramas pequeñas de un árbol retroceden hasta las mismas ramas principales (Baum *et al.*, 2005). Desde el punto de vista de la teoría de grafos, los árboles filogenéticos son grafos acíclicos no dirigidos en los que dos nodos pueden estar conectados como máximo por un camino con el objetivo de representar relaciones de parentesco (Brandes y Cornelsen, 2009).

El concepto de *redes filogenéticas* engloba cualquier red utilizada para representar relaciones evolutivas entre un conjunto de entidades biológicas representados en los nodos de la red, generalmente los de ramas terminales (Rupp *et al.*, 2010). Comúnmente se utiliza para referirse a esquemas en los que los nodos pueden estar conectados por más de un camino, a diferencia de los árboles. El mayor número de posibilidades para establecer dichos caminos, en comparación a los árboles filogenéticos, presenta algunas ventajas cuando la evolución involucra relaciones reticuladas (Syvanen, 1985; Rieseberg, 1997), como duplicaciones genéticas, transferencia genética horizontal, pérdida de genes o segmentación genómica. Por la gran cantidad de flujo horizontal en la evolución de los genomas virales, dichas ventajas pueden reflejar sus complejas relaciones de forma más explícita (Koonin y Dolja, 2014).

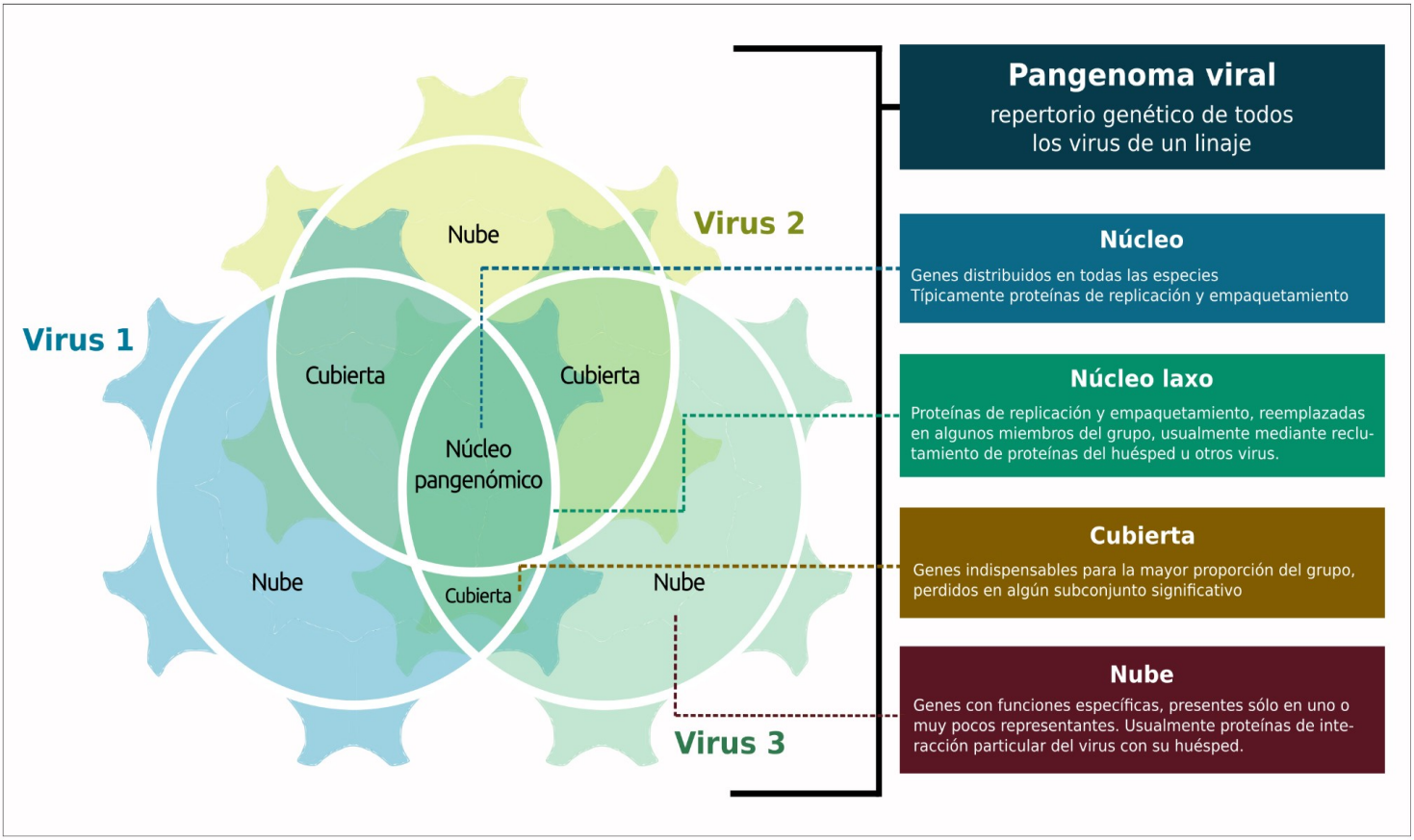


Figura 2. Adaptación del concepto de pangenoma viral. Esquema señalando las diferentes categorías del pangenoma viral, en función de la prevalencia de los grupos homólogos (izquierda). Definición del pangenoma viral, grupos por prevalencia y tipos de secuencias más comunes en estudios previos de pangenómica en virus de procariontes. Modificado de Kristensen *et al.*, 2013.

OBJETIVOS

General

El presente estudio tiene como objetivo principal identificar los grupos de genes ortólogos de los virus de DNA de cadena sencilla (ssDNA) y categorizarlos en función de su prevalencia. A partir de ello, obtener anotaciones funcionales e identificar dominios representativos de cada grupo de secuencias para extender búsquedas de homólogos en linajes remotos para indagar su historia evolutiva.

Particulares

- ➔ Identificar grupos de genes ortólogos en los genomas de referencia de virus ssDNA.
- ➔ Estandarizar el manejo de sus archivos genómicos, tomando en consideración las particularidades genómicas, hasta constituir una base de datos apta para el análisis pangenómico.
- ➔ Conocer la prevalencia de los grupos de genes homólogos en las familias de virus ssDNA, categorizándolos en núcleo, cubierta y nube pangenómica.
- ➔ Anotar funcionalmente los grupos de secuencias homólogas.
- ➔ Enriquecer los atributos conocidos de los grupos de homólogos recopilando información de sus dominios.
- ➔ Elaborar un catálogo de secuencias representativas de los dominios de cada grupo de genes homólogos.
- ➔ Extender búsquedas de genes homólogos fuera de los genomas de virus ssDNA, incorporando tanto otros grupos virales como organismos celulares.
- ➔ Representar mediante redes y filogenias las similitudes entre los homólogos encontrados para el caso de la helicasa *Rep*, comparándolo con las topologías afines a las distintas hipótesis del origen de los virus.

ANTECEDENTES

De acuerdo con los criterios de clasificación del Comité Internacional de Taxonomía de Virus (**International Committee on Taxonomy of Viruses o ICTV**; Lefkowitz *et al.*, 2017), las familias de virus ssDNA son monofiléticas. Las aproximaciones evolutivas se han centrado en comparaciones de las estructuras primarias de proteínas implicadas en la replicación (Ilyna y Koonin, 1992; Kuprovic y Koonin, 2014), organización genómica (Roux *et al.*, 2013) y estructuras de sus cápsides (Chapman y Rossmann, 1993).

En el campo metodológico de la pangenómica existen sólo algunos antecedentes directos de virus. Un análisis enfocado en detectar grupos de genes ortólogos de virus de procariontes (Kristensen *et al.*, 2013), incluyendo en su muestra a las familias de virus ssDNA *Inoviridae* y *Microviridae,* permitió detectar genes de dichas familias sin homólogos evidentes en otros grupos virales o de sus hospederos. La mayoría de las incorporaciones de la pangenómica en estudios sobre virus se han orientado a virus de DNA de doble cadena. Ha sido descrito que, en respuesta a presiones selectivas, algunos virus de insectos pueden adquirir genes accesorios (Brito *et al.*, 2016); así como también se han encontrado variaciones geográficas del genoma accesorio de virus de amebas (Assis *et al.*, 2015).

Respecto a las hipótesis del origen de virus ssDNA existen trabajos que describen una alta complejidad en sus relaciones. Destaca la postulación de eventos de transferencia horizontal, desde diversos grupos de virus dsDNA y de RNA de cadena sencilla que dieron origen a la familia *Bidnaviridae* (Kuprovic y Koonin, 2014), y desde virus ssRNA en la evolución de la cápside de la familia *Bacilladnaviridae* (Kazlauskas *et al.*, 2017).

La reciente incorporación de herramientas como la construcción de redes de similitud (Krupovic *et al.*, 2009; Krupovic, 2013; Kazlauskas *et al.*, 2019), ha posibilitado describir imbricadas interacciones evolutivas de elementos genéticos de virus CRESS DNA con elementos del *mobiloma* bacteriano. El estado del conocimiento actual sugiere que estos eventos corresponden a la hipótesis de escape.

El estudio de la evolución temprana de los virus es una línea del Laboratorio de Origen de la Vida, UNAM. Entre sus aportes al presente trabajo, destacan: indicios de homología entre proteínas virales y celulares (Campillo-Balderas, 2018), uso de estructuras cristalográficas en comparaciones de homólogos (Jácome *et al.*, 2015) y métricas genómicas como indicativos de la antigüedad de los linajes virales (Campillo-Balderas *et al.*, 2015).

METODOLOGÍA

El código implementado, así como las dependencias a otros repositorios se encuentran disponibles bajo una licencia de uso MIT a través del repositorio de github https://github.com/arqueaodv/ssDNA_viral_pangenomics. Este incluye una guía textual y visual para su implementación.

Configuración de la base de datos: descarga, concatenado y filtrado

Se incorporaron las familias virales del reino Monodnaviria, del 9° reporte del Comité Internacional de Taxonomía de Virus (Lefkowitz *et al.*, 2017), ICTV, por sus siglas en inglés. Para ello se realizaron búsquedas en el sitio web del NCBI (National Center for Biotechnology Information; https://www.ncbi.nlm.nih.gov/) con la siguiente estructura booleana:

- **Criterios de inclusión:**
  - `Familia [ORGANISM]`
  - `AND srcdb_refseq[PROP]`
- **Criterios de exclusión:**
  - `NOT wgs[prop]`
  - `NOT cellular organisms[Organism]`
  - `NOT AC_000001:AC_999999[pacc]`

- **Búsqueda general:**

```
Family [ORGANISM] AND srcdb_refseq[PROP] NOT wgs[prop] NOT cellular organisms[Organism] NOT
                                    AC_000001:AC_999999[pacc]
```

Esto permitió obtener genomas del recurso de secuencias virales de NCBI Refseq (Brister *et al.*, 2015), excluyendo datos provenientes de *whole genome shotgun* (WGS) y *high throughput sequencing* (HTS). El resultado de búsqueda por familia se descargó en un archivo único en formato *genbank*, que fue dividido para cada especie.

Hasta este punto se contaba con genomas completos de la mayoría de las familias; sin embargo, entre los *Nanoviridae*, *Bidnaviridae* y *Geminiviridae* existen representantes con genomas segmentados (Hulo *et al.*, 2020). En estos casos, sus archivos de segmentos fueron agrupados en archivos genómicos mediante su identificador taxonómico del NCBI. Este fue un proceso indispensable para el análisis pangenómico, pues opera bajo el supuesto de que la carpeta de entrada contiene genomas completos y de no cumplirse favorece la subestimación de la prevalencia de los grupos de homólogos.

Los genomas tanto en formato *GenBank*, como en formato *Fasta* cuentan con elementos textuales que indican el inicio de un producto proteico. Sin embargo, existen errores tipográficos que pueden caer sobre estas regiones del archivo y que impiden distinguir correctamente los atributos de las secuencias. Si bien estos son raros, las variaciones incluso de una proteína son altamente significativas cuando tomamos en

consideración los reducidos tamaños de los genomas de virus ssDNA.Para mitigar su efecto, se contaron las proteínas por genoma, descartando aquellos que variaron en ± el 50% del conteo más observado por familia y los conteos para las cepas de referencia.

Agrupamientos pre-pangenómicos

Por la escala del estudio, particularmente en las familias con cientos de genomas, consideramos la generación de subgrupos, con el fin de evitar la subestimación del núcleo pangenómico comparando virus muy distantes.

Para ese propósito se elaboraron matrices de distancias entre pares. Por su alto puntaje en pruebas de rendimiento (Zielezinski *et al.*, 2019), fue utilizado el algoritmo *alignment-free* de comparación *Cumulative Power Spectrum* (Pei et al., 2019). Este método construye un vector 28-dimensional a partir de la transformación de la secuencia de nucleótidos a señales de covarianza y su posterior descomposición en componentes oscilatorios mediante Transformada Rápida de Fourier (FFT). El vector construido para una secuencia es siempre el mismo por lo que es un método determinista que facilita la inclusión de nuevas secuencias disponibles.

Las matrices de disimilitud de cada familia fueron analizadas para la determinación del mejor número de grupos y algoritmo de agrupamiento. En un *pipeline* de decisión del paquete de *R, NbClust* (Charrad *et al.*, 2014), se estimaron múltiples índices de evaluación, pruebas estadísticas y métodos gráficos, De esta forma se generaron los subgrupos más relevantes y se generaron consensos absolutos. Los subgrupos con menos de 5 representantes fueron considerados *outliers* y descartados de los análisis posteriores al no sobrepasar el umbral de tamaño muestral requerido por los algoritmos pangenómicos. El paquete de *R*, taxize (Chamberlain y Szöcs, 2021) fue utilizado para anotar taxonómicamente a los integrantes de cada subgrupo.

Obtención de pangenomas

El *software* utilizado para el análisis pangenómico fue *GET_HOMOLOGUES* ver. 3.4.3 (Contreras-Moreira y Vinuesa, 2013), sirviendo como *input* cada uno de los subgrupos aptos. El rigor de la agrupación fue aumentado (opción **-D**), requiriendo una composición similar de dominios de Pfam (El-Gebali *et al.*, 2019). Los parámetros para considerar a dos secuencias parte del mismo grupo fueron modificados a un *query coverage* mínimo de 50% (**-C 50**) y un *e-value* máximo de 1e-3 (**-E 0.001**) , ambos más laxos que los valores por defecto y ajustados a partir de ensayos iniciales. La identificación de dominios se realizó con HMMER3 (Howard Hughes Medical Institute, 2020), y los algoritmos de ortología seleccionados fueron COGtriangles (Kristensen *et al.*, 2010) y OrthoMCL (Li *et al.*, 2003), opciones **-G** y **-M** respectivamente. Se incorporaron además las banderas **-t 0** para solicitar al programa reportar todos los posibles *clusters*, y la bandera **-z** para incluir un análisis de *soft core*.

El comando general de uso de GET_HOMOLOGUES fue:

```
~$ get_homologues.pl -d $input -D -G/M -n 12 -t 0 -z -C 50 -E 0.001
```

Posteriormente, con el programa **compare_clusters.pl** se construyeron matrices de intersección de pangenomas, para comparar entre dos iteraciones.

Para cada grupo de homólogos se obtuvo un valor de prevalencia, definido como el porcentaje de virus aportando una secuencia, respecto al total del subgrupo. Haciendo uso de este valor de prevalencia, se asignaron a categorías pangenómicas: núcleo >98%, núcleo laxo >90%, cubierta >30% y nube <30%. Una vez obtenidos los pangenomas con los algoritmos COGtriangles, *OrthoMCL*, sus intersecciones y categorías por prevalencia, cada conjunto de secuencias homólogas fue analizado funcional y evolutivamente. La Figura 3 muestra el diagrama de flujo desde la descarga de los genomas hasta la obtención de pangenomas.

Descripción funcional de los pangenomas

El principal objetivo de este enfoque fue la asignación de funciones a los grupos homólogos, con términos de *Gene Ontology* (Ashburner *et al.*, 2000;The Gene Ontology Consortium, 2019,2021). Este se abordó mediante dos estrategias: anotación basada en similitud y anotación por asociación estadística de términos. Las principales etapas y procesos de la anotación funcional se resumen en el diagrama de flujo de la Figura 4.

Anotación basada en similitud

Se construyó una base de datos con proteínas virales de las cepas de referencia, nombrada localmente " **Viralrefs** ", para las que se tienen asociados términos de *Gene Ontology* (GO), *InterPro* (Blum *et al.*, 2021), *Pfam* (El-Gebali *et al.*, 2019) y PDB (Berman *et al.*, 2000). Las cepas de referencia fueron seleccionadas desde el sitio *web* de *Viralzone* (Hulo *et al.*, 2020). Las secuencias y sus metadatos fueron descargadas de *UniProt* (The UniProt Consortium, 2021) en formato *fasta*.

Cada proteína de la base de secuencias inicial, no anotada funcionalmente, fue comparada contra Viralrefs mediante *Protein-Protein BLAST 2.9.0* (Camacho *et al.*, 2009), parámetros por defecto. Se descartaron todos aquellos *hits* con valor de *e-value* por encima de 1e-10 y porcentaje de identidad menor a 25%. Entre las secuencias con resultados positivos, para aquellas con más de una anotación posible, los mejores *bitscores* de los contendientes fueron normalizados respecto al máximo y se tomaron como válidos al 10% mejor puntuado. Dentro de cada grupo de homólogos, se realizó un conteo de los términos de GO asignados a sus integrantes.

Anotación por asociación estadística de términos

En primer lugar, se listaron todos los dominios *pfam* encontrados entre las secuencias ssDNA, haciendo uso del programa *hmmscan* (Howard Hughes Medical Institute, 2020). Por cada dominio encontrado se realizó una consulta en dos recursos: 1) ***pfam2GO*** (Blum *et al.*, 2021), un mapeo específico de dominios de *pfam* a términos de *GO* a través de co-ocurrencias, y 2) **GODomainMiner (GODM)** (Alborzi *et al.*, 2017) un modelo estadístico para inferir de relaciones entre *pfam* y *GO*, basado en "vecinos comunes" y que clasifica estas asociaciones según su certeza en "oro", "plata" y "bronce". En caso de ambigüedad en la anotación, se priorizó en el siguiente orden: GODM oro, pfam2GO, GODM plata, GODM bronce.

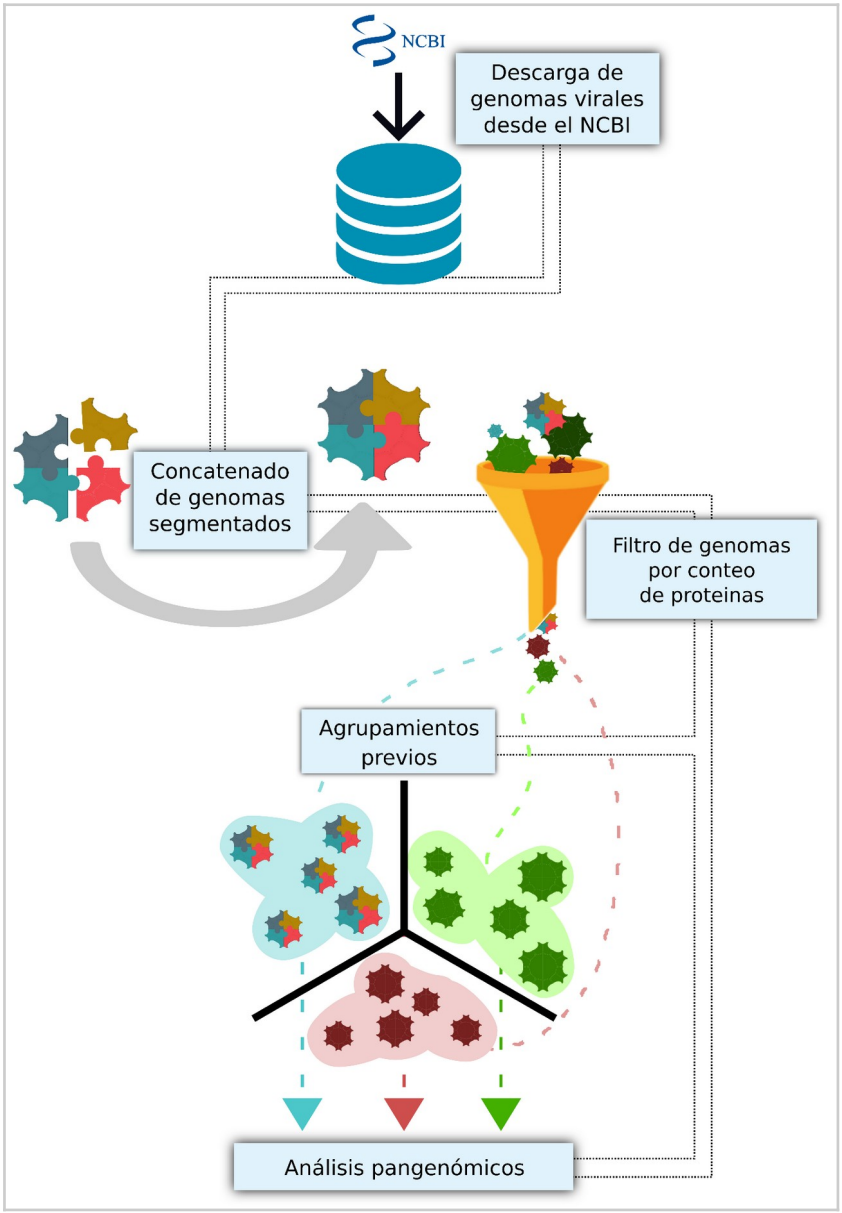


Figura 3. De la descarga de genomas a los análisis pangenómicos. Diagrama de flujo ilustrando los principales procesos y etapas desde la obtención de genomas virales de referencia en NCBI, manipulación y filtrado de los datos, hasta el uso de ***Get_homologues***.

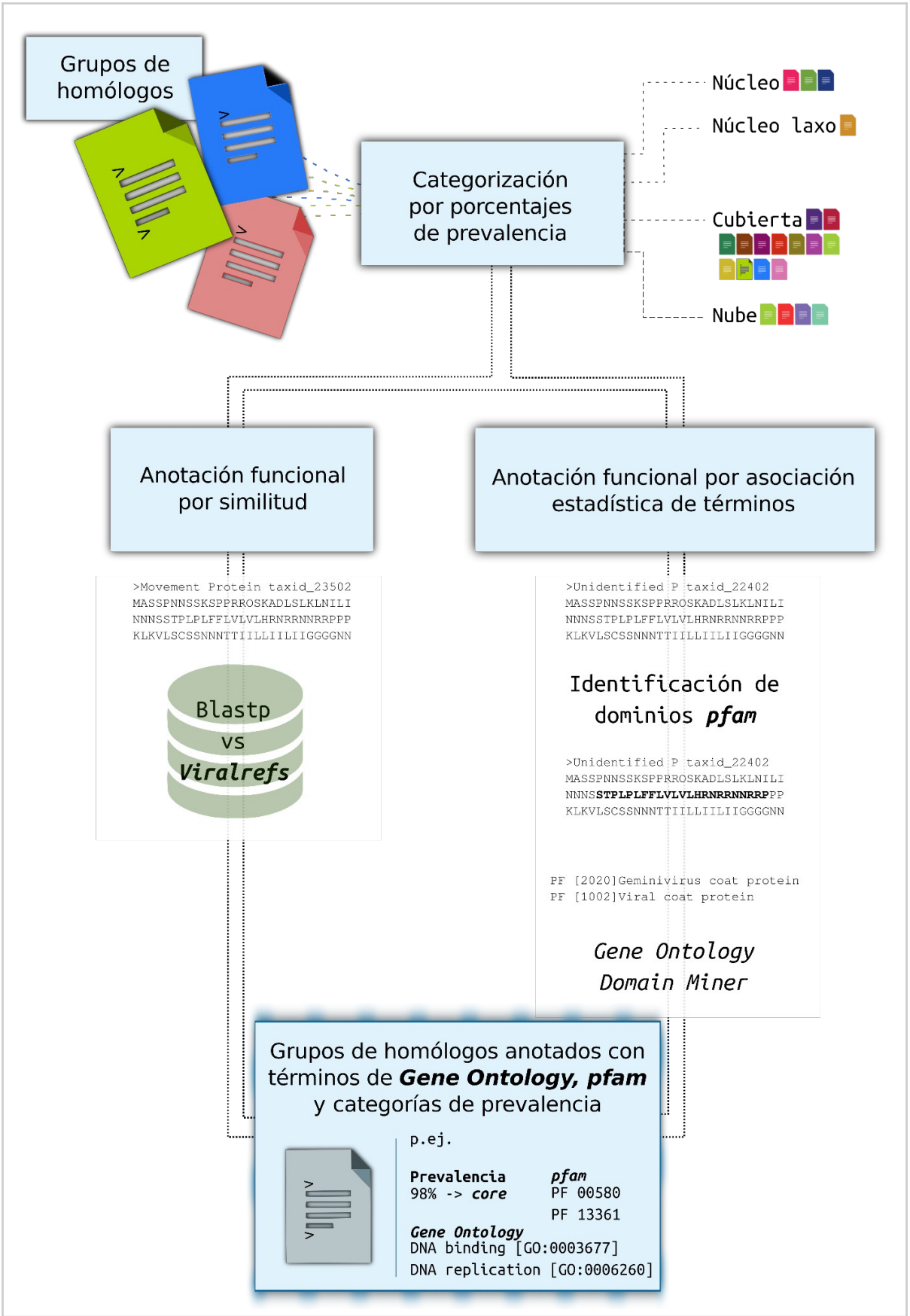


Figura 4. Diagrama de flujo de la anotación funcional. Se muestran los pasos seguidos en las dos estrategias de anotación funcional: Anotación basada en similitud y Anotación por asociación estadística de términos.

Análisis evolutivo de los pangenomas

Para estudiar el posible origen de los grupos de homólogos virales, se construyeron filogenias incluyendo homólogos virales y homólogos celulares. La primera tarea consistió en encontrarlos, para lo cual se extendieron búsquedas a nivel de dominio, cuyo diagrama de flujo se muestra en la Figura 5 y se explica a continuación.

Segmentación de proteínas por dominio

Los grupos de secuencias homólogas existen como archivos en formato *fasta*. Las proteínas dentro de uno de estos archivos pangenómicos comparten su estructura de dominios; sin embargo, no es de esperarse que otros homólogos remotos mantengan su composición. Por esta razón, la búsqueda de homólogos en otros grupos virales y celulares no tomó como punto de partida las proteínas completas, sino cada uno de los dominios. Para ello, los archivos *fasta* de proteínas completas fueron descompuestos en archivos cortados a cada dominio.

Selección de representantes

Con la finalidad de disminuir el tiempo de cómputo y evitar la redundancia, las búsquedas de homólogos partieron de representantes por dominio. Para seleccionarlos, los representantes de cada dominio de un mismo grupo de homólogos fueron alineados de forma óptima con *Biostrings* (Pagès *et al.*, 2021) y se calcularon los medioides[2], que fueron tomados como los representantes.

Búsqueda de homólogos virales y celulares

Los homólogos se recuperaron realizando búsquedas en dos bases de datos de secuencias de proteínas filtradas al 50%:

- ***UniRef50*** **(Suzek et al., 2014).** Las bases de datos *UniRef* proporcionan un conjunto de secuencias agrupadas, provenientes de *UniProt* y *UniParc*, seleccionados para brindar una cobertura completa del espacio de secuencias conocido. *UniRef50* es la síntesis de esa enorme diversidad a una redundancia de 50%. Descargada vía ftp desde `ftp.uniprot.org/pub/databases/uniprot/uniref/uniref50/`
- ***ssDNA50.*** Fue construida localmente a partir de la muestra completa de secuencias de virus de DNA de cadena sencilla del recurso de genomas virales de NCBI (Brister *et al.*, 2015). En dicho conjunto se redujo la redundancia al 50% mediante cd-hit (Li y Godzik, 2006) .

Por cada secuencia representativa se realizaron cinco iteraciones de *jackhmmer* (Eddy, 2011) contra ambas bases de datos. Los resultados positivos o *hits* de cada búsqueda fueron etiquetados con su familia en caso de ser virales, las animales como Metazoa, las fúngicas como Fungi, las del supergrupo Archaeplastida con su phylum y las procariotas por su dominio.

2.- Un medioide es un elemento representativo de un conjunto de datos. La suma de sus diferencias con todos los demás elementos del grupo es la mínima.

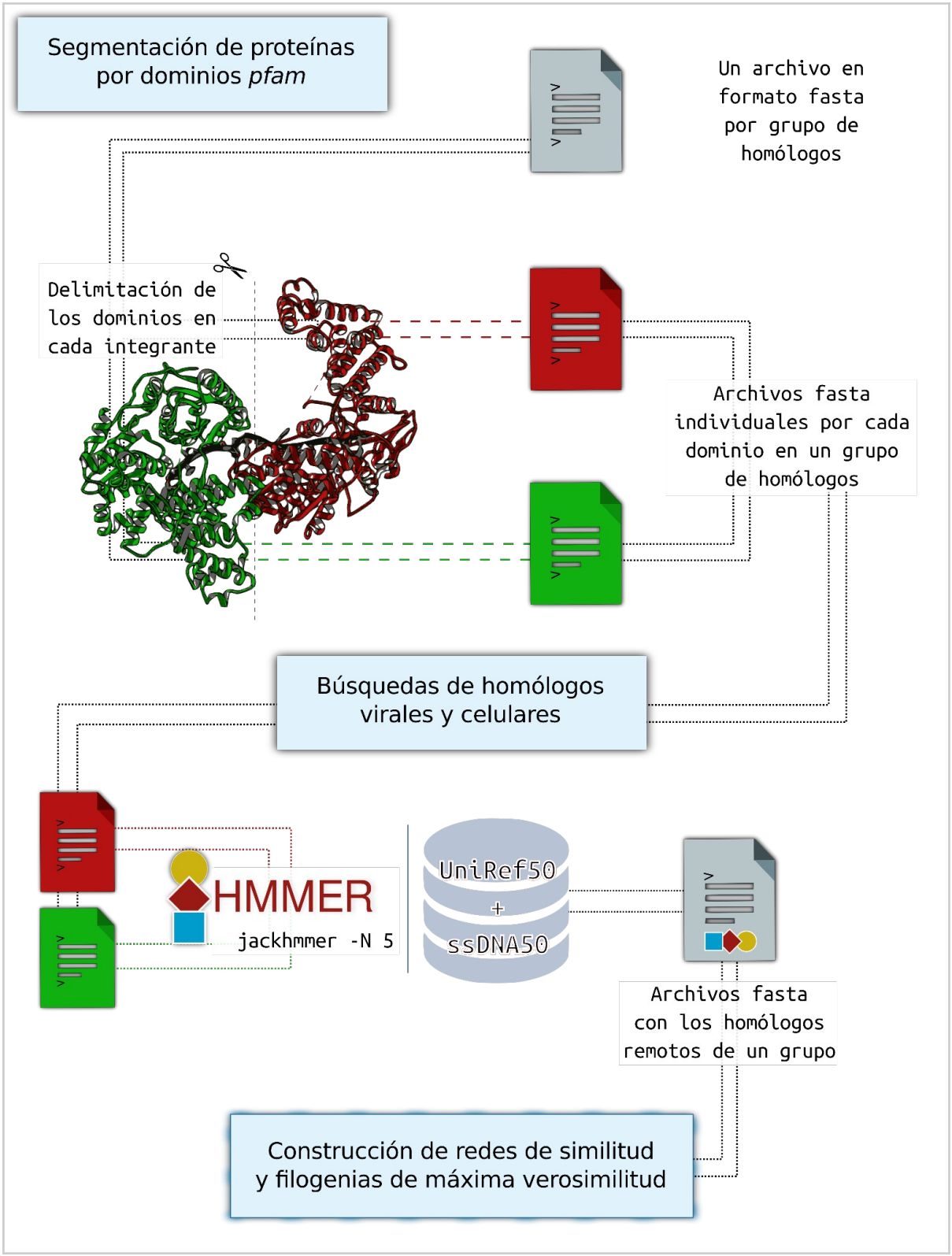


Figura 5. Diagrama de flujo del enfoque evolutivo. Se resumen de forma gráfica los pasos hasta la búsqueda de homólogos remotos mediante el uso de jackhmmer.

Redes de similitud

Los homólogos virales y celulares resultantes de cada búsqueda se analizaron con *CLANS* (Frickey y Lupas, 2004), una implementación del algoritmo de *layout* Fruchterman–Reingold que simula un espacio multidimensional donde las proteínas se atraen o repelen de acuerdo con su similitud (*P-value*), construyendo una red de similitud. La red generada fue estructurada mediante *clustering* jerárquico. Los grupos "transversales", es decir que contienen tanto secuencias de diversas familias virales, como aquellas que incorporaron representantes celulares fueron seleccionados para construir filogenias.

Filogenias de Máxima Verosimilitud

Se realizaron alineamientos múltiples de las secuencias de aminoácidos con *MAFFT* (Katoh *et al.*, 2002), métodos precisos (G-INS-i, L-INS-i y E-INS-i) para conjuntos con menos de 600 secuencias; y método automático para conjuntos con más. Los alineamientos resultantes se procesaron con *TrimAl* (Capella-Gutiérrez *et al.*, 2009), con la opción **-automated1** optimizada para máxima verosimilitud. Los modelos evolutivos fueron estimados haciendo uso de *ModelFinder* (Kalyaanamoorthy *et al.*, 2017). El soporte de las filogenias se estimó con *Ultrafast Bootstrap* de 1000 réplicas (Minh *et al.*, 2013), inferencia de topologías mediante *IQ-TREE 2* (Minh *et al.*, 2020). Los árboles consenso, para describir las historias evolutivas contadas por los grupo de ortólogos virales y sus homólogos en linajes distantes, fueron contrastados con las distintas hipótesis respecto al origen de los virus.

RESULTADOS

A continuación, se muestran las principales tablas y figuras que soportan las conclusiones alcanzadas en el proyecto. Debido al volumen de genomas analizados, algunas muestran lo obtenido solo para familias específicas. En dichos casos se indica el grupo seleccionado, y se describe lo observado contemplando el panorama general.

Los resultados correspondientes a la búsqueda de homólogos remotos y sus análisis subsecuentes (redes de similitud y árboles filogenéticos) corresponden a lo obtenido hasta un corte realizado el día 10 de enero de 2022. Consideramos necesario mencionar que lo anterior no debe interpretarse como un incumplimiento de los objetivos, pues se cuenta con el catálogo pangenómico completo, sino como una extensión en curso de los alcances del proyecto. Todos los resultados crudos del análisis se encuentran disponibles en carpetas por familia en el repositorio antes mencionado, mediante el enlace:

https://github.com/arqueaodv/ssDNA_viral_pangenomics/tree/main/results.

Descripción de la base de datos

Se obtuvo un total de 1796 secuencias iniciales (Tabla 1R), catalogadas como genomas de referencia para integrantes de las familias ssDNA. La concatenación de los genomas segmentados de las familias *Bidnaviridae*, *Geminiviridae* y *Nanoviridae*, redujo esta cantidad a 1533 genomas completos. Finalmente, el filtrado a partir del número de proteínas supuso una reducción de la muestra hasta 1445 genomas finales incorporados en el análisis pangenómico (Tabla 1R). La distribución de dichos genomas se concentra en las familias *Geminiviridae* (512), *Circoviridae*(191) y *Parvoviridae* (143). Por el contrario *Bidnaviridae* y *Spiraviridae* fueron descartadas del análisis pangenómico por contar con menos de cinco representantes.

De acuerdo con sus hospederos (Figura 1R), los virus CRESS DNA abarcan ~80% de las secuencias. Los hospederos mejor representados son las plantas, vertebrados e invertebrados, y por el contrario destaca la poca disponibilidad de virus de arqueas. Esta disponibilidad de datos afectó directamente qué tan estricta fue la variación permitida para el fiiltrado por conteo (Tabla 2R), requiriendo ser más permisivos en las familias con menos genomas para no perder su muestra por completo. Los rangos más amplios se observaron en las familias de fagos *Pleolipoviridae, Microviridae e Inoviridae.*

Tabla 1R. Conteo de secuencias de la base de datos. Se muestran los números totales correspondientes a cada familia, así como el porcentaje de reducción por concatenación y filtrado por conteo de proteínas.

Familia	Genomas descargados	Porcentaje de reducción por concatenado	Genomas concatenados	Porcentaje de reducción por filtrado	Genomas finales
Alphasatellitidae	100	4%	96	3%	93
Anelloviridae	108	3%	105	7%	98
Bacilladnaviridae	9	0%	9	11%	8
Bidnaviridae	**4**	**50%**	**2**	**0%**	**2**
Circoviridae	217	1%	215	11%	191
Geminiviridae	704	24%	538	5%	512
Genomoviridae	108	1%	107	3%	104
Inoviridae	45	4%	43	2%	42
Microviridae	62	0%	62	8%	57
Nanoviridae	91	87%	12	0%	12
Parvoviridae	146	2%	143	0%	143
Pleolipoviridae	14	0%	14	29%	10
Smacoviridae	51	2%	50	2%	49
Spiraviridae	**1**	**0%**	**1**	**0%**	**1**
Tolecusatellitidae	136	0%	136	10%	123
Total	1796	15%	1533	6%	1445

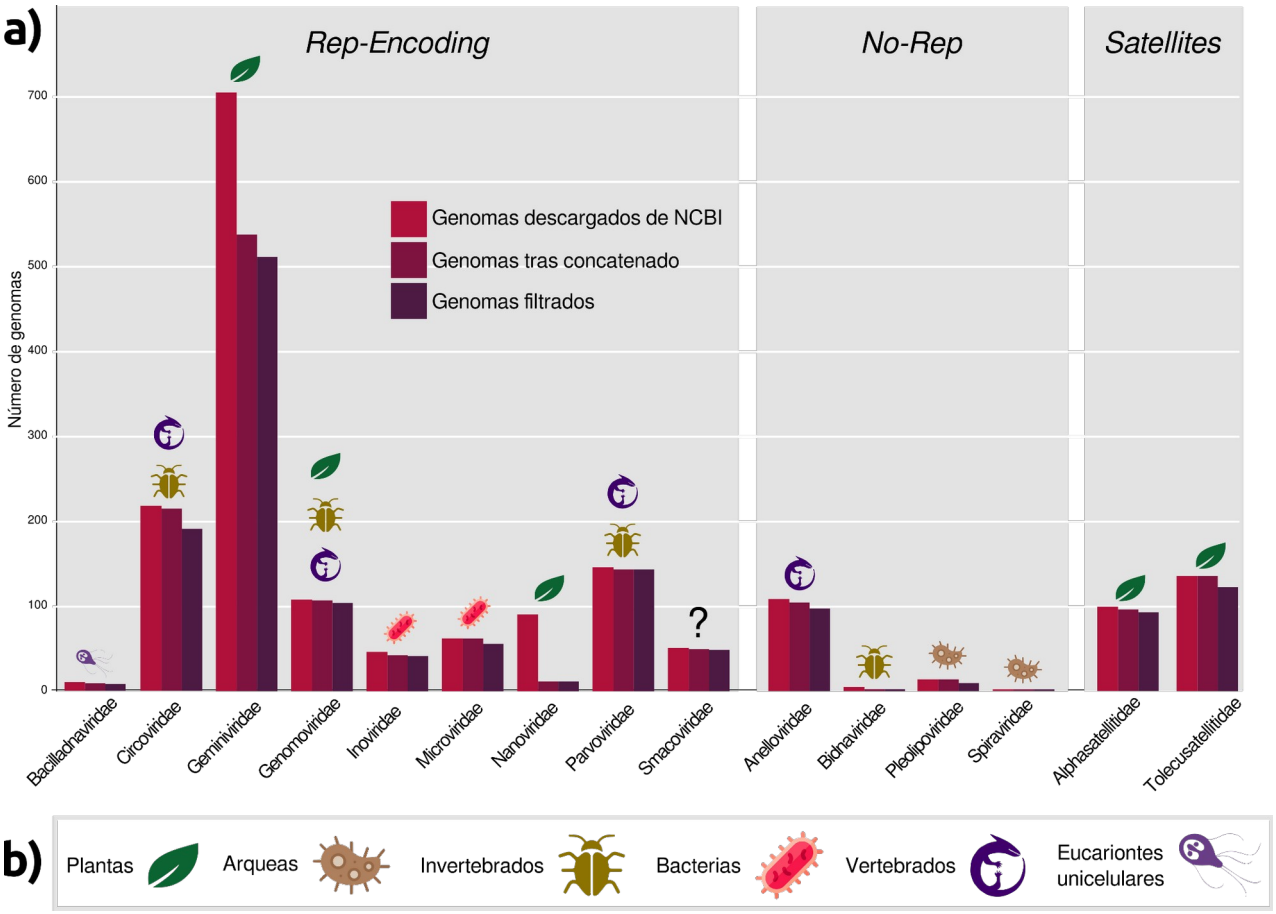


Figura 1R. Distribución de las secuencias en la base de datos de los principales grupos ssDNA. a) Se muestran los números totales correspondientes a cada familia, agrupados por la expresión o no de la helicasa Rep. **b)** Linaje principal de hospederos por familia viral.

Tabla 2R. Valores de corte en el filtro por conteo de proteínas. Se muestran los parámetros considerados para la determinación de los umbrales por cada familia. Fueron tomados en cuenta el número de proteínas entre las cepas de referencia, así como el promedio dentro de la muestra.

Familia	Mínimo observado en cualquier cepa de referencia	Máximo observado en cualquier cepa de referencia	#promedio de proteínas	Variación permitida ±%	Conteo máximo permitido	Conteo mínimo permitido	Porcentaje de reducción por filtrado
Alphasatellitidae	1	2	**2**	50	**3**	**1**	3%
Anelloviridae	2	4	**4**	30	**6**	**2**	7%
Bacilladnaviridae	3	6	**4**	95	**8**	**1**	11%
Circoviridae	2	3	**3**	50	**5**	**2**	11%
Geminiviridae	4	8	**7**	50	**11**	**4**	5%
Genomoviridae	2	4	**3**	50	**5**	**2**	3%
Inoviridae	9	15	**11**	50	**17**	**6**	2%
Microviridae	8	19	**8**	50	**19**	**4**	8%
Nanoviridae	7	9	**8**	50	**12**	**4**	0%
Parvoviridae	3	9	**4**	50	**9**	**2**	0%
Pleolipoviridae	9	15	**17**	50	**26**	**9**	29%
Smacoviridae	6	6	**3**	50	**6**	**2**	2%
Tolecusatellitidae	1	1	**1**	50	**2**	**1**	10%

Agrupamientos pre-pangenómicos

La determinación de los métodos de *clustering* concluyó en su mayoría con el algoritmo *complete* (Tabla 3R), que favorece un menor número de subgrupos con diferenciación completa. No se observaron patrones evidentes en la selección de métodos.

Tabla 3R. Resumen de los pre-agrupamientos por familia. Se indican los métodos seleccionados mediante el paquete de R, el número de subgrupos por familia y cuántos de ellos son aptos para el análisis pangenómico.

Familia	Método de clustering	Número de subgrupos	Aptos para pangenómica (# de subgrupos con más de cinco genomas)
Alphasatellitidae	*complete*	5	2
Anelloviridae	*complete*	5	3
Bacilladnaviridae	*complete*	2	1
Circoviridae	*complete*	3	3
Geminiviridae	*complete*	2	2
Genomoviridae	*centroid*	12	3
Inoviridae	*centroid*	7	3
Microviridae	*single*	4	2
Nanoviridae	*complete*	3	1
Parvoviridae	*median*	4	2
Pleolipoviridae	*mcquitty*	2	1
Smacoviridae	*median*	7	2
Tolecusatellitidae	*single*	4	1

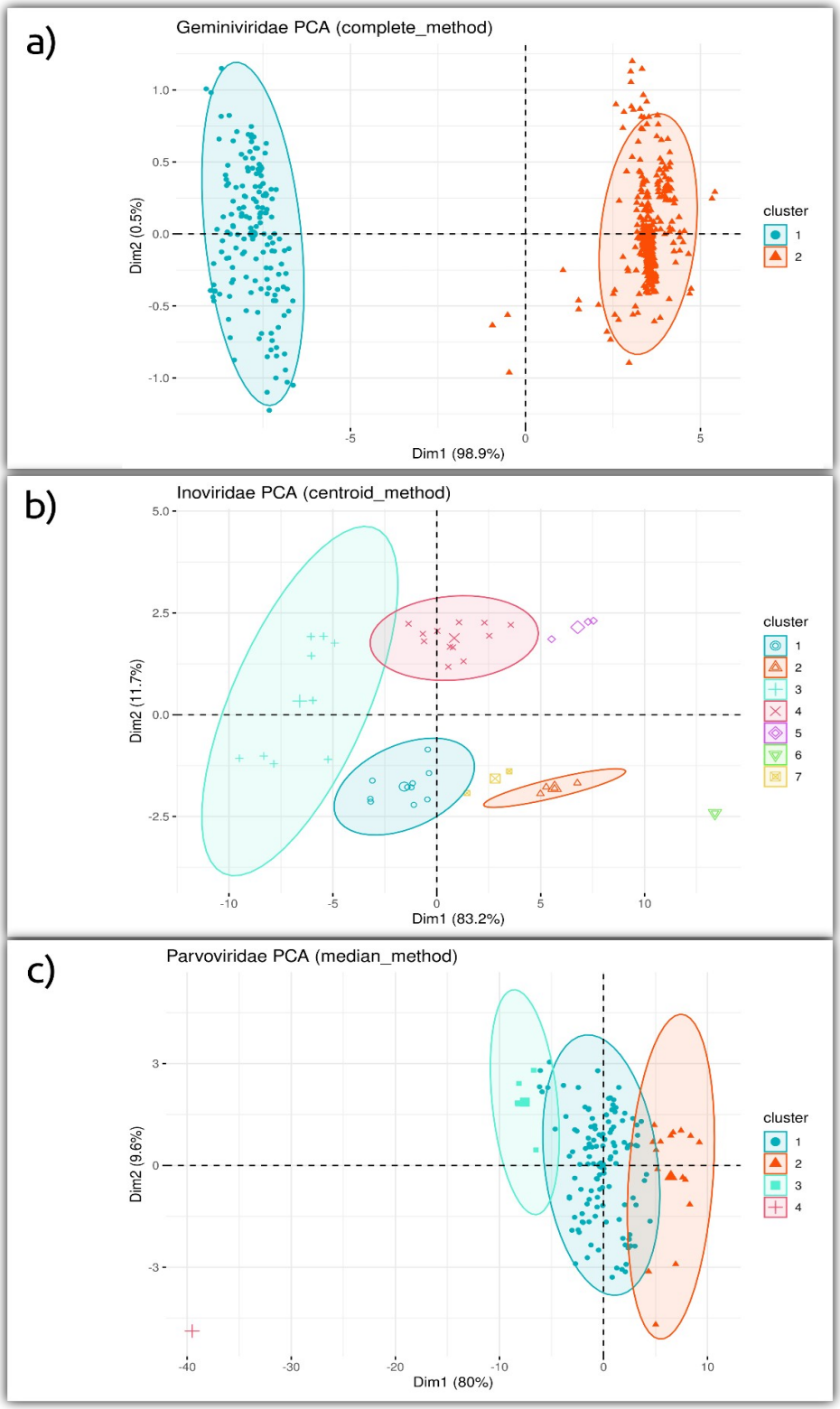


Figura 2R. Esquemas de partición generales. Análisis de componentes principales de los distintos escenarios de partición observados al interior de algunas familias ssDNA, cada punto en los PCA representa un genoma, las elipsoides delimitan los clusters de genomas encontrados. **a)** Segregación completa, sin solapamiento entre los rangos de inclusión. **b)** Segregación mixta, con solapamiento entre algunos subgrupos. **c)** Segregación incompleta, sin claro distanciamiento entre los límites de cada subgrupo.

Las familias mostraron tres escenarios generales de esquemas de partición (Figura 2R), de mayor a menor segregación estos son:

- **Segregación completa** (Figura 2R a)**:** Los genomas al interior de las familias *Alphasatellitidae*, *Anelloviridae*, *Nanoviridae*, *Bacilladnaviridae*, *Circoviridae* y *Geminiviridae*, se agruparon en *clusters* cuyas elipsoides no se traslapan. De manera coincidente, este fue el caso de todas las familias con genomas multipartitos.

- **Segregación mixta** (Figrua 2R b): Los genomas al interior de las familias *Genomoviridae*, *Inoviridae*, *Microviridae*, *Pleolipoviridae*, *Smacoviridae*, *Tolecusatellitidae* se agruparon en *clusters* cuyas elipsoides se encontraban en algunos casos.

- **Segregación incompleta** (Figura 2R c): Únicamente los genomas de la familia *Parvoviridae* dispusieron segregación incompleta, es decir que las elipsoides de los subgrupos se traslapan. Además, este esquema de partición es incompatible con la clasificación de los integrantes de esta familia.

Pangenomas

El total de conjuntos analizados por GET_HOMOLOGUES fue de 24, obteniendo un total de 48 entre los algoritmos pangenómicos COG y OMCL. El pre-agrupamiento de los genomas tuvo impactos positivos sobre la estimación del núcleo y núcleo laxo pangenómico (Figura 3R y Figura 4R). Para la familia *Geminiviridae* y *Genomoviridae*, la manipulación previa determinó la existencia del núcleo pangenómico aunado a un aumento considerable de la proporción correspondiente al núcleo laxo (Figura 3R).

Destacan los casos de las familias *Parvoviridae*, *Pleolipoviridae* e *Inoviridae*, para las cuales los parámetros de análisis utilizados no concluyeron en un núcleo pangenómico (Tabla 4R). De forma contrastante, en la familia *Nanoviridae* todos los grupos de homólogos fueron asignados al núcleo. Como se esperaba, el número de grupos ortólogos que conforman las categorías pangenómicas varían entre familias (Tabla 4R).

Otro impacto del manejo y filtrado de los genomas iniciales se dio sobre la consistencia de los algoritmos COG y OMCL (Figura 4R); pues los grupos de homólogos hallados por ambos aumentaron en compatibilidad y consistencia, aunque el número total se redujo.

Tabla 4R. Número de grupos ortólogos por categoría pangenómica. Se muestra el conteo por categoría pangenómica de los grupos ortólogos de cada familia.

Familia	Core (# grupos ortólogos)	Softcore (# grupos ortólogos)	Shell (# grupos ortólogos)	Cloud (# grupos ortólogos)
Alphasatellitidae	0	0	3	8
Anelloviridae	1	2	35	18
Bacilladnaviridae	1	1	7	2
Circoviridae	0	1	21	8
Geminiviridae	4	8	3	13
Genomoviridae	2	3	9	4
Inoviridae	0	0	55	0
Microviridae	10	2	16	21
Nanoviridae	6	0	0	0
Parvoviridae	0	0	31	25
Pleolipoviridae	1	4	18	0
Smacoviridae	3	1	0	5
Tolecusatellitidae	0	1	0	1

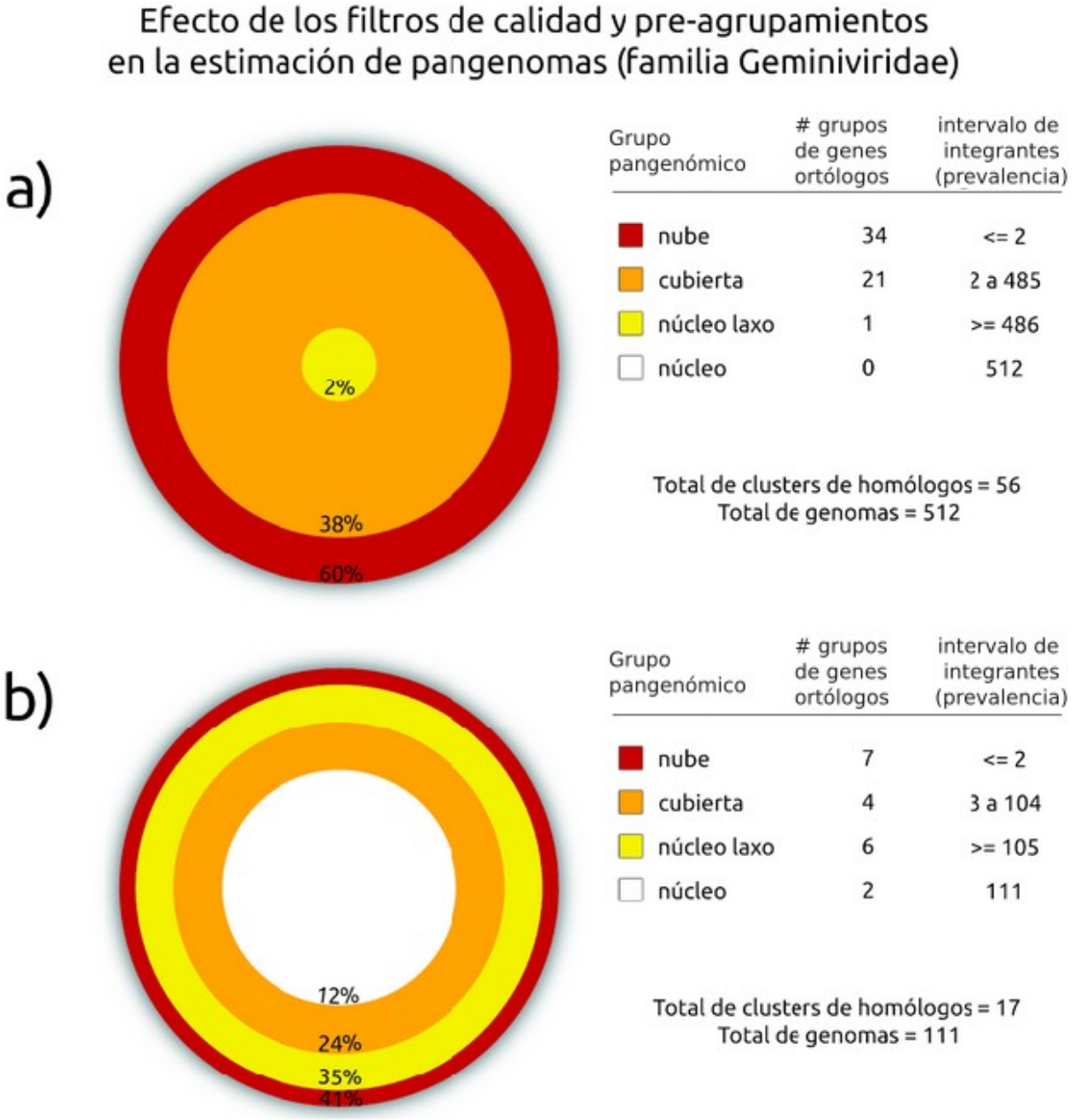


Figura 3R. Efecto de la manipulación previa sobre la categorización por prevalencia. Se ilustra el efecto del concatenado, filtro por conteo y pre-agrupamiento de los genomas, sobre la clasificación por prevalencia de los grupos homólogos. **a)** escenario de análisis directo de la descarga de los genomas con mayor número de genomas y grupos homólogos, sin núcleo pangenómico. **b)** distribución por prevalencia de los grupos de homólogos bajo el impacto de la manipulación previa de los datos. El pangenoma corresponde al subgrupo más inclusivo de la familia *Geminiviridae*.

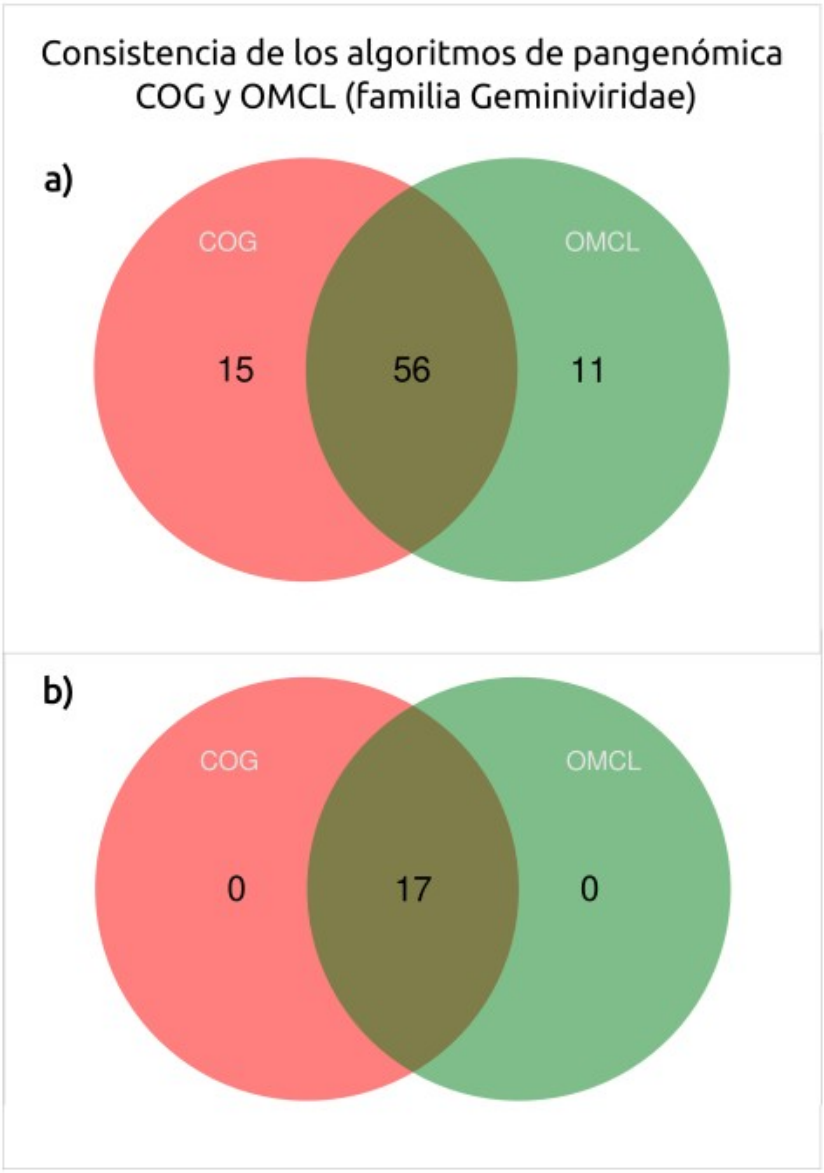


Figura 4R. Efecto de la manipulación previa en la consistencia de los algoritmos pangenómicos. Los diagramas de Venn muestran los grupos de homólogos compartidos y exclusivos de los algoritmos COG (rojo) y OMCL (verde) en dos escenarios: **a)** sin manipulación previa, y **b)** tras concatenado, filtrado y agrupamiento.

Descripción funcional de los pangenomas

Anotación por similitud. Se anotaron ~40% de las proteínas en términos de *Gene Ontology*. 25% de forma exacta, es decir que la proteína era parte del conjunto de referencias de *Uniprot* y 15% mediante *hits* significativos. Los valores límite de *e-value* e identidad para considerar dicha significancia, fueron analizados en el contexto de gráficas de densidad (Figura 5R). Se estableció así un valor máximo de *e-value* = 1e-10 y porcentaje de identidad de al menos 25%, siendo el primero el principal determinante.

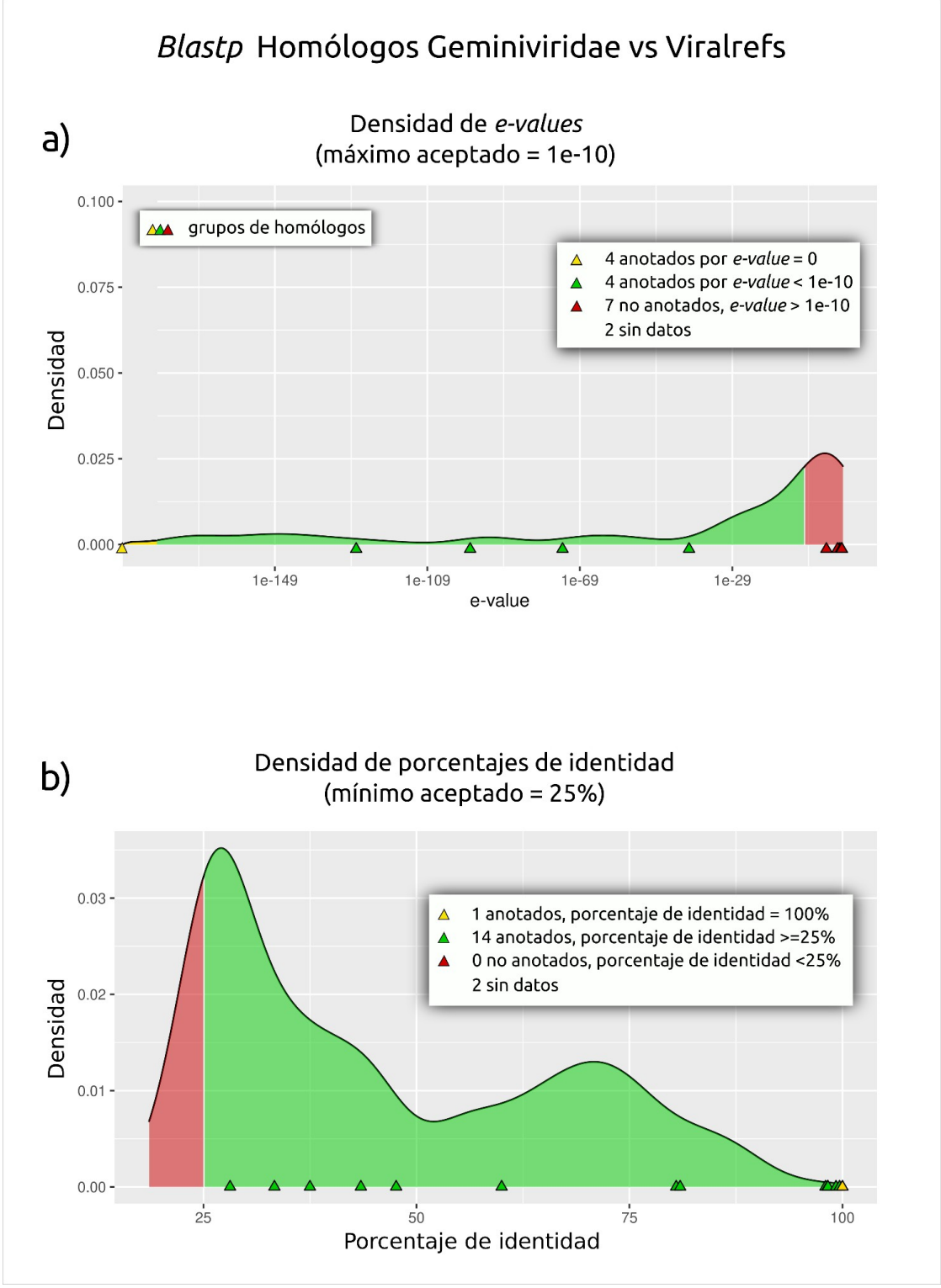


Figura 5R. Valores de aceptación de la anotación por homología. Se muestran las gráficas de densidad de los estadísticos de *Blastp* de la familia *Geminiviridae* contra la base de datos *Viralrefs*. **a)** Densidad de los valores resultantes de *e-value* y **b)** Densidad de los porcentajes de identidad encontrados. En ambos casos se muestran las zonas de aceptación (verde) y rechazo (rojo).

El conteo para todas las familias de los grupos de homólogos anotados exitosamente se muestra en la Figura 6R. En comparación, la estrategia de anotación funcional basada en dominios y el algoritmo GODM devolvió una cantidad ligeramente

menor de secuencias anotadas (Figura 6R). La completitud de la anotación funcional fue del 100% únicamente para la familia *Nanoviridae* (Figura 6R). La diversidad de anotaciones se relaciona positivamente con el tamaño del genoma, relevancia médica del virus y relevancia médica del hospedero.

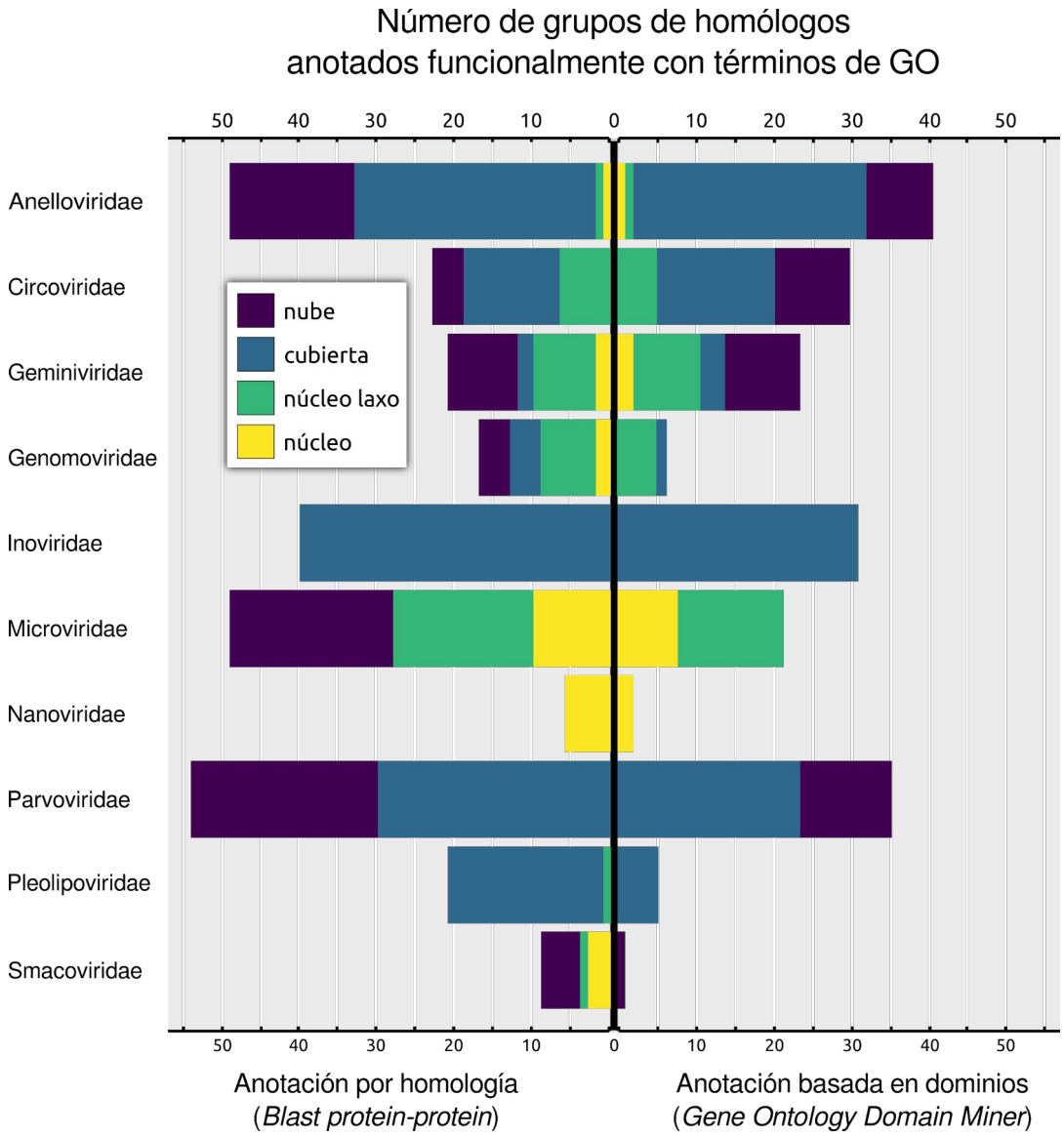


Figura 6R. Número de grupos de homólogos anotados funcionalmente. Cada barra corresponde al conteo de grupos anotados exitosamente mediante las estrategias: basada en similitud (izquierda) y en dominios (derecha). Asimismo, el color indica la proporción correspondiente a cada grupo pangenómico: núcleo (amarillo), núcleo laxo (verde), cubierta (azul) y nube (violeta).

El porcentaje de grupos de ortólogos no anotados por grupo pangenómico (Figura 7R) es asimétrico. Los grupos de alta prevalencia, núcleo y núcleo laxo únicamente aportan el 4% y 9% respectivamente, mientras que los de baja prevalencia, cubierta y nube constituyen el 54% y 33% respectivamente. Dicho de otra forma, los grupos homólogos del núcleo pudieron ser anotados funcionalmente casi en su totalidad (96%).

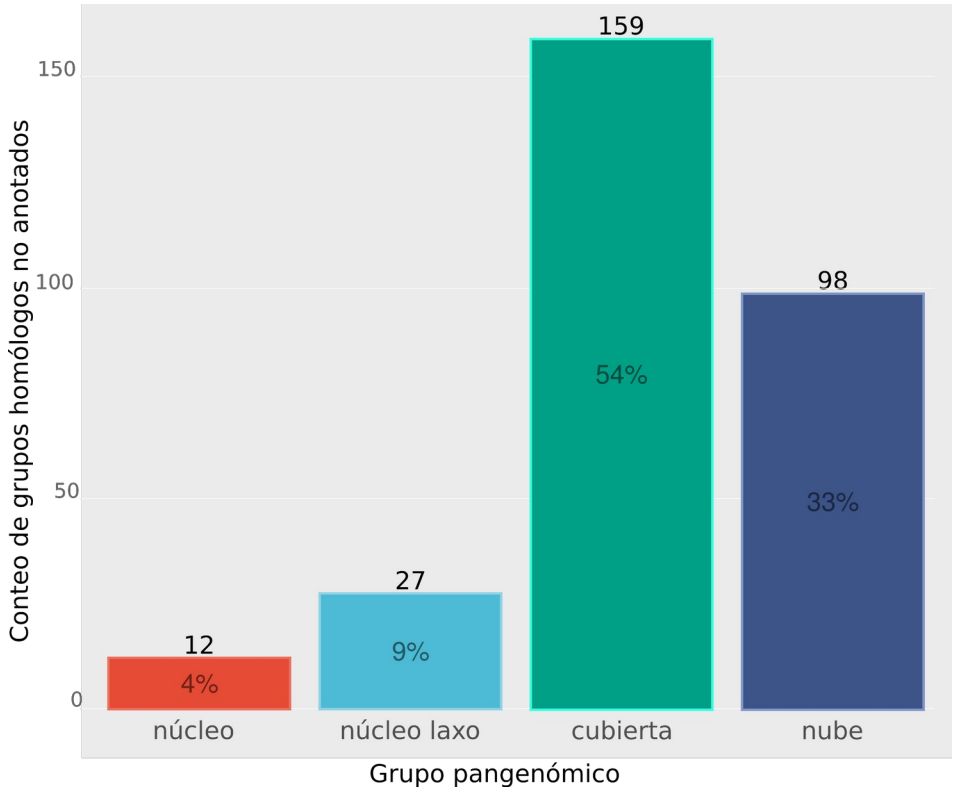
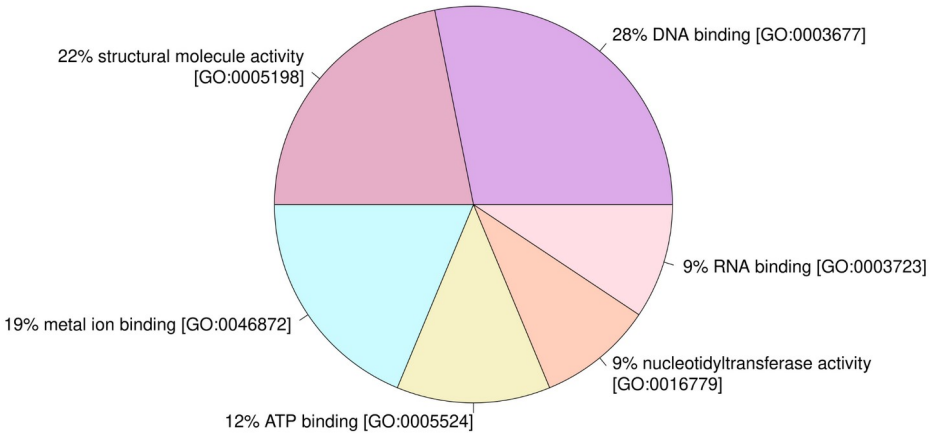


Figura 7R. Número de grupos de genes ortólogos no asociados a una función, por categoría pangenómica. Gráfica de barras que describe el conteo de grupos de secuencias homólogas que no pudieron ser anotadas bajo ninguna estrategia. Se distinguen los grupos pangenómicos por prevalencia, de izquierda a derecha: núcleo, núcleo laxo, cubierta y nube.

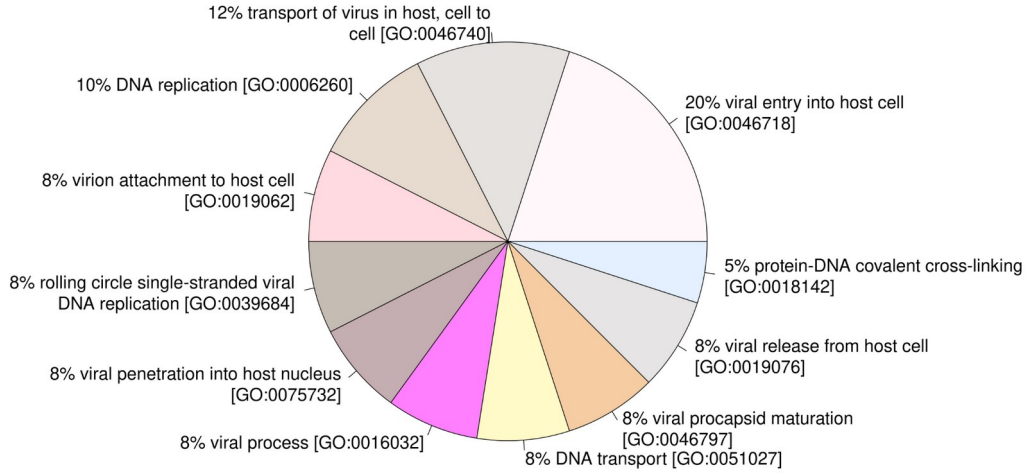
Al universo de conjuntos de homólogos de todas las familias fueron asociadas 81 funciones moleculares, 96 componentes celulares y 174 procesos biológicos, para un total de 351 términos de *Gene Ontology*. Entre las que conforman al núcleo pangenómico, el 65% de las funciones moleculares corresponden a interacción con ácidos nucleicos (DNA, RNA y enlace coordinado a través de iones metálicos), 88% de los procesos biológicos pertenecen a procesos virales característicos (maduración de la procápside, y entrada, empaquetamiento y liberación del material genético viral), y 40% de los componentes celulares se refieren a la cápside viral (Figura 8R).

Anotaciones principales del núcleo pangenómico ssDNA [Gene Ontology]

a) Función molecular



b) Proceso biológico



c) Componente celular

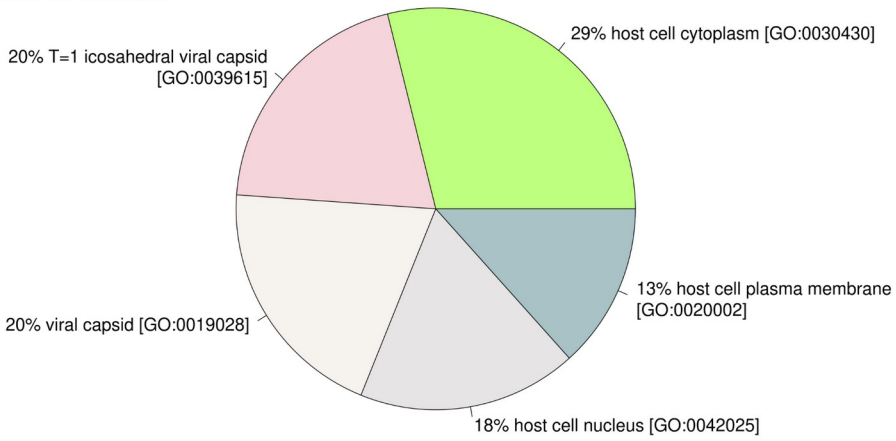


Figura 8R. Anotación funcional del núcleo pangenómico. Se grafican las proporciones de las principales anotaciones funcionales del núcleo pangenómico por cada categoría general de *Gene Ontology*: **a)** *Molecular function*, **b)** Biological process y **c)** *Cellular component*.

Búsqueda de homólogos

Las bases de datos UniRef50 y ssDNA50 conformaron un conjunto de más de 20 millones de secuencias (Tabla 5R) no redundantes de virus (286,202), bacterias (13,308,576), arqueas (960,626), eucariontes (6,748,041), provenientes de más de 88 mil especies.

Tabla 5R. Universo de secuencias para la búsqueda de homólogos virales y celulares. Son indicados los totales de secuencias de proteínas y especies incorporadas a la base de datos global para la búsqueda de homólogos. Se distingue entre las pertenecientes a Archaea, Bacteria, Eukaryota, Virus y No clasificadas.

Grupo	Número de secuencias	Número de especies
Archaea	960626	2511
Bacteria	13308576	42920
Virus	286202	13991
Eukaryota	6748041	28661
No clasificados	372160	172
Total	**21,675,605**	**88255**

Hasta el corte realizado el día 10 de enero de 2022, se cuenta con los resultados de 183 búsquedas de 70 dominios Pfam distintos (Tabla 6R) mediante *jackhmmer* (Eddy, 2011). Esto representa un 69% de avance, y la totalidad para las familias *Bacilladnaviridae*(1), *Circoviridae*(26), *Geminiviridae*(38), *Inoviridae*(71), *Nanoviridae*(6), *Smacoviridae*(2) y *Anelloviridae*(39).

Debido a que cada uno de los 183 resultados puede dar pie a descripciones detalladas de las relaciones entre las secuencias presuntamente homólogas, a partir de esta etapa se optó por enfocar los recursos de procesamiento en explorar los casos con mayores rangos de inclusión taxonómica, es decir aquellos cuyas interpretaciones puedan ser más transversales.

Este es el caso de los dominios "*Geminivirus Rep catalytic domain*" y "*Geminivirus rep protein central domain*", ambos presentes en las familias *Circoviridae* y *Geminiviridae*. Y también el caso de los dominios "*RNA helicase*" y "*Putative viral replication protein*", ambos encontrados en las familias "*Nanoviridae*" y "*Smacoviridae*" (Tabla 6R).

Tabla 6R. Búsquedas de homólogos por dominio. Se disponen los totales por familia (gris oscuro), destacando los casos transversales (verde).

Dominio Pfam	Conteo	*Bacilladnaviridae*	*Circoviridae*	*Geminiviridae*	*Inoviridae*	*Nanoviridae*	*Smacoviridae*	*Anelloviridae*
RNA helicase	7	**1**	**4**	0	0	**1**	**1**	0
Circovirus capsid protein	6	0	**6**	0	0	0	0	0
Putative viral replication protein	6	0	**4**	0	0	**1**	**1**	0
Geminivirus Rep catalytic domain	8	0	**3**	**5**	0	0	0	0
Phospholipase A2-like domain	2	0	**2**	0	0	0	0	0
Viral coat protein (S domain)	2	0	**2**	0	0	0	0	0
Geminivirus coat protein/nuclear export factor BR1 family	7	0	**1**	**6**	0	0	0	0
Geminivirus rep protein central domain	5	0	**1**	**4**	0	0	0	0
Circovirus ORF3	1	0	**1**	0	0	0	0	0
Protein of unknown function (DUF681)	1	0	**1**	0	0	0	0	0
Satellite tobacco necrosis virus coat protein	1	0	**1**	0	0	0	0	0
Geminivirus AC4/5 conserved region	4	0	0	**4**	0	0	0	0
Geminivirus C4 protein	4	0	0	**4**	0	0	0	0
Geminivirus V2 protein	4	0	0	**4**	0	0	0	0
Geminivirus AL2 protein	3	0	0	**3**	0	0	0	0
Geminivirus AL3 protein	2	0	0	**2**	0	0	0	0
WCCH motif	2	0	0	**2**	0	0	0	0
Curtovirus V2 protein	1	0	0	**1**	0	0	0	0
Curtovirus V3 protein	1	0	0	**1**	0	0	0	0
Geminivirus BL1 movement protein	1	0	0	**1**	0	0	0	0
Geminivirus putative movement protein	1	0	0	**1**	0	0	0	0
Protein of unknown function (DUF2523)	7	0	0	0	**7**	0	0	0
Zonular occludens toxin (Zot)	7	0	0	0	**7**	0	0	0
Inovirus Coat protein B	5	0	0	0	**5**	0	0	0
Phage X family	5	0	0	0	**5**	0	0	0
Family of unknown function (DUF5455)	3	0	0	0	**3**	0	0	0
Phage Coat Protein A	3	0	0	0	**3**	0	0	0
AT hook motif	2	0	0	0	**2**	0	0	0
Bacterial type II and III secretion system protein	2	0	0	0	**2**	0	0	0
Helix-turn-helix domain	2	0	0	0	**2**	0	0	0
Phage major coat protein, Gp8	2	0	0	0	**2**	0	0	0
Phage protein	2	0	0	0	**2**	0	0	0
Phage related protein	2	0	0	0	**2**	0	0	0
Phage replication protein CRI	2	0	0	0	**2**	0	0	0
Tail virion protein G7P	2	0	0	0	**2**	0	0	0
Bacterial TSP3 repeat	1	0	0	0	**1**	0	0	0
Bacterial type II/III secretion system short domain	1	0	0	0	**1**	0	0	0
Bacteriophage CI repressor helix-turn-helix domain	1	0	0	0	**1**	0	0	0
Bacteriophage replication gene A protein (GPA)	1	0	0	0	**1**	0	0	0
Cro/C1-type HTH DNA-binding domain	1	0	0	0	**1**	0	0	0
CTX phage RstB protein	1	0	0	0	**1**	0	0	0
Domain of unknown function (DUF3850)	1	0	0	0	**1**	0	0	0
Domain of unknown function (DUF4124)	1	0	0	0	**1**	0	0	0
Domain of unknown function DUF29	1	0	0	0	**1**	0	0	0
Family of unknown function (DUF5447)	1	0	0	0	**1**	0	0	0
Heat-labile enterotoxin alpha chain	1	0	0	0	**1**	0	0	0
Heat-labile enterotoxin beta chain	1	0	0	0	**1**	0	0	0
Helix-destabilising protein	1	0	0	0	**1**	0	0	0
Helix-turn-helix	1	0	0	0	**1**	0	0	0
Inovirus Gp2	1	0	0	0	**1**	0	0	0
N-terminal N1 domain of Vibrio phage CTXphi pIII	1	0	0	0	**1**	0	0	0
Neisseria meningitidis TspB protein	1	0	0	0	**1**	0	0	0
Probable transposase	1	0	0	0	**1**	0	0	0
Protein of unknown function (DUF1293)	1	0	0	0	**1**	0	0	0
Putative Gamma DNA binding protein G5P	1	0	0	0	**1**	0	0	0
Putative transposase DNA-binding domain	1	0	0	0	**1**	0	0	0
Replication initiation factor	1	0	0	0	**1**	0	0	0
Replication protein	1	0	0	0	**1**	0	0	0
Resolvase, N terminal domain	1	0	0	0	**1**	0	0	0
TraX protein	1	0	0	0	**1**	0	0	0
Cell cycle link protein Clink	1	0	0	0	0	**1**	0	0
Movement and RNA silencing protein	1	0	0	0	0	**1**	0	0
Nanovirus coat protein	1	0	0	0	0	**1**	0	0
Nanovirus component 8 (C8) protein	1	0	0	0	0	**1**	0	0
TT viral ORF2	15	0	0	0	0	0	0	**15**
Domain of unknown function (DUF755)	12	0	0	0	0	0	0	**12**
TT viral orf 1	8	0	0	0	0	0	0	**8**
Gyrovirus capsid protein (VP1)	2	0	0	0	0	0	0	**2**
Chicken anaemia virus VP-3 protein	1	0	0	0	0	0	0	**1**
pORF2a truncated protein	1	0	0	0	0	0	0	**1**
Total	**183**	**1**	**26**	**38**	**71**	**6**	**2**	**39**

Al profundizar en las secuencias anotadas con los dominios identificados como casos transversales, se encontró que todos ellos referían a la helicasa Rep. Esta proteína multidominio fue asignada por el *pipeline* de análisis al núcleo o núcleo laxo de las familias *Geminiviridae*, *Genomoviridae*, *Circoviridae*, *Microviridae*, *Nanoviridae*, *Smacoviridae*. Además, la anotación funcional "*rolling circle single-stranded viral DNA replication [GO:0039684]*", proceso en el que participa dicha proteína, destaca entre los de mayor prevalencia de los núcleos pangenómicos (Figura 8R). Por lo anterior y por ser esta

la proteína más documentada entre los virus ssDNA, se presentan los análisis evolutivos de sus búsquedas de homólogos. En particular se escogió la búsqueda proveniente de la familia *Circoviridae*, nombrada "*Circoviridae_OMCL_3_putative_replication..*".(ver Tabla 6R celda verde brillante).

Homólogos del grupo "*Circoviridae_OMCL_3_putative_replication..*"

Los homólogos recuperados en la búsqueda mediante *jackhmmer* para los medioides del grupo *"Circoviridae_OMCL_3_putative_replication.."* abarcaron secuencias de nueve de las trece familias de virus ssDNA, *Circoviridae*(108), *Geminiviridae*(146), *Genomoviridae*(174), *Nanoviridae*(18), *Inoviridae*(2), *Parvoviridae*(2), *Smacoviridae*(75), *Alphasatellitidae*(77) y *Microviridae*(20), además de secuencias de Bacteria(98), Eukaryota(2) y *Mimiviridae*(1).

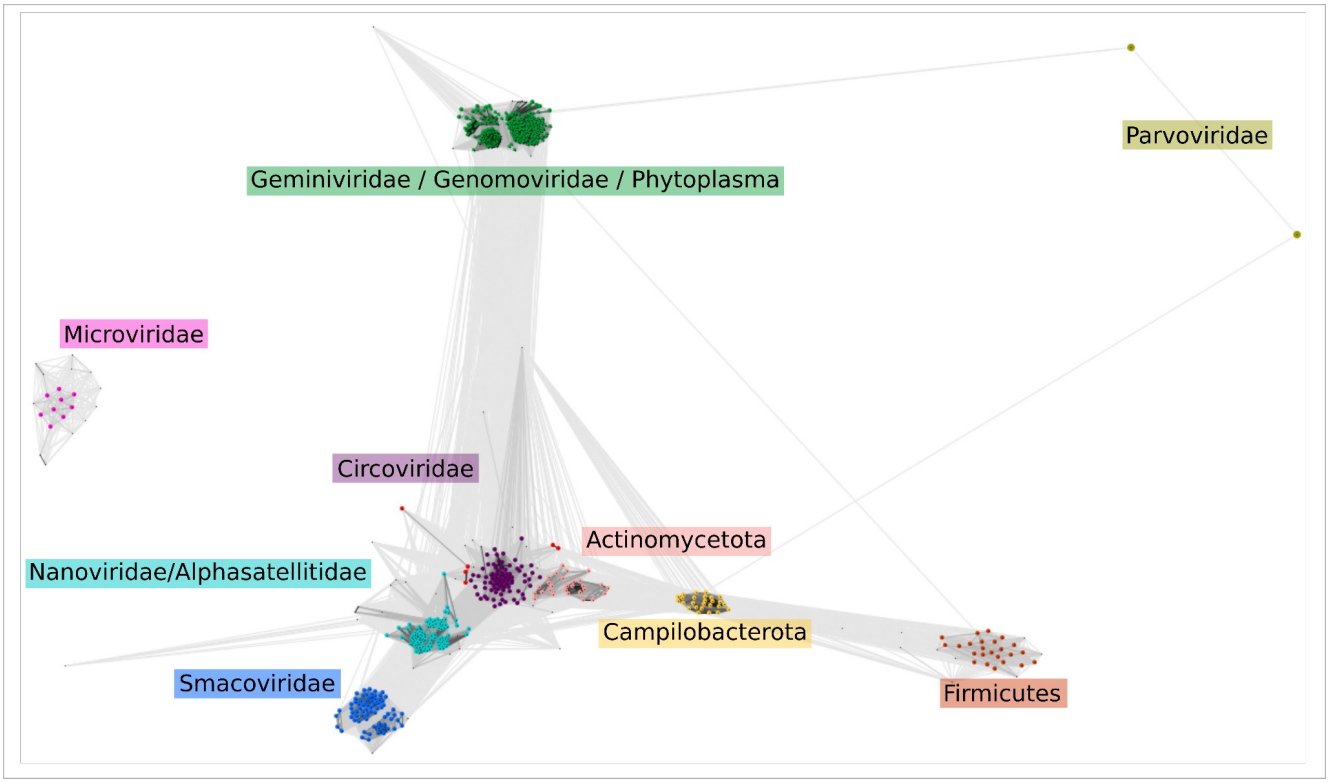


Figura 9R. Red de similitud CLANS de los homólogos Rep. Red construida mediante el software CLANS con las 723 secuencias de proteínas recuperadas en la búsqueda de homólogos de "*Circoviridae_OMCL_3_putative_replication..*". Cada punto es una secuencia, el color de los puntos señala pertenencia al grupo de *convex clustering* indicados con el color de etiqueta correspondiente. Los ejes entre pares de puntos muestran un valor de P>0.8 para el establecimiento de dicha relación.

No obstante, el alineamiento múltiple de dichas secuencias, aún con parámetros de alta precisión, resultó en alineamientos deficientes de "no convergencia", que a su vez generaron filogenias por máxima verosimilitud que no satisficieron las pruebas de saturación; es decir, era imposible distinguir las sustituciones puntuales de aquellas generadas por múltiples eventos.

La red de similitud construida con el programa CLANS (Figura 9R) permitió identificar 9 grupos mediante la conectividad de elementos coalescentes. Al etiquetar dichos elementos con las clasificaciones de sus integrantes, se hizo evidente que estos reflejan algunas relaciones entre familias virales y phyla bacterianos. Existen dos observaciones destacables. La primera es la existencia de un *cluster* que agrupa proteínas de procedencia viral y bacteriana (*Geminiviridae*/*Genomoviridae*/Phytoplasma). Y la segunda, es la posición de la familia *Microviridae*, pues se encuentra desconectada del resto (Figura 9R).

El componente mayor de la red fue separado y con él se construyó finalmente un alineamiento adecuado, así como filogenias por máxima verosimilitud que mostraron valores de soporte en su mayoría por encima de 80% y un menor grado de saturación.

El árbol consenso de homólogos Rep (Figura 10R) muestra una distribución de los grupos virales en 4 grandes clados:

1) Los *Geminiviridae*, *Genomoviridae*, *Parvoviridae* y una especie de *Inovirus* forman un grupo junto a bacterias del género *Phytoplasma* (Figura 10R a). Destaca la agrupación del género *Masterivirus* de la familia *Geminiviridae* con bacterias del género *Phytoplasma* al ser ambos grupos de parásitos obligados de plantas.

2) La familia *Circoviridae* aparece segmentada en dos nodos, uno de ellos formando un grupo con bacterias del phylum Actinomycetota. El segundo nodo de la familia *Circoviridae* se agrupa con bacterias Firmicutes (relación no resuelta al colapsar los nodos con bootstrap < 80).

3) Un grupo exclusivamente viral que contempla a los *Nanoviridae*/*Alphasatellitidae*, el género *Hyperionvirus* de virus gigantes de la familia *Mimiviridae* y todos los *Smacoviridae*.

4) Bacterias del phylum Campilobacterota formando un grupo con un *Inovirus* y estos, a su vez, relacionados a las bacterias del phylum Firmicutes.

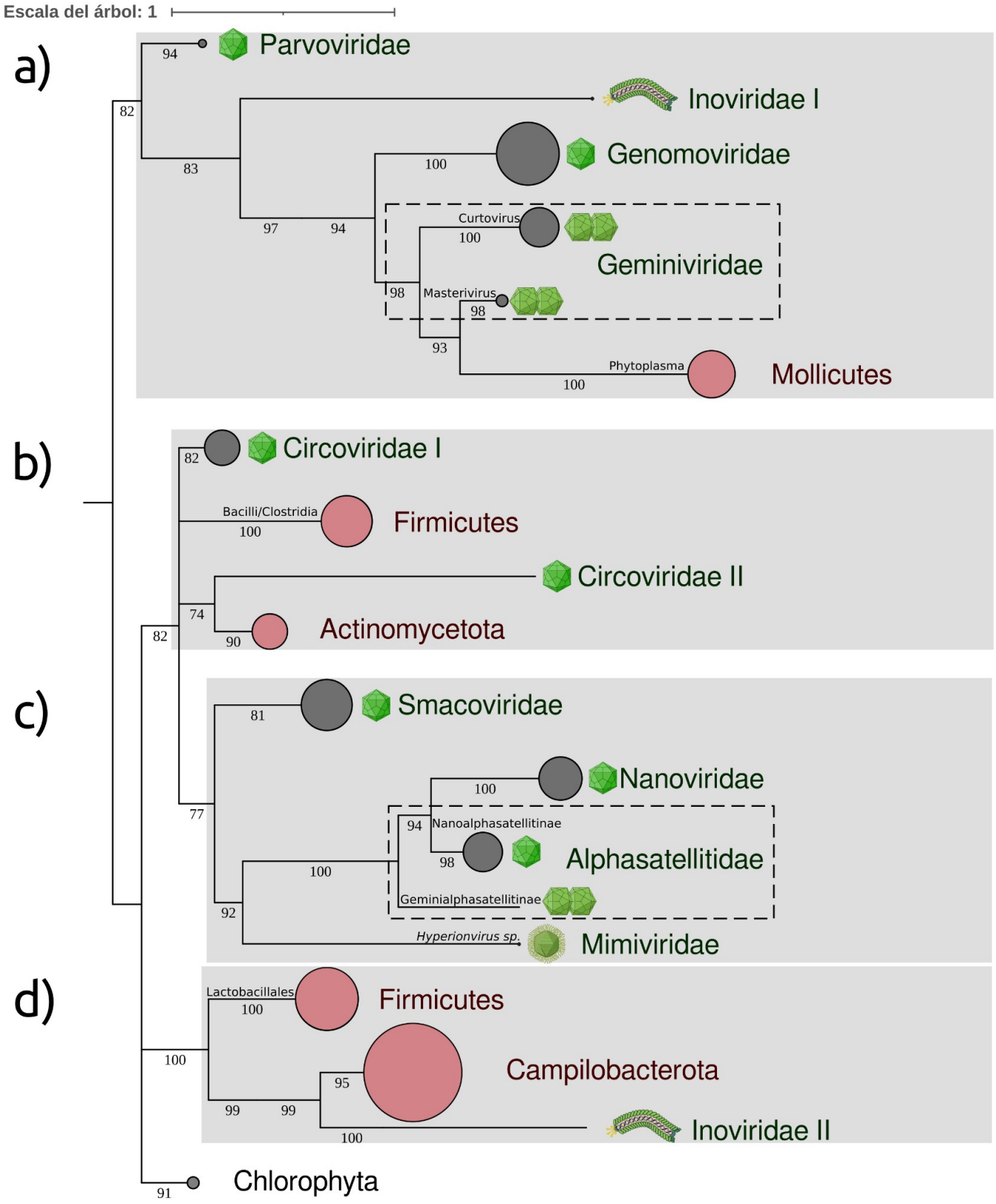


Figura 10R. Árbol consenso de máxima verosimilitud de los homólogos Rep. Grupos principales formados entre secuencias de origen viral y origen bacteriano (a-d). Los círculos en ramas terminales representan grupos colapsados de secuencia de organismos del mismo grupo. El tamaño del círculo es proporcional al número de secuencias del nodo colapsado. Las ramas correspondientes a linajes virales se indican con un esquema del tipo de cápside del grupo. Los valores de bootstrap escalados de 0 a 100 se muestran debajo de cada rama interna.

DISCUSIÓN

Características y distribución de los datos genómicos

El conocimiento de los virus se encuentra en constante cambio. Familias completas han sido descubiertas recientemente, gracias a la implementación de nuevas técnicas de secuenciación, así como *pipelines* metagenómicos enfocados en la reconstrucción de genomas virales, o muestreos en ambientes poco conocidos. Estos, y otros factores como la importancia médica y/o económica del virus y su hospedero, influenciaron directamente la disponibilidad de los datos descargados y su estructuración.

Entre los virus de DNA de cadena sencilla, dos de las familias más representadas fueron *Geminiviridae* y *Tolecusatellitidae*, con 538 y 136 genomas completos respectivamente (Tabla 1R). Esta alta disponibilidad muy probablemente se deba a que entre sus miembros existen patógenos del maíz y tomate, además de otros cultivos de alto consumo, que provocan enfermedades y pérdidas millonarias cada año por su alta transmisibilidad (Morales y Anderson, 2001; Zúñiga-Vega y Ramírez, 2002). Por su parte, el resto de los grupos de alta representatividad (>100 genomas completos), son *Circoviridae*, *Parvoviridae*, *Genomoviridae* y *Anelloviridae* (Figura 1R a), todas ellas con miembros patógenos de humanos,;como los virus *Torque teno virus* (*Anelloviridae*) (Hino y Miyata, 2007) y *Primate erythroparvovirus 1* (*Parvoviridae*) (Landry, 2016) o asociados a humanos, pero sin afecciones evidentes (*Genomoviridae*) (Krupovic *et al.*, 2016). Por otra parte, existe baja disponibilidad de datos para virus cuyos hospederos pertenecen a linajes diversos, pero menos estudiados (Figura 1R b) como las arqueas (*Pleolipoviridae* y *Spiraviridae*) o bacterias de vida libre (*Inoviridae*), familias virales que atraviesan frecuentemente cambios en su clasificación.

Durante el curso de esta investigación, algunos miembros de la familia *Inoviridae* fueron reasignados a la familia *Plectroviridae* (Adriaenssens *et al.*, 2020), así como también se creó la familia *Finnlakeviridae*, cuyo único miembro (*Finnlakevirus FLiP)*, es el primer virus ssDNA conocido con cápside icosaédrica y membrana interna. Otro grupo reciente es la familia *Spiraviridae*, que infecta arqueas hipertermófilas y que muestra características antes desconocidas (Mochizuki *et al.*, 2012). El estado del conocimiento de las familias virales incorporadas al presente estudio es heterogéneo, especialmente cambiante y creciente para virus cuyos hospederos habitan los límites fisiológicos de la vida. Esto tiene un peso sobre los métodos y estrategias para investigar su origen y evolución temprana, pues los nuevos grupos pueden modificar los pangenomas encontrados y la prevalencia de los grupos de homólogos. Métodos como la metagenómica no dirigida, sobre todo en ambientes sub-muestreados, abordan la diversidad conocida de los virus ssDNA (Simmonds *et al.*, 2017; Wang *et al.*, 2018). Por ello, los métodos implementados en estudios evolutivos deben facilitar la incorporación de secuencias de dicha proveniencia. Esto fue parcialmente logrado para la familia *Smacoviridae*, representada en la base de datos final con genomas de ensambles metagenómicos de muestras fecales (Varsani y Krupovic, 2018). Podemos considerar que el tamaño de la base de datos construida es

representativa y suficiente para la realización de análisis pangenómicos en relación con el estado actual de conocimiento.

Tratamiento de los datos descargados

El filtrado de los genomas afectó de forma distinta la cantidad de secuencias para cada familia. En general, supuso cambios significativos (>24% tamaño muestral) para las familias *Bidnaviridae*, *Geminiviridae* y *Nanoviridae* (Tabla 1R). Esto corresponde a las particularidades genómicas de dichos grupos, pues empaquetan su material genético de forma segmentada en distintas partículas virales, hasta 8 de ellas en la familia *Nanoviridae*, que coincidentemente dispuso el mayor impacto, ya que 87% de los genomas fueron concatenados. Esta modularidad tuvo un impacto significativo en la consideración del supuesto del análisis pangenómico que asume que todos los elementos a comparar representan la totalidad del repertorio genético de una especie, pues cada genoma segmentado constituye sólo una fracción del total. Esto es similar, aunque en mucho mayor proporción a lo ocurrido con los plásmidos bacterianos (Contreras-Moreira y Vinuesa, 2013). La principal limitante que esto significa para las familias *Bidnaviridae, Geminiviridae* y *Nanoviridae* consiste en que al conformar el *input* por segmento genómico, los algoritmos pangenómicos *COG* y *OMCL* interpretan que los genes de un segmento no están relacionados con el otro, considerándolos como dos conjuntos independientes que concluyen en conteos erróneos. Como resultado los núcleos pangenómicos son inexistentes, pues no existen segmentos que porten al menos un representante de todos los grupos ortólogos. Desde el punto de vista metodológico, este déficit fue totalmente resuelto con la concatenación de segmentos. Sin embargo, existen implicaciones teóricas sobre cómo dicha manipulación modifica el concepto de pangenoma viral, pues en estos casos es insuficiente para retratar la modularidad, o incluso de forma extensiva, son también insuficientes para virus satélites (*Alphasatellitidae* y *Tolecusatellitidae*) al no tomar en cuenta los genes del virus auxiliar.

Los valores límite para el filtro por conteo tuvieron que ser ajustados en el contexto de cada familia, ponderando el mantener una muestra viable para los análisis pangenómicos. En el caso más notable, de la familia Baciladnaviridae, esto significó elevar el valor permitido de variación hasta el 95% (Tabla 2R), valor mínimo para mantener una muestra de más de 5 genomas, procesable por los algoritmos COGtriangles y OrthoMCL. Para el resto de las familias el impacto fue simplemente la reducción de la muestra, orientada a la homogeneización.

Agrupamientos pre-pangenómicos

Los esquemas de agrupamiento al interior de cada familia reflejaron principalmente a las subfamilias y géneros. El patrón observado fue que los subgrupos eran más distantes y consistentes cuando el rango de hospederos de la familia era más restringido. Como fue mencionado anteriormente, el método de *Cumulative Power Spectrum* es sensible a diferencias globales en los genomas, como pueden ser variación en la sintenia, genes

presentes y si el genoma está o no segmentado. Por ello, las muestras de familias como *Geminiviridae* (Figura 2R a) y *Nanoviridae* resultaron en agrupamientos contundentes, entre sus miembros segmentados y monopartitos. La familia de virus satélites *Alphasatellitidae* fue segmentada en las subfamilias *Geminialphasatellitinae* y *Nanoalphasatellitinae*, que reflejan al virus del cual son satélites, dando importancia a estas interacciones.

Los escenarios más complejos como los 7 subgrupos de *Inoviridae* (Figura 2R b) no son fácilmente relacionados con su clasificación. Este tipo de agrupamientos mixtos (Figura 2R b) o solapados (Figura 2R c) con un gran número de subgrupos, muy heterogéneos en tamaño, fue observado en familias cuya clasificación atraviesa cambios más frecuentes, como las familias *Inoviridae* (Adriaenssens *et al.*, 2020), *Pleolipoviridae* (Bamford et al., 2017) y *Parvoviridae* (Cotmore et al., 2019) . Esto denota las problemáticas de su clasificación que han llevado el planteamiento de nuevas familias. Los pangenomas encontrados para estas familias son más propensos a cambiar conforme se amplíe el conocimiento de su diversidad. Además, estos son grupos para los cuales no pudieron ser identificados núcleos pangenómicos (Figura 6R).

La comparación de la consistencia obtenida bajo los dos distintos algoritmos de análisis permitió identificar el conjunto de parámetros que favorece la robustez de los grupos de ortólogos recuperados.

Distribución de las funciones de los grupos homólogos

Si bien se incorporaron especies virales ampliamente investigadas, cuyas anotaciones funcionales son conocidas, anotar mediante las cepas de referencia puede traer consigo un cuello de botella a la diversidad de funciones; por ejemplo, favoreciendo la presencia de funciones asociadas a la patogenicidad, por ser estos los elementos genéticos más estudiados. Sin embargo, este enmascaramiento de la diversidad funcional puede ser heterogéneo. Se sabe que las funciones relacionadas a procesos de flujo de la información, empaquetamiento e interacción con la célula huésped, muestran poca variación entre virus de la misma familia (Kristensen *et al.*, 2013). Por lo anterior, es probable que las anotaciones funcionales del núcleo pangenómico tengan una certeza suficiente y mayor al resto, lo cual es muy favorable debido a su importancia para la búsqueda de homólogos celulares y virales no ssDNA.

Como puede observarse en la Figura 6R, la cubierta y la nube contienen la mayor diversidad de anotaciones funcionales entre las categorías pangenómicas. De forma contrastante, ambas categorías también son aquellas con más grupos de homólogos no anotados (Figura 7R). Al igual que otros resultados, estas aseveraciones deben precisarse a nivel de familia. Por ejemplo, la familia *Nanoviridae* se constituye únicamente por núcleo pangenómico. Así mismo, para las familias *Microviridae* y *Smacoviridae*, las anotaciones funcionales del núcleo y núcleo laxo exceden las de cubierta y nube. Consideramos que muy probablemente esto se debe a que la muestra de genomas para dichos grupos es reducida y poco diversa.

Funciones principales del núcleo pangenómico

En consonancia con lo esperado, la mayoría de las funciones moleculares del núcleo (65%) están relacionadas con la interacción con ácidos nucleicos y empaquetamiento del genoma (Figura 8R a). Las funciones moleculares, *DNA binding* [GO:0003677], *RNA binding* [GO:0003723], *nucleotydiltransferase activity* [GO:0016779] y, *metal ion binding* [GO:0046872] están asociadas, principalmente, con las proteínas "*Replication-associated protein*", "*Replication enhancer protein*". Ambas desempeñan un papel central en la replicación *rolling circle replication*, también parte de los principales procesos biológicos de Gene Ontology (GO) (Figura 8R b) y concentran información evolutiva, perspectiva con la que ya se ha investigado el o los posibles orígenes de los virus CRESS DNA desde genes celulares (Liu *et al.*, 2011; Kuprovic y Koonin, 2014; Kazlauskas *et al.*, 2019).

Por su parte, las funciones moleculares implicadas en el empaquetamiento del genoma están representadas por múltiples proteínas de cápside y de movimiento, comúnmente referidas como "*Capsid protein*" y "*Movement protein*", las cuales también han sido estudiadas en perspectiva del origen y evolución temprana, pero sólo en el caso de la familia *Bacilladnaviridae* (Kazlauskas *et al.*, 2017).

En general, y como era esperado, el núcleo pangenómico concentra funciones moleculares relacionadas a los procesos de flujo de la información genética, procesos biológicos específicos a los virus, incluso a los virus ssDNA, pero que a su vez están ampliamente compartidos y conservados entre las familias.

La categoría de GO cuyos términos son menos aplicables a las familias virales, son los referidos como "Componente celular", pues únicamente funge como referencia para la localización física de las proteínas virales en las partes celulares del hospedero (Figura 8R c). Para el núcleo pangenómico, las anotaciones fácilmente relacionales son "*T=1 icosahedral viral capsid* [GO:0039615]" y "*viral capsid* [GO:0019028]", ambas refiriéndose a las mismas proteínas de cápside ya mencionadas, pero sin ser lo suficientemente específicas.

Homólogos celulares y virales

La base de datos utilizada para la búsqueda de homólogos incorporó un total de 21,675,605 secuencias (Tabla 5R). Este tamaño muestral y representatividad es similar al utilizado por Kazlauskas y colaboradores en 2019, para identificar homólogos a los dominios HUH endonucleasa de virus ssDNA (Kazlauskas *et al.*, 2017), y representa una ampliación a los antecedentes del Laboratorio de Origen de la Vida (Campillo-Balderas, 2018). Debido a ello consideramos suficiente el tamaño de muestra utilizado. Sin embargo, para mantener este volumen de datos, fue necesario indexar la base de datos de forma

local. Como consecuencia de la pandemia de COVID-19, todas las búsquedas de homólogos fueron realizadas en una computadora personal, lo que significó una limitación importante debido a la alta demanda de recursos de procesamiento.

Para evitar cambiar los parámetros de *jackhmmer* o utilizar una base de datos menos representativa, se optó por realizar un corte con las búsquedas completadas hasta el 10 de enero de 2022 y profundizar en casos específicos (Tabla 6R). Como puede observarse, sólo algunas familias han sido completadas. A pesar de ello, los casos denominados "transversales" (Tabla 6R verde), por abarcar más de una familia, permitieron construir búsquedas con un panorama más amplio y rangos de inclusión taxonómica mayor.

Como era esperado, las búsquedas que cumplían dichos criterios partieron de los dominios de las proteínas virales Rep y Rep-A, al ser las de mayor prevalencia entre los virus ssDNA (Zhao *et al.*, 2019), y más comunes en los núcleos pangenómicos. Aunque esto parezca redundante respecto a trabajos ya publicados (véase Liu *et al.*, 2011, Kazlauskas *et al.*, 2017; 2019) que postulan relaciones entre secuencias virales y celulares de las Rep, debe tomarse en consideración que las semillas de búsqueda de homólogos en los trabajos mencionados fueron definidas "manualmente", a partir del conocimiento *a priori* de estudios estructurales (Chapman y Rossmann, 1993), fisiológicos (Krupovic y Forterre, 2015) y comparaciones de secuencias (Ilyna y Koonin, 1992; Kuprovic y Koonin, 2014). En nuestro caso, la definición de los conjuntos de secuencias se realizó mediante un proceso de filtrado, pre-agrupamiento, pangenómica e identificación de dominios, automatizado. Al llevarse de esta forma, la comparación con las filogenias previamente publicadas es la mejor vía de validación de nuestros procesos de análisis.

Estudio de caso: Homólogos de las helicasas Rep de virus ssDNA

Por su transversalidad, fueron elegidos los resultados de búsqueda de homólogos correspondientes a los dominios "*Geminivirus rep protein central domain*" y *"Geminivirus Rep catalytic domain"*, particularmente aquellos del *cluster* pangenómico etiquetado como "*Circoviridae_OMCL_3_putative_replication*". Acorde a lo esperado, el uso de *jackhmmer* de dichos representantes contra "UniRef50 + ssDNA50", devolvió un conjunto de proteínas de diferentes grupos, encontrándose resultados positivos entre todas las familias CRESS DNA y algunos elementos genéticos móviles de bacterias e incluso algas.

El conjunto de presuntos homólogos mostró una gran similitud con las secuencias del estudio de Kazlauskas y colaboradores (2019), así como Zhao y colaboradores (2021). Entre los grupos incluidos las diferencias principales fueron: (i)las familias *Rudiviridae*, *Sphaerolipoviridae*, *Myoviridae*, *Corticoviridae*, ausentes de nuestros resultados. Todas ellas son familias de virus de DNA de doble cadena que infectan bacterias, con replicación vía *rolling circle replication (RCR).* Su omisión probablemente se deba a que las proteínas Rep de estos grupos forman parte de la familia de proteínas Rep_trans, cuyas similitudes

son notorias solo a nivel de la estructura de sus sitios catalíticos (Wawrzyniak *et al.*, 2017). **(ii)**Destaca la ausencia de secuencias provenientes de helitrones del grupo *IS91*, elementos móviles con transposasas que se asemejan a las proteínas iniciadoras de RCR (Thomas y Pritham, 2015), cuya ausencia puede deberse a que no conservan los mismos dominios que las virales. **(iii)**No fueron obtenidas secuencias endógenas de genomas eucariontes. Esto va en contra de lo esperado, puesto que han sido descritas secuencias endógenas de invertebrados de origen trazable a la familia *Circoviridae* (Zhao *et al.*, 2021). Esta discordancia pudo originarse en el estado cambiante de la clasificación de la familia (Rosario *et al.*, 2017).

En su mayoría, las secuencias recuperadas corresponden a endonucleasas del tipo HUH, con representantes de las cuatro categorías identificadas entre virus (Wawrzyniak *et al.*, 2017). Estas son las familias de proteínas **(i)***Phage_GAP;* de la familia *Microviridae*, **(ii)***Viral_Rep*; que incluye a diversos virus ssDNA circulares, **(iii)***Gemini_AL1*; característica de *Geminiviridae* y *Genomoviridae*, y **(iv)** *Rep_N*; encontrada en virus ssDNA de genoma lineal. El resto de las secuencias obtenidas que pudieron ser asignadas a una familia de proteínas se distribuye entre los grupos *Rep_1* y *Rep_2*, ambos de procedencia bacteriana. Otros subgrupos de endonucleasas no fueron incorporados, un escenario previsible dado que son considerados como evolutivamente independientes (Francia y Clewell, 2002).

La red de CLANS (Figura 9R) permitió establecer un escenario general del grado de similitud entre secuencias. Para su interpretación, debe tomarse en cuenta que ésta proviene del modelado en un espacio tridimensional y que no deben tomarse las distancias lineales aparentes entre puntos. El principal aporte de dicha red fueron los grupos generados mediante *convex clustering* con base en su conectividad, destacando la congruencia que estos muestran con los phyla bacterianos y las familias virales, con excepción de los *Inoviridae*. Aún más interesante es el agrupamiento entre secuencias de virus de las familias *Geminiviridae* y bacterias del género *Phytoplasma*, ambos parásitos intracelulares obligados de plantas. Una observación ya conocida, por el alto grado de similitud conservado a nivel de secuencia entre ambos grupos (Krupovic *et al.*, 2009), que ha generado debate sobre cómo debe interpretarse (Ver abajo).

Es importante hacer notar la condición de las proteínas Rep de *Microviridae* (Figura 9R izquierda), único subgrupo desconectado del resto y cuya inclusión en los árboles de máxima verosimilitud causaba el incumplimiento de las pruebas de saturación. No obstante, la información dispuesta en la red es insuficiente para dar por hecho que dichas proteínas no estén relacionadas evolutivamente al resto. Una posible vía para explicar su desconexión puede ser la ausencia ya mencionada de otros bacteriófagos con las que muestran mayor similitud (Kazlauskas *et al.*, 2019), que pudieran constituir puntos intermedios que los anclen a la red global.

El árbol consenso de homólogos Rep (Figura 10R) muestra algunos de los grupos reportados anteriormente. En primer lugar, las familias *Geminiviridae* y *Genomoviridae* forman un grupo junto con las bacterias del género *Phytoplasma* (Figura 10R a), lo cual suma en favor de un origen común (Zhao *et al.*, 2019). La polaridad de dicha relación aún no

ha sido resuelta, quedando abiertos dos escenarios: **(i)** las bacterias *Phytoplasma* y otras Mollicutes adquirieron sus proteínas Rep desde algún geminivirus mediante transferencia genética horizontal ocurrida en hospederos comunes. Y las Rep de *Genomoviridae* y *Geminiviridae* surgieron desde otras helicasas virales (Saccardo *et al.*, 2011), ancestro que,para nuestros resultados, sería común entre *Parvoviridae*, *Inoviridae*, *Genomoviridae* y *Geminiviridae*. O **(ii)** las helicasas virales de *Geminiviridae* surgieron por escapes genéticos de bacterias (Ilyna y Koonin, 1992), particularmente desde elementos extracromosómicos de plásmidos de *Phytoplasma* de origen bacteriano (Krupovic *et al.*, 2009). La información que nosotros obtuvimos (Figura 10R a) se acerca a la propuesta de Saccardo y colaboradores (2011); puesto que no pudieron ser establecidas relaciones desde Reps de *Phytoplasma* y el resto de homólogos bacterianos.

Las Rep de *Circoviridae* aparecen en dos grupos distintos del árbol, relacionado a dos grupos distintos de bacterias, lo cual es incompatible con un escape genético para explicar su origen, una propuesta que se debe robustecer con un mayor nivel de soporte de rama. A pesar de ello más del 90% de las secuencias se agrupan en un sólo nodo, por la condición monofilética del género *Cyclovirus*, que abarca la mayor diversidad, mientras que el resto de integrantes de la familia tienen una clasificación debatida dentro de la familia (Rosario *et al.*, 2017). Su cercanía con bacterias Firmicutes y Actinomycetota muestra una topología distinta a la descrita por Kazlauskas y colaboradores (2019), en la que estas helicasas parecen compartir un ancestro común con *Smacoviridae*, *Nanoviridae* y *Alphasatellitidae*. La omisión de otras familias de bacteriófagos de dsDNA puede ser la causa de la pérdida de dicho agrupamiento, pues algunos de sus integrantes muestran semejanzas tanto con *Smacoviridae* como con *Circoviridae* (Zhao *et al.*, 2019), que pudiera actuar como puntos intermedios entre ambos linajes.

El grupo conformado por *Smacoviridae*, *Nanoviridae* y *Alphasatellitidae* también ha sido reportado previamente (Varsani y Krupovic, 2018). Destaca la inclusión de un integrante de la familia *Mimiviridae*, la cual forma parte del orden con los genomas virales de mayor tamaño (Nucleocytoviricota), que han mostrado un papel central de la transferencia horizontal en el origen de muchos de sus genes (Moreira y Brochier-Armanet, 2008). La ausencia de helicasas Rep en otros mimivirus aunado a su alta similitud con las representantes de *Nanoviridae*/*Aplhasatellitidae*, sugieren un evento de transferencia genética horizontal reciente.

Como puede observarse, los *Inoviridae* se distribuyen en dos grupos distantes entre sí, uno asociado a *Parvoviridae*, *Geminiviridae*, *Genomoviridae* y Phytoplasma, y el otro a helicasas bacterianas. Este posicionamiento es compatible con que *Inoviridae* sea una familia polifilética (Roux *et al.*, 2019), lo que puede ser observado también en la red de similitud (Figura 9R).

Las helicasas Rep de virus ssDNA mostradas en el árbol de máxima verosimilitud parecen tener historias evolutivas diversas. El caso particular que más ha sido estudiado es el de *Geminiviridae*/*Phytoplasma*, para el que nuestro árbol favorece la condición plesiomórfica de las helicasas de *Geminiviridae*/*Genomoviridae* y una adquisición

posterior vía transferencia horizontal hacia bacterias Phytoplasma (Saccardo *et al.*, 2011). Este escenario reivindica el papel de los virus en la evolución de otros linajes celulares, pero a diferencia del virocentrismo, en este caso para explicar el origen reciente de un grupo particular de elementos genéticos.

El grupo (a) suma en favor de un ancestro viral común a las helicasas Rep de *Parvoviridae*, *Geminiviridae* y *Genomoviridae*.

Para el conjunto formado entre los subgrupos (b) y (c) del árbol mostrado, puede aceptarse la hipótesis de escape,en este caso de helicasas bacterianas relacionadas a las Firmicutes y Actinomycetota, para dar origen a las de la familia *Circoviridae*. No obstante, la posibilidad de extender dicha ancestría bacteriana a los grupos *Smacoviridae*, *Nanoviridae* y *Alphasatellitidae* dependerá de la posibilidad de resolver los nodos colapsados, posiblemente incorporando análisis estructurales posteriores.

Otro evento atribuible a la hipótesis de escape permitiría explicar el surgimiento de los *Inoviridae* II (Figura 10 d) como fenómenos de escape desde el phylum Campilobacterota, aceptando la condición polifilética de la familia (Roux *et al.*, 2019).

Conclusiones

- Las particularidades genómicas de los virus ssDNA pudieron ser sorteadas para llevar una muestra de sus genomas de referencia a un análisis pangenómico, identificando los grupos de proteínas homólogas de 13 de las 15 familias.

- Los procesos de edición de los archivos genómicos crudos, pre-agrupamiento por similitud y concatenado se automatizaron en un *pipeline* bioinformático que puede ser llevado a otros grupos virales con características similares.

- La categorización de los grupos de homólogos, por su prevalencia, permitió construir un núcleo pangenómico en 6 de las 15 familias, identificando las proteínas características y compartidas.

- Anotar funcionalmente utilizando dos estrategias independientes hizo posible la descripción funcional del núcleo pangenómico, encontrando una mayor presencia de proteínas involucradas en la replicación y empaquetamiento del genoma.

- El estudio evolutivo de los elementos de alta prevalencia agrupados en el núcleo pangenómico permitió describir relaciones transversales entre las familias de virus ssDNA y organismos celulares. Esto resalta la importancia de continuar una estrategia de estudio evolutivo sistematizada a partir de catálogos pangenómicos.

- Se identificaron los dominios de las proteínas de mayor prevalencia, y la selección de sus representantes permitió extender una búsqueda amplia de sus homólogos en linajes remotos.

- La filogenia de homólogos de helicasas Rep / endonucleasas HUH virales y celulares plantea que la diversidad actual es producto de múltiples historias evolutivas. Dos de ellas se inclinan en favor de la hipótesis de escape para explicar el origen de los *Inoviridae* y *Circoviridae*.

- La posición en la filogenia de los *Geminiviridae*/*Genomoviridae* respecto a bacterias del género *Phytoplasma* reivindica a los virus como entes biológicos centrales en la evolución reciente de linajes celulares, en contraposición al origen antiguo planteado por el virocentrismo.

- Debido al reducido tamaño de los genomas de virus ssDNA, y sus núcleos pangenómicos, una historia más completa y detallada deberá incorporar la información de proteínas asignadas a otras categorías pangenómicas.

- La construcción de pangenomas, la anotación funcional por homología, así como la completitud de las historias evolutivas de los virus ssDNA están abiertas al descubrimiento de nuevos virus.

LITERATURA CITADA

Abrahão, J. S., Araújo, R., Colson, P., & La Scola, B. (2017). The analysis of translation-related gene set boosts debates around origin and evolution of mimiviruses. *PLoS Genetics*, *13*(2), e1006532. 10.1371/journal.pgen.1006532

Adriaenssens, E. M., Sullivan, M. B., Knezevic, P., van Zyl, L. J., Sarkar, B. L., Dutilh, B. E., Alfenas-Zerbini, P., Lobocka, M., Tong, Y., Brister, J. R., Switt, A. I., Klumpp, J., Aziz, R. K., Barylski, J., Uchiyama, J., Edwards, R. A., Kropinski, A. M., Petty, N. K., Clokie, M. R.J., … Krupovic, M. (2020). Taxonomy of prokaryotic viruses: 2018-2019 update from the ICTV Bacterial and Archaeal Viruses Subcommittee. *Archives of Virology*, *165*, 1253–1260. 10.1007/s00705-020-04577-8

Agrelli, A., de Moura, R. R., Crovella, S., & Brandão, L. A.C. (2019). Mutational landscape of Zika virus strains worldwide and its structural impact on proteins. *Gene*, *708*, 57-62. 10.1016/j.gene.2019.05.039

Alborzi, S. Z., Devignes, M.-D., & Ritchie, D. (2017, April). Associating Gene Ontology Terms with Pfam Protein Domains. *5th International Work-Conference on Bioinformatics and Biomedical Engineering - IWBBIO 2017*, 127-138. 10.1007/978-3-319-56154-7_13

Anisimova, M. (Ed.). (2019). *Evolutionary Genomics: Statistical and Computational Methods*. Springer New York.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J.M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., & Sherlock, G. (2000). Gene Ontology: tool for the unification of biology. *Nature Genetics*, *25*(1), 25-29. 10.1038/75556

Assis, F. L., Bajrai, L., Abrahao, J. S., Kroon, E. G., Dornas, F. P., Andrade, K. R., Boratto, P. V., Pilotto, M. R., Robert, C., Benamar, S., Scola, B. L., & Colson, P. (2015, July). Pan-Genome Analysis of Brazilian Lineage A Amoebal Mimiviruses. *Viruses*, *7*(7), 3483–3499. 10.3390/v7072782

Avery, O. T., MacLeod, C. M., & McCarty, M. (1994). Studies on the chemical nature of the substance inducing transformation of pneumococcal types. Induction of transformation by a desoxyribosenucleic acid fraction isolated from pneumococcus type III. *Journal of Experimental Medicine*, *79*(2), 137-158. 10.1084/jem.79.2.137

Aylward, F. O., Moniruzzaman, M., Ha, A. D., & Koonin, E. V. (2021). A phylogenomic framework for charting the diversity and evolution of giant viruses. *PLOS Biology*, *19*(10), e3001430. 10.1371/journal.pbio.3001430

Baltimore, D. (1971, 9). Expression of Animal Virus Genomes. *Bacteriological Reviews*, *35*(3), 235-241. 10.1128/br.35.3.235-241.1971

Bamford, D., Pietila, M., Roine, E., Atanasova, N., Dientsiber, A., & Oksanen, H. (2017). ICTV Virus Taxonomy Profile: Pleolipoviridae. J Gen Virol, 98(12), 2916-2917.

Bándea, C. I. (1983). A new theory on the origin and the nature of viruses. *Journal of Theoretical Biology*, *105*(4), 591-602.

Barocchi, M. A., Masignani, V., & Rappuoli, R. (2005). Cell entry machines: a common theme in nature? *Nature Reviews Microbiology*, *3*, 349–358. 10.1038/nrmicro1131

Baum, D. A., Smith, S. D., & Donovan, S. S.S. (2005). The Tree-Thinking Challenge. *Science*, *310*(5750), 979-980. 10.1126/science.1117727

Becerra, A. C. I., Delaye, L., Islas, S., & Lazcano, A. (2007). The Very Early Stages of Biological Evolution and the Nature of the Last Common Ancestor of the Three Major Cell Domains. *Annual Review of Ecology, Evolution, and Systematics*, *38*, 361-379. 10.1146/annurev.ecolsys.38.091206.095825

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., & Bourne, P. E. (2000, January 1). The Protein Data Bank. *Nucleic Acids Research*, *28*(1). 10.1093/nar/28.1.235

Blum, M., Chang, H.-Y., Chuguransky, S., Grego, T., Kandasaamy, S., Mitchell, A., Nuka, G., Paysan-Lafosse, T., Qureshi, M., Raj, S., Richardson, L., Salazar, G. A., Williams, L., Bork, P., Bridge, A., Gough, J., Haft, D. H., Letunic, I., Marchler-Bauer, A., … Finn, R. (2021, January 8). The InterPro protein families and domains database: 20 years on. *Nucleic Acids Research*, *49*(D1). 10.1093/nar/gkaa977

Brandes, U., & Cornelsen, S. (2009). Phylogenetic graph models beyond trees. *Discrete Applied Mathematics*, *157*(10), 2361-2369. 10.1016/j.dam.2008.06.031

Breitbart, M., & Rohwer, F. (2005, June). Here a virus, there a virus, everywhere the same virus?

Trends in Microbiology, *13*(6), 278-284. 10.1016/j.tim.2005.04.003.

Brister, R. J., Ako-Adjei, D., Bao, Y., & Blinkova, O. (2015, January). NCBI viral genomes resource. *Nucleic Acids Research*, *43*(Database Issue), D571-577. 10.1093/nar/gku1207

Brito, A. F., Brito, A. F. d., Braconi, C. T., Weidmann, M., Dilcher, M., Pereira-Alves, J. M., Gruber, A., & de Andrade Zanotto, P. M. (2016). The Pangenome of the Anticarsia gemmatalis Multiple Nucleopolyhedrovirus (AgMNPV). *Genome Biology and Evolution*, *8*(1), 94-108. 10.1093/gbe/evv231

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009, December 15). BLAST+: architecture and applications. *BMC Bioinformatics*, *10*(421). 10.1186/1471-2105-10-421

Campillo-Balderas, J. A. (2018, Junio). *Origen y evolución temprana de los virus y su relación con el último ancestro común de los seres vivos* [Tesis de Doctorado, Posgrado en Ciencias Biológicas, Universidad Nacional Autónoma de México].

Campillo-Balderas, J. A., Lazcano, A., & Becerra, A. (2015, December). Viral genome size distribution does not correlate with the antiquity of the host lineages. *3*(143). 10.3389/fevo.2015.00143

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. doi:10.1093/bioinformatics/btp348

Carroll, I. P., & Rein, A. (2016). Viral Nucleic Acids. In *Encyclopedia of Cell Biology* (pp. 517-524). Ralph A. Bradshaw & Philip D. Stahl. 10.1016/B978-0-12-394447-4.10061-6

Chamberlain, S. A., & Szöcs, E. (2021, July 26). taxize: taxonomic search and retrieval in R. *F1000 Research*, *191*(2). https://doi.org/10.12688/f1000research.2-191.v2

Chapman, M. S., & Rossmann, M. G. (1993, June). Structure, Sequence, and Function Correlations among Parvoviruses. *Virology*, *194*(2), 491-508. 10.1006/viro.1993.1288

Charrad, M., Ghazzall, N., Bolteau, V., & Niknafs, A. (2014). NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set. *Journal of Statistical Software*, *61*(6), 1-36. 10.18637/jss.v061.i06

Chenuil, A. (2006). Choosing the right molecular genetic markers for studying biodiversity: from molecular evolution to practical aspects. *Genetica*, *127*(1-3), 101-120. 10.1007/s10709-005-2485-1

Chirico, N., Vianelli, A., & Belshaw, R. (2010). Why genes overlap in viruses. *Proceedings of The Royal Society Biological Sciences*, *277*(1701), 3809-3817. 10.1098/rspb.2010.1052

Chuong, E. B. (2013, October). Retroviruses facilitate the rapid evolution of the mammalian placenta. *Bioessays*, *35*(10). 10.1002/bies.201300059

Contreras-Moreira, B., & Vinuesa, P. (2013, December). GET_HOMOLOGUES, a Versatile Software Package for Scalable and Robust Microbial Pangenome Analysis. *Applied and Environmental Microbiology*, *79*(24), 7696-7701. 10.1128/AEM.02411-13

Delaye, L., Becerra, A., & Lazcano, A. (2005). The Last Common Ancestor: What's in a name? *Origins of Life and Evolution of Biospheres*, *35*(6), 537–554. 10.1007/s11084-005-5760-3

D'Herelle, F. (1926). The Bacteriophage and its Behaviour. *Nature*, *118*, 183–185. 10.1038/118183a0

Duffy, S., & Shackelton, L. A. (2008). Rates of evolutionary change in viruses: patterns and determinants. *Nature Reviews Genetics*, *9*, 267-276. 10.1038/nrg2323

Eddy, S. R. (2011, October). Accelerated Profile HMM Searches. *PLOS Computational Biology*, *7*(10), 1-16. 10.1371/journal.pcbi.1002195

Edson, K. M., Vinson, B., Stoltz, D. B., & Summers, M. S. (1981, Feb 6). Virus in a Parasitoid Wasp: Suppression of the Cellular Immune Response in the Parasitoid's Host. *Science*, *211*(4482), 582-583. 10.1126/science.7455695

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., & Finn, R. D. (2019, January 08). The Pfam protein families database in 2019. *Nucleic Acids Research*, *47*(D1), D427-D432. 10.1093/nar/gky995

Flint, J., Racaniello, V., Rall, G., Hatziioannou, T., & Skalka, A. M. (2020). *Principles of Virology* (Fifth ed., Vol. I). American Society for Microbiology & Wiley. ISBN: 978-1-683-67358-3

Forterre, P. (2006). The origin of viruses and their possible roles in major evolutionary

transitions. *Virus Research*, *117*(1), 5-16. 10.1016/j.virusres.2006.01.010

Forterre, P. (2013). The virocell concept and environmental microbiology. *The ISME Journal*, *7*, 233–236. 10.1038/ismej.2012.110

Forterre, P. (2015). The two ages of the RNA world, and the transition to the DNA world, a story of viruses and cells. *Biochimie*, *87*(9-10), 793-803. 10.1016/j.biochi.2005.03.015

Forterre, P., & Krupovic, M. (2012). The Origin of Virions and Virocells: The Escape Hypothesis Revisited. In G. Witzany (Ed.), *Viruses: Essential Agents of Life* (pp. 43-60). Springer Netherlands. 10.1007/978-94-007-4899-6_3

Francia, M. V., & Clewell, D. B. (2002). Transfer origins in the conjugative Enterococcus faecalis plasmids pAD1 and pAM373: identification of the pAD1 nic site, a specific relaxase and a possible TraG-like protein. *Molecular Microbiology*, *45*(2), 375-395. 10.1046/j.1365-2958.2002.03007.x

Frickey, T., & Lupas, A. (2004, December 12). CLANS: a Java application for visualizing protein families based on pairwise similarity. *Bioinformatics*, *20*(18), 3702-3704. 10.1093/bioinformatics/bth444

Fry, I. (2006). The origins of research into the origins of life. *Endeavour*, *30*(1), 24-28. 10.1016/j.endeavour.2005.12.002

The Gene Ontology Consortium. (2019, January). The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Research*, *47*(D1), D330-D338. 10.1093/nar/gky1055

The Gene Ontology Consortium. (2021). The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Research*, *49*(D), D325-D334. 10.1093/nar/gkaa1113

Hansen, J. L., Long, A. M., & Schultz, S. C. (1997). Structure of the RNA-dependent RNA polymerase of poliovirus. *Structure*, *5*(8), 1109-1122. 10.1016/S0969-2126(97)00261-X

Harris, J. R. (1991). The evolution of placental mammals. *Federation of European Biochemical Societies*, *295*(1,2,3), 3-4.

Hino, S., & Miyata, H. (2007). Torque teno virus (TTV): current status. *Rev Med Virol*, *17*(1), 45-57. 10.1002/rmv.524

Howard Hughes Medical Institute. (2020, November). *HMMER: biosequence analysis using profile hidden Markov models*. HMMER. http://hmmer.org/

Hulo, C., de Castro, E., Masson, P., Bougueleret, L., Bairoch, A., Xenairos, I., & Le Mercier, P. (2020). *ssDNA*. ViralZone. Retrieved January 14, 2021, from https://viralzone.expasy.org/283

Ilyina, T. V., & Koonin, E. V. (1992). Conserved sequence motifs in the initiator proteins for rolling circle DNA replication encoded by diverse replicons from eubacteria, eucaryotes and archaebacteria. *Nucleic Acids Research*, *20*(13), 3279–3285. 10.1093/nar/20.13.3279

Ilyna, T. V., & Koonin, E. V. (1992, July). Conserved sequence motifs in the initiator proteins for rolling circle DNA replication encoded by diverse replicons from eubacteria, eucaryotes and archaebacteria. *Nucleic Acids Research*, *20*(13), 3279-3285. 10.1093/nar/20.13.3279

Jackson, A. P. (2015). The evolution of parasite genomes and the origins of parasitism. *Parasitology*, *142*(S1), S1 - S5. 10.1017/S0031182014001516

Jácome, R., Becerra, A., Ponce de León, S., & Lazcano, A. (2015, September). Structural analysis of monomeric RNA-dependent polymerases: evolutionary and therapeutic implications. *PloS one*, *10*(9), 29. 10.1371/journal.pone.0139001

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A., & Jermiin, L. S. (2017, June). ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature Methods*, *14*(6), 587-589. 10.1038/nmeth.4285

Karp, G. (2013). *Biología celular y molecular: conceptos y experimentos* (Séptima ed.). McGraw-Hill.

Katoh, K., Misawa, K., Kuma, K., & Miyata, T. (2002, July 15). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, *30*(14), 3059-3066. 10.1093/nar/gkf436

Kazlauskas, D., Dayaram, A., Kraberger, S., Goldstien, S., Varsani, A., & Kuprovic, M. (2017, April). Evolutionary history of ssDNA bacilladnaviruses features horizontal acquisition of the capsid gene from ssRNA nodaviruses. *Virology*, *504*, 114-121. 10.1016/j.virol.2017.02.001

Kazlauskas, D., Varsani, A., Koonin V, E., & Krupovic, M. (2019, July). Multiple origins of prokaryotic and eukaryotic single-stranded DNA viruses from bacterial and archaeal plasmids. *Nature Communications*, *10*(1), 3425. 10.1038/s41467-019-11433-0

Keeling, P. J., & Palmer, J. D. (2008). Horizontal gene transfer in eukaryotic evolution. *Nature Reviews, Genetics*, *9*(8), 605-18. 10.1038/nrg2386

Koonin, E., Krupovic, M., & Agol, V. (2021, 8). The Baltimore Classification of Viruses 50 Years Later: How Does It Stand in the Light of Virus Evolution? *Microbiology and Molecular Biology Reviews*, *85*(3), 18. 10.1128/MMBR.00053-21

Koonin, E. V., & Dolja, V. V. (2014). Virus World as an Evolutionary Network of Viruses and Capsidless Selfish Elements. *Microbiology and Molecular Biology Reviews*, *78*(2), 278-303. 10.1128/MMBR.00049-13

Koonin, E. V., Krupovic, M., & Yutin, N. (2015). Evolution of double-stranded DNA viruses of eukaryotes: from bacteriophages to transposons to giant viruses. *Annals of the New York Academy of Sciences*, *1341*(1), 10-24. 10.1111/nyas.12728

Koonin, E. V., Senkevich, T. G., & Dolja, V. V. (2006). The ancient Virus World and evolution of cells. *BioMed Central Page 1 of 27 (page number not for citation purposes) Biology Direct*, *1*(29). 10.1186/1745-6150-1-29

Kristensen, D. M., Kannan, L., Coleman, M. K., Wolf, Y. I., Sorokin, A., Koonin, E. V., & Mushegian, A. (2010, June). A low-polynomial algorithm for assembling clusters of orthologous groups from intergenomic symmetric best matches. *Bioinformatics*, *26*(12), 1481-1484. 10.1093/bioinformatics/btq229

Kristensen, D. M., Walker, A. S., Yamada, T., Bork, P., Musheigan, A. R., & Koonin, E. V. (2013). Orthologous Gene Clusters and Taxon Signature Genes for Viruses of Prokaryotes. *Journal of Bacteriology*, *195*(5), 941-950. 10.1128/JB.01801-12

Krupovic, M. (2013, October). Networks of evolutionary interactions underlying the polyphyletic origin of ssDNA viruses. *Current Opinion in Virology*, *3*(5), 578-586. 10.1016/j.coviro.2013.06.010

Krupovic, M., & Forterre, P. (2015, 4). Single-stranded DNA viruses employ a variety of mechanisms for integration into host genomes. *Annals of The New York Academy of Sciences*, *1341*(1), 41-53. 10.1111/nyas.12675

Krupovic, M., Ghabrial, S. A., Jiang, D., & Varsani, A. (2016). Genomoviridae: a new family of widespread single-stranded DNA viruses. *Archives of Virology*, *161*, 2633–2643. 10.1007/s00705-016-2943-3

Krupovic, M., Ravantti, J. J., & Bamford, D. H. (2009). Geminiviruses: a tale of a plasmid becoming a virus. *BMC Ecology and Evolution*, *9*, 112. 10.1186/1471-2148-9-112

Kuprovic, M., & Koonin, E. (2014). Evolution of eukaryotic single-stranded DNA viruses of the Bidnaviridae family from genes of four other groups of widely different viruses. *Scientific Reports*, *18*(4), 5347. 10.1038/srep05347

Landry, M. L. (2016). Parvovirus B19. *Microbiol Spectr*, *4*(3). 10.1128/microbiolspec

Lefkowitz, E. J., Dempsey, D. M., Curtis Hendrickson, R., Orton, R. J., Siddekk, S., & Smith, D. (2017, October). Virus taxonomy: the database of the International Committee on Taxonomy of Viruses (ICTV). *Nucleic Acids Research*, *46*(14), D708 - D717. 10.1093/nar/gkx932

Leimester, C.-A., Sohrabi-Jahromi, S., & Morgenstrem, B. (2017, April). Fast and accurate phylogeny reconstruction using filtered spaced-word matches. *Bioinformatics*, *33*(7), 971-979. 10.1093/bioinformatics/btw776

Li, L., Stoeckert Jr, C. J., & Roos, D. S. (2003, September). OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. *Gemome Research*, *13*(9), 2178-2189. 10.1101/gr.1224503

Li, W., & Godzik, A. (2006, July 1). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, *22*(13), 1658-1659. 10.1093/bioinformatics/btl158

Liu, H., Fu, Y., Yu, X., Xie, J., Cheng, J., Ghabrial, S. A., Li, G., Yi, X., & Jiang, D. (2011, September). Widespread Horizontal Gene Transfer from Circular Single-stranded DNA Viruses to Eukaryotic Genomes. *BMC Ecology and Evolution*, *11*, 276. 10.1186/1471-2148-11-276

López-García, P. (2012). The Place of Viruses in Biology in Light of the Metabolism-versus-replication-first Debate. *History and Philosophy of the Life Sciences*, *34*(3), 391-406. http://www.jstor.org/stable/43831419

Lucía-Sanz, A., & Manrubia, S. (2017, November). Multipartite viruses: adaptive trick or evolutionary treat? *npj Systems Biology and Applications*, *3*(Article number: 34). 10.1038/s41540-017-0035-y

Lukashev, A. N. (2010). Recombination among picornaviruses. *Reviews in Medical Virology*, *20*(5), 327-337. 10.1002/rmv.660

Mann, N. H., Cook, A., Millard, A., Bailey, S., & Clokie, M. (2003). Bacterial photosynthesis genes in a virus. *Nature*, *424*(741). 10.1038/424741a

Márquez, L. M., Redman, R. S., Rodríguez, R. J., & Roossinck, M. J. (2007, January). A virus in a fungus in a plant: three-way symbiosis required for thermal tolerance. *Science*, *315*(5822), 513-5. 10.1126/science.1136237

Martin, D. P., Biagini, P., Lefeuvre, P., Golden, M., Roumagnac, P., & Varsani, A. (2011, September). Recombination in Eukaryotic Single Stranded DNA Viruses. *Viruses*, *3*(9), 1699-1738. 10.3390/v3091699

Mihara, T., Nishimura, Y., Shimizu, Y., Nishiyama, H., Yoshikawa, G., Uehara, H., Hingamp, P., Goto, S., & Ogata, H. (2016, March 1). Linking Virus Genomes with Host Taxonomy. *Viruses*, *8*(3), 66. 10.3390/v8030066

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., & Lanfear, R. (2020, May). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution*, *37*(5), 1530–1534. 10.1093/molbev/msaa015

Minh, B. Q., Thi Nguyen, M. A., & von Haeseler, A. (2013, May). Ultrafast Approximation for Phylogenetic Bootstrap. *Molecular Biology and Evolution*, *30*(5), 1188–1195. 10.1093/molbev/mst024

Mochizuki, T., Krupovic, M., Pehau-Arnaudet, G., Sako, Y., Forterre, P., & Prangishvili, D. (2012). Archaeal virus with exceptional virion architecture and the largest single-stranded DNA genome. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(33), 13386–13391. 10.1073/pnas.1203668109

Morales, F. J., & Anderson, P. K. (2001). The emergence and dissemination of whitefly-transmitted geminiviruses in Latin America. *Arch Virol*, *146*(3), 415-441. 10.1007/s007050170153

Moreira, D. (2000). Multiple independent horizontal transfers of informational genes from bacteria to plasmids and phages: implications for the origin of bacterial replication machinery. *Molecular Microbiology*, *35*(1), 1-5. 10.1046/j.1365-2958.2000.01692.x

Moreira, D., & Brochier-Armanet, C. (2008). Giant viruses, giant chimeras: The multiple evolutionary histories of Mimivirus genes. *BMC Evolutionary Biology*, *2*(12). 10.1186/1471-2148-8-12

Moreira, D., & López-García, P. (2009, March 9). Ten reasons to exclude viruses from the tree of life. *Nature Reviews Microbiology*, *7*, 306–311. 10.1038/nrmicro2108

Nasir, A., Forterre, P., Kim, K. M., & Caetano-Anollés, G. (2014, April 30). The distribution and impact of viral lineages in domains of life. *Frontiers in Microbiology*, *5*. 10.3389/fmicb.2014.00194

Olson, A. J., Hu, Y. H.E., & Keinan, E. (2007). Chemical mimicry of viral capsid self-assembly. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(52), 20731-20736. 10.1073/pnas.0709489104

Onafuwa-Nuga, A., & Telesnitsky, A. (2009). The remarkable frequency of human immunodeficiency virus type 1 genetic recombination. *Microbiology and Molecular Biology Reviews*, *73*(3), 451-480. 10.1128/MMBR.00012-09

Pagès, H., Aboyoun, P., Gentleman, R., & DebRoy, S. (2021). *Biostrings: Efficient manipulation of biological strings.* (3.14) [R package version 3.14]. https://bioconductor.org/packages/Biostrings

Payne, S. (2017). *Viruses: From Understanding to Investigation*. Elsevier Science. 10.1016/C2014-0-03894-4

Pei, S., Dong, R., He, R. L., & Yau, S. (2019, July). Large-Scale Genome Comparison Based on Cumulative Fourier Power and Phase Spectra: Central Moment and Covariance Vector. *Computational and Structural Biotechnology Journal*, *17*, 982-994. 10.1016/j.csbj.2019.07.003

Pérez-Losada, M., Arenas, M., & Galán, J. C. (2015). Recombination in viruses: Mechanisms, methods of study, and evolutionary consequences. *Infection, Genetics and Evolution*, *30*, 296-307. 10.1016/j.meegid.2014.12.022

Raoult, D., Audic, S., Robert, C., Abergel, C., Renesto, P., Ogata, H., La Scola, B., Suzan, M., & Claverie, J.-M. (2004). The 1.2-Megabase Genome Sequence of Mimivirus. *Science*, *306*(5700), 1344-1350. 10.1126/science

Raoult, D., & Forterre, P. (2008). Redefining viruses: lessons from Mimivirus. *Nature Reviews Microbiology*, *6*(4), 315-319. 10.1038/nrmicro1858

Retel, C., Märkle, H., Becks, L., & Feulner, P. G.D. (2019, March). Ecological and Evolutionary

Processes Shaping Viral Genetic Diversity. *Viruses*, *11*(3), 220. 10.3390/v11030220

Rieseberg, L. H. (1997). Hybrid Origins of Plant Species. *Annual Review of Ecology and Systematics*, *28*(1), 359-389. 10.1146/annurev.ecolsys.28.1.359

Robson, F., Khan, K. S., Le, T. K., Paris, C., Demirbag, S., Barfuss, P., Rocchi, P., & Ng, W.-L. (2020). Coronavirus RNA Proofreading: Molecular Basis and Therapeutic Targeting. *Molecular Cell*, *79*(5), 710-727. 10.1016/j.molcel.2020.07.027

Roossinck, M., & Bazán, E. R. (2017). Symbiosis: Viruses as Intimate Partners. *Annual Review of Virology*, *4*, 123-139. 10.1146/annurev-virology-110615-042323

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

Roossinck, M. J. (2015, July). Move Over, Bacteria! Viruses Make Their Mark as Mutualistic Microbial Symbionts. *Journal of Virology*, *89*(13), 6532-5. 10.1128/JVI.02974-14

Rosario, K., Breitbart, M., Harrach, B., Seagles, J., Delwart, E., Biagini, P., & Varsani, A. (2017). Revisiting the taxonomy of the family Circoviridae: Establishment of the genus Cyclovirus and removal of the genus Gyrovirus. *Archives of Virology*, *162*, 1447-1463. 10.1007/s00705-017-3247-y

Rosario, K., Duffy, S., & Breitbart, M. (2012). A field guide to eukaryotic circular single-stranded DNA viruses: insights gained from metagenomics. *Archives of Virology*, *157*, 1851–1871. 10.1007/s00705-012-1391-y

Roux, S., Enault, F., Bronner, G., Vaulot, D., Forterre, P., & Krupovic, M. (2013). Chimeric viruses blur the borders between the major groups of eukaryotic single-stranded DNA viruses. *Nature Communications*, *4*, 2700. 10.1038/ncomms3700

Roux, S., Krupovic, M., Daly, R. A., Borges, A. L., Nayfach, S., Schulz, F., Sharrar, A., Matheus-Carnevali, P. B., Cheng, J.-F., Ivanova, N. N., Bondy-Denomy, J., Wrighton, K. C., Woyke, T., Visel, A., Kyrpides, N. C., & Eloe-Fadrosh, E. A. (2019). Cryptic inoviruses revealed as pervasive in bacteria and archaea across Earth's biomes. *Nature Microbiology*, *4*, 1895–1906. 10.1038/s41564-019-0510-x

Rupp, R., Scornavacca, C., & Huson, D. H. (2011). *Phylogenetic Networks: Concepts, Algorithms and Applications*. Cambridge University Press. 10.1017/CBO9780511974076

Saccardo, F., Cettul, E., Palmano, S., Noris, E., & Firrao, G. (2011). On the alleged origin of geminiviruses from extrachromosomal DNAs of phytoplasmas. *BMC Evolutionary Biology*, *11*(185). 10.1186/1471-2148-11-185

Sanjuán, R., & Domingo-Calap, P. (2016, December). Mechanisms of viral mutation. *Cellular and Molecular Life Sciences*, *73*(23). 10.1007/s00018-016-2299-6

Sanjuán, R., & Pilar, D. C. (2016, December). Mechanisms of viral mutation. *Cellular and Molecular Life Sciences*, *73*(23), 4433-4448. 10.1007/s00018-016-2299-6

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). Consensus statement: Virus taxonomy in the age of metagenomics. *Nature Reviews Microbiology*, *15*(3), 161-168. 10.1038/nrmicro.2016.177

Simon-Loriere, E., & Holmes, E. C. (2011). Why do RNA viruses recombine? *Nature Reviews Microbiology*, *9*, 617–626. 10.1038/nrmicro2614

Smits, S. L., Zijlstra, E. E., van Hellemond, J. J., Schapendonk, C. M.E., Bodewes, R., Schürch, A. C., Haagmans, B. L., & Osterhaus, A. D.M.E. (2013). Novel Cyclovirus in Human Cerebrospinal Fluid, Malawi, 2010–2011. *Emerging Infectious Diseases*, *19*(9), 1511-1513. 10.3201/eid1909.130404

Suzek, B. E., Wang, Y., Huang, H., McGarvey, P. B., Wu, C. H., & The UniProt Consortium. (2014, November 13). UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, *31*(6), 926-932. 10.1093/bioinformatics/btu739

Syvanen, M. (1985). Cross-species gene transfer; implications for a new theory of evolution. *Journal of Theoretical Biology*, *112*(2), 333–343. 10.1016/s0022-5193(85)80291-5

Tello Lacal, C. (Ed.). (2019). *Principles of Molecular Virology*. Arcler Education Incorporated. https://eds.p.ebscohost.com/eds/ebookviewer/ebook/bmxlYmtfXzIwMTM5OTZfX0FO0?sid=e03ecada-24ec-4f4a-a1b0-3d930dbc43e9@redis&vid=7&format=EB&rid=2

Thomas, J., & Pritham, E. J. (2015). Helitrons, the eukaryotic rolling-circle transposable elements. *Microbiology Spectrum*, *3*(MDNA3-0049-2014). 10.1128/microbiolspec.MDNA3-0049-2014

The UniProt Consortium. (2021, January 8). UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research*, *49*(D1), D480-D489. 10.1093/nar/gkaa1100

Van Regenmortel, M. H.V. (2018). The Species Problem in Virology (M. Kielian, T. C. Mettenleiter, & M. J. Roossinck, Eds.). *Advances in Virus Research*, *100*, 1-18. 10.1016/bs.aivir.2017.10.008

Varsani, A., & Krupovic, M. (2018). Smacoviridae: a new family of animal-associated single-stranded DNA viruses. *Archives of Virology*, *163*, 2005–2015. 10.1007/s00705-018-3820-z

Vernikos, G., Medini, D., Riley, D. R., & Tettelin, H. (2015). Ten years of pan-genome analyses. *Current Opinion in Microbiology*, *23*, 148-154. 10.1016/j.mib.2014.11.016

Wang, H., Wu, S., Li, K., Pan, Y., Yan, S., & Wang, Y. (2018). Metagenomic analysis of ssDNA viruses in surface seawater of Yangshan Deep-Water Harbor, Shanghai, China. *Marine Genomics*, *41*, 50-53. 10.1016/j.margen.2018.03.006

Watson, J. D., & Crick, F. H.C. (1953). Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature*, (171), 737–738. 10.1038/171737a0

Wawrzyniak, P., Grażyna, P., & Bartosik, D. (2017). The Different Faces of Rolling-Circle Replication and Its Multifunctional Initiator Proteins. *Frontiers in Microbiology*, *8*. 10.3389/fmicb.2017.02353

Woese, C. R., Kandler, O., & Wheelis, M. L. (1990). Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proceedings of the Natural Academy of Sciences United States of America*, *87*(12), 4576–4579.

Wolf, Y. I., Kazlauskas, D., Iranzo, J., Lucía-Sanz, A., Kuhn, J. H., Krupovic, M., Dolja, V. V., & Koonin, E. (2018). Origins and Evolution of the Global RNA Virome. *mBio*, *9*(6), e02329-18. 10.1128/mBio.02329-18

Zhao, L., Lavington, E., & Duffy, S. (2021). Truly ubiquitous CRESS DNA viruses scattered across the eukaryotic tree of life. *Journal of Evolutionary Biology*, *34*(12), 1901-10916. 10.1111/jeb.13927

Zhao, L., Rosario, K., Breitbart, M., & Duffy, S. (2019). Eukaryotic Circular Rep-Encoding Single-Stranded DNA (CRESS DNA) Viruses: Ubiquitous Viruses With Small Genomes and a Diverse Host Range. *Advances in Virus Research*, *103*, 71-133. 10.1016/bs.aivir.2018.10.001

Zielezinski, A., Girgis, H. Z., Bernard, G., Leimester, C. A., Tang, K., Dencker, T., Katharina Lau, A., Röhling, S., Choi, J. J., Waterman, M., Comin, M., Kim, S. H., Vinga, S., Almeida, J. S., Chan, C., James, B. T., Sun, F., Morgenstern, B., & Karlowski, W. (2019, July). Benchmarking of alignment-free sequence comparison methods. *Genome Biology*, *20*(1), 144. 10.1186/s13059-019-1755-7

Zúñiga-Vega, C., & Ramírez, P. (2002). Los geminivirus, patógenos de importancia mundial. *Manejo Integrado de Plagas y Agroecología*, (64), 25-33.