



**UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO**  
**POSGRADO EN CIENCIAS BIOLÓGICAS**  
**FACULTAD DE CIENCIAS**

**ORIGEN Y EVOLUCIÓN TEMPRANA DE LOS VIRUS Y SU RELACIÓN  
CON EL ÚLTIMO ANCESTRO COMÚN DE LOS SERES VIVOS**

**TESIS**

QUE PARA OPTAR POR EL GRADO DE:  
**DOCTOR EN CIENCIAS**

PRESENTA:  
**JOSÉ ALBERTO CAMPILLO BALDERAS**

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

COMITÉ TUTOR: **DR. JOSÉ LUIS DELAYE ARREDONDO**  
CINVESTAV UNIDAD IRAPUATO

COMITÉ TUTOR: **DR. LEÓN PATRICIO MARTÍNEZ CASTILLA**  
FACULTAD DE QUÍMICA, UNAM

**MÉXICO, CD. MX.**

**JUNIO 2018**



Universidad Nacional  
Autónoma de México

Dirección General de Bibliotecas de la UNAM

**Biblioteca Central**



**UNAM – Dirección General de Bibliotecas**  
**Tesis Digitales**  
**Restricciones de uso**

**DERECHOS RESERVADOS ©**  
**PROHIBIDA SU REPRODUCCIÓN TOTAL O PARCIAL**

Todo el material contenido en esta tesis esta protegido por la Ley Federal del Derecho de Autor (LFDA) de los Estados Unidos Mexicanos (México).

El uso de imágenes, fragmentos de videos, y demás material que sea objeto de protección de los derechos de autor, será exclusivamente para fines educativos e informativos y deberá citar la fuente donde la obtuvo mencionando el autor o autores. Cualquier uso distinto como el lucro, reproducción, edición o modificación, será perseguido y sancionado por el respectivo titular de los Derechos de Autor.





**UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO**  
**POSGRADO EN CIENCIAS BIOLÓGICAS**  
**FACULTAD DE CIENCIAS**

**ORIGEN Y EVOLUCIÓN TEMPRANA DE LOS VIRUS Y SU RELACIÓN  
CON EL ÚLTIMO ANCESTRO COMÚN DE LOS SERES VIVOS**

**TESIS**

**QUE PARA OPTAR POR EL GRADO DE:  
DOCTOR EN CIENCIAS**

**PRESENTA:  
JOSÉ ALBERTO CAMPILLO BALDERAS**

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

**COMITÉ TUTOR: DR. JOSÉ LUIS DELAYE ARREDONDO**  
CINVESTAV UNIDAD IRAPUATO

**COMITÉ TUTOR: DR. LEÓN PATRICIO MARTÍNEZ CASTILLA**  
FACULTAD DE QUÍMICA, UNAM

**MÉXICO, CD. MX.**

**JUNIO 2018**



POSGRADO EN CIENCIAS BIOLÓGICAS  
FACULTAD DE CIENCIAS  
DIVISIÓN ACADÉMICA DE INVESTIGACIÓN Y POSGRADO

OFICIO FCIE/DAIP/470/2018

ASUNTO: Oficio de Jurado


Lic. Ivonne Ramírez Wence  
Directora General de Administración Escolar, UNAM  
Presente

Me permito informar a usted que en la reunión ordinaria del Comité Académico del Posgrado en Ciencias Biológicas, celebrada el día 12 de marzo de 2018, se aprobó el siguiente jurado para el examen de grado de DOCTOR EN CIENCIAS del (la) alumno (a) CAMPILLO BALDERAS JOSÉ ALBERTO con número de cuenta 505017473 con la tesis titulada: "ORIGEN Y EVOLUCIÓN TEMPRANA DE LOS VIRUS Y SU RELACIÓN CON EL ÚLTIMO ANCESTRO COMÚN DE LOS SERES VIVOS", realizada bajo la dirección del (la) DR. ARTURO CARLOS II BECERRA BRACHO:

Presidente:	DR. ANTONIO EUSEBIO LAZCANO-ARAUJO REYES
Vocal:	DR. LUIS DAVID ALCARAZ PERAZA
Secretario:	DRA. BEATRIZ GÓMEZ GARCÍA
Suplente:	DR. LUIS JOSÉ DELAYE ARREDONDO
Suplente:	DR. CARLOS CABELLO GUTIÉRREZ

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

ATENTAMENTE  
"POR MI RAZA HABLARA EL ESPIRITU"  
Ciudad Universitaria, Cd. Mx., a 08 de mayo de 2018

  
DR. ADOLFO GERARDO NAVARRO SIGÜENZA  
COORDINADOR DEL PROGRAMA



AGNS/MMVA/ASR/ipp

## **AGRADECIMIENTOS INSTITUCIONALES**

Al Posgrado en Ciencias Biológicas (PCB) de la Universidad Nacional Autónoma de México (UNAM) por todo el apoyo académico otorgado para mi formación científica.

Al Consejo Nacional de Ciencia y Tecnología (Conacyt) por concederme la beca de manutención para realizar mis estudios de doctorado (CVU 165264).

Al Programa de Apoyo para Estudios de Posgrado (PAEP) del PCB por otorgarme el apoyo financiero para asistir a dos congresos internacionales y a través de la Convocatoria de Mejoras a la Tasa de Graduación del Doctorado en Ciencias Biológicas.

Al Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica (PAPIIT) por apoyarnos económicamente a través del proyecto de investigación IN223916.

A mi Tutor Principal, el Dr. Arturo Carlos Il Becerra Bracho, y a los miembros de mi Comité Tutor, el Dr. José Luis Delaye Arredondo y el Dr. León Patricio Martínez Castilla, por aceptar la dirección de esta tesis y por su apoyo académico en la elaboración de la misma.

## AGRADECIMIENTOS A TÍTULO PERSONAL

A mis padres y demás familia por su amor incondicional y por su paciencia sobre mi ausencia en muchos momentos familiares importantes debido a la realización de esta tesis.

Al Dr. Arturo Becerra y al Al Dr. Antonio Lazcano por su sabiduría, amistad, apoyo, consejos, paciencia, sentido del humor mexicano e inglés y por su interminable deseo de tratar de conquistar al mundo con rotíferos gigantes.

A mi Comité Tutor por sus comentarios, llamadas de atención, sugerencias y amistad.

A mis queridos Macacos por su amistad, apoyo, afecto y acoso psicológico.

A mi amada UNAM por su cobijo académico.

A mis profesores que impartieron cátedra en el PCB por sus enseñanzas.

Y, por supuesto...



# ÍNDICE

<b>RESUMEN</b>	<b>8</b>
<b>ABSTRACT</b>	<b>9</b>
<b>I. INTRODUCCIÓN</b>	<b>10</b>
1.1 Características generales de los virus	10
1.2 Características genómicas y genéticas de los virus	11
1.3 Características moleculares y ecológicas de los virus en relación con sus hospederos	12
1.4 Características evolutivas	13
1.4.1 Origen de los virus	14
1.4.1.1 Hipótesis del virocentrismo (origen precelular)	14
1.4.1.2 Hipótesis de la regresión celular (origen post-celular temprano)	15
1.4.1.3 Hipótesis del escape (origen post-celular tardío)	15
1.4.1.4 Hipótesis de la coevolutiva a largo plazo (origen simultáneo o precelular)	16
1.4.1.5 Hipótesis de las vesículas (origen simultáneo y posterior)	16
1.4.2 Estrategias metodológicas para abordar el problema sobre su origen	17
1.4.2.1 Estrategia basada en secuencia primaria de proteínas	17
1.4.2.2 Estrategia basada en organización del genoma	18
1.4.2.3 Estrategia basada en estructura terciaria	18
1.4.3 Estudios pangenómicos	19
1.5 El último ancestro común de los seres vivos y los virus	20
1.6 Los megavirus y el LCA	20
<b>II. MATERIALES Y MÉTODOS</b>	<b>22</b>
2.1 Construcción de la base de datos con información biológica y ecológica de los virus	22
2.2 Análisis de los datos biológicos y ecológicos de los virus	22
2.3 Construcción de la base de datos pangenómica de los virus	23
2.4 Análisis pangenómicos de los megavirus	24
2.5 Clasificación funcional de los grupos de homólogos del pangenoma de megavirus	25
2.6 Búsqueda de homólogos en bases de datos celulares y virales	28
2.7 Análisis filogenéticos basados en estructura primaria del repertorio del pangenoma de megavirus	28
2.8 Construcción de la base de datos de estructuras terciarias	29
2.9 Análisis filogenéticos basados en estructuras terciarias	29
<b>III. RESULTADOS</b>	<b>31</b>
3.1 Bases de datos biológicos y ecológicos	31
3.2 Bases de datos de proteomas virales	31
3.3 Pagenoma viral de los megavirus	31



3.4 Composición funcional del pangenoma de los megavirus	37
3.5 Análisis filogenéticos basados en la estructura primaria	41
3.6 Análisis filogenéticos basados en la estructura terciaria	52
<b>IV. DISCUSIÓN</b>	<b>53</b>
<b>V. CONCLUSIONES</b>	<b>60</b>
<b>VI. PERSPECTIVAS</b>	<b>61</b>
<b>VII. REFERENCIAS</b>	<b>62</b>

# RESUMEN

La genómica comparada ha permitido trazar la historia evolutiva de todos los seres vivos y ha proporcionado evidencia indirecta de la existencia del último ancestro común a todos ellos llamado LCA (*Last Common Ancestor*). Sin embargo, la descripción de las relaciones evolutivas entre los virus a través del análisis de datos genómicos no ha permitido determinar con claridad su origen. Por otro lado, la reciente disponibilidad de más datos biológicos, genómicos, estructurales y ecológicos de los virus de DNA y RNA en las bases de datos públicas proporciona la oportunidad de inferir y analizar, con más detalle, las relaciones evolutivas que guardan entre ellos mismos y con sus hospederos.

En la presente tesis, nosotros hemos comparado y analizado los datos biológicos y ecológicos recientes, la composición pangenómica y la filogenómica de varias familias virales para determinar su posible origen y evolución temprana con respecto a sus hospederos Bacteria, Archaea y Eukarya. Algunos virólogos sostienen que debido al tamaño y a la composición química de los virus de RNA, éstos surgieron en el Mundo del RNA. Sin embargo, nuestros resultados han revelado que dichas características no muestran una correlación con la distribución de estos virus y la filogenia de los hospederos correspondientes. Nuestros resultados han mostrado que la mayoría de los virus de RNA infectan solo a Eukarya, con excepción de los Cystoviridae y los Leviviridae que sólo infectan a proteobacterias que forman parte de la microbiota de algunos animales. Aún no se han encontrado virus de RNA en Archaea. Por otra parte, los virus de DNA de doble cadena (dsDNA), como los fagos, sólo infectan a Bacteria y a Archaea, pero no a Eukarya. Otros virus de genomas grandes de dsDNA, como los virus citoplásmicos gigantes (megavirus), solo infectan amibas (protistas) y no a otros linajes eucariontes evolutivamente más recientes.

Por otro lado, nuestros análisis pangenómicos y la construcción de filogenias han mostrado que las proteínas que se encuentran altamente conservadas en los megavirus (“núcleo pangenómico”) intervienen en procesos de replicación y reparación del DNA, transcripción y señalización. La mayoría de estas proteínas tienen un origen celular (protistas, plantas, hongos y animales) y, por lo tanto, son probablemente las más antiguas al resto del pangenoma. Estos mismos resultados han revelado que aquellas proteínas virales que no están muy conservadas (“cubierta y nube pangenómica”) intervienen en algunos procesos genéticos, celulares y metabólicos, pero mayoritariamente tienen funciones desconocidas. La mayoría de las filogenias del resto del pangenoma ha revelado un posible origen celular reciente y otras proteínas virales podrían ser homólogos distantes de células de acuerdo a estudios preliminares de comparación de estructuras terciarias. Estos resultados podrían indicar que los virus están relacionados con la historia evolutiva de sus hospederos celulares, es decir, los virus de RNA tienen un origen más reciente que debe ser visto como una coevolución con los eucariontes lo cual indicaría que no tuvieron un origen en el Mundo del RNA. Por otro lado, los virus de DNA podrían tener un origen más antiguo que se remonta posiblemente al origen mismo del LCA.

# ABSTRACT

The comparative genomics has allowed to trace the evolutionary history of all organisms and it has also given some indirect evidence on the traits of the last common ancestor (LCA) of Bacteria, Archaea, and Eukarya. However, the phylogenetic analysis based on sequence data to determine the origin and early evolution of viruses has been severely compromised by their highly divergent nature. On the other hand, the recent availability of more biological, genomic, structural, and ecological data of DNA and RNA viruses has provided the opportunity to infer and analyze, in detail, their evolutionary relationships among them and their hosts.

In a first approach to understand the origin of viruses, we compared and analyzed recent biological and ecological data, determined and characterized the pangenomic composition, and made a phylogenomic exploration of several viral families. While some researchers argue that viruses are the missing link between the non-living, the RNA world, and the first cells due to their morphological and genomic “simplicity”, our results reveal that the size distribution and chemical nature of the viral genome do not exhibit a correlation with the phylogeny of their hosts. We found that the supposedly “more complex” and longest viral genomes are found in phages, which infect only ancient domains of life (Bacteria and Archaea) and in giant viruses, as megaviruses, which infect ancient lineages of eukaryotes (protists). A rather significant majority of the RNA viruses infect only the Eukarya domain. No RNA viruses have been found in Archaea yet. There are only two RNA viral families in prokaryotes, but they infect Proteobacteria of animal microbiota. Our pangenomic analysis and phylogenetic trees have shown that the highly-conserved proteins (core genes) in megaviruses intervene in the most of DNA replication and repair processes, probably have either a bacteria, protist, fungi, plant, and/or an animal origin, and therefore, could be the most antique proteins than the rest of the pangenome. These same results have indicated that less-conserved proteins (shell genes) and unique proteins specific to single viral strains (cloud genes) intervene in some genetic, cellular, metabolic, and unknown functions, have mainly a viral and eukaryotic origin, and therefore, they could be the most recent proteins of the viral pangenome. These preliminary results might suggest that the evolutionary history of viruses is related to the phylogeny of their host cells, that is to say, DNA viruses could have a more antique origin that goes back to the LCA stage, while the origin of RNA viruses may be explained by a coevolutionary process with their eukaryotic hosts. These asseverations could confirm the hypothesis that viruses are escaping genes from cell genomes; and hence, viruses can be antique, but not primitive.

# I. INTRODUCCIÓN

## 1.1 Características generales de los virus

De acuerdo a los análisis metagenómicos, los virus son las entidades biológicas universales más abundantes de la biósfera con un estimado de  $10^{31}$  virus (Breitbart & Rohwer, 2005). La virósfera (Abroi & Gough, 2011) muy probablemente infecta a todos los tipos celulares de los tres dominios del árbol de la vida (Bacteria, Archaea y Eukarya) y tiene una influencia extraordinaria en procesos biogeoquímicos y geológicos (Edwards & Rohwer, 2005). Los virus presentan propiedades peculiares que los definen como agentes infecciosos intracelulares que dependen de la maquinaria enzimática para replicarse. Se caracterizan por tener una arquitectura muy simple que puede resguardar a un genoma de DNA o de RNA en una cubierta llamada cápside (icosaédrica, helicoidal o compleja) y que, en algunos casos, adicionalmente presenta una membrana lipídica. Las partículas virales de la progenie infecciosa, llamadas viriones, se forman por el autoensamblaje *de novo* a partir de los componentes recién sintetizados en la célula durante su ciclo infeccioso (Flint, Rall, Racaniello, & Skalka, 2015). Algunas características son equivalentes a las de las células como la posesión de genes, la capacidad de crear múltiples copias de sí mismos y la habilidad de evolucionar por selección natural (Gibbs & Calisher, 2005). Sin embargo, ellos no realizan autopoiesis (por sí mismos no pueden autoreplicarse porque no tienen ribosomas ni tienen un metabolismo propio) y no comparten un ancestro común. Esto ha dado pie a que haya una discusión sobre si los virus pueden incluirse en la definición de vida o simplemente son estructuras orgánicas que pueden interactuar con los seres vivos (Koonin & Starokadomskyy, 2016; Moreira & López-García, 2009). Hasta febrero de 2018, se cuenta con casi 7,500 genomas de referencia en el *GenBank*, un poco más de 2,600 proteomas y más de 16,700 proteínas virales de referencia en el *ViralZone* y se tiene la descripción de 121 familias, 142 géneros y 9 especies sin clasificar en el Comité Internacional de Taxonomía de Virus (ICTV, por sus siglas en inglés).

## 1.2 Características genómicas y genéticas de los virus

El sistema de clasificación que maneja el ICTV se basa en la naturaleza química de ácido nucleico del genoma viral (DNA o RNA, circular o lineal, de una o dos cadenas,

segmentado o no), en la arquitectura y dimensión de la cápside y en la presencia o ausencia de una membrana lipídica (Flint et al., 2015). El ICTV también hace uso de la clasificación de Baltimore (1971) que se basa en la manera en que los virus producen su RNA mensajero que será traducido por los ribosomas de sus hospederos. Así, existen siete tipos de genomas para todas las familias virales: de DNA de doble cadena (dsDNA) y de cadena sencilla (ssDNA); de RNA de doble cadena (dsRNA), de cadena sencilla positiva [(+)ssRNA] y de cadena sencilla negativa [(-)ssRNA] y retrovirus de RNA (ssRNA-RT) y de DNA (dsDNA-RT) (Baltimore, 1971).

El tamaño del genoma varía enormemente entre los grupos virales. El genoma de mayor longitud es el de los Pandoravirus (dsDNA) con 2,500 kilopares de bases (kbp) (Philippe et al., 2013), y el de menor longitud es el de los circovirus de ssDNA con apenas 1 kbp (Belyi, Levine, & Skalka, 2010). Los genomas de los virus de RNA en promedio son más pequeños que los de DNA con límite de 35 kpb (Campillo-Balderas, Lazcano, & Becerra, 2015). A diferencia de los virus de dsDNA, los virus de RNA y de ssDNA se caracterizan por tener una tasa alta de mutación debido a que carecen de mecanismos de corrección, generar un gran número de individuos por progenie, presentar una duplicación génica y una transferencia horizontal de genes muy bajas, mantener niveles de recombinación genética relativamente poco frecuentes, tener un solapamiento de genes muy común y poseer genomas segmentados (Brandes & Linial, 2016; Duffy & Holmes, 2009; Holmes, 2009; Sanjuán, Nebot, Chirico, Mansky, & Belshaw, 2010).

La replicación del genoma de algunos virus de DNA generalmente tiene lugar en el núcleo de la célula y dependen de los mecanismos de procesamiento de DNA y de RNA de la célula, mientras que la replicación de los megavirus se lleva a cabo en el citoplasma y generalmente codifican algunas proteínas que intervienen en la maquinaria de replicación, transcripción y síntesis de proteínas. La replicación del genoma de los virus de RNA se presenta en el citoplasma con diferentes modos de replicación (RNA con polaridad positiva, negativa o ambas) y en la presencia de una polimerasa de RNA dependiente de RNA (RdRp). Finalmente, la replicación del genoma de los retrovirus se da con la presencia de una transcriptasa reversa (RT) para producir un RNA intermediario (virus de dsDNA-RT) o un DNA intermediario (virus de ssRNA-RT) que después es integrado al genoma celular recibiendo el nombre de provirus.

### **1.3 Características moleculares y ecológicas de los virus en relación con sus hospederos**

Los virus tienen distintos tipos de interacción con sus respectivos hospederos desde un nivel bioquímico y celular hasta un nivel ecológico para poder infectar, replicarse y diseminarse. A nivel molecular y celular, existen determinantes específicos que permiten una exitosa infección viral como puede ser la unión a la célula hospedera mediada por proteínas virales unidas a receptores celulares tales como proteínas membranales, lípidos, carbohidratos, glicoproteínas, polisacáridos, glicoesfingolípidos y lipopolisacáridos, entre otros (Grove & Marsh, 2011; Rakhuba, Kolomiets, Dey, & Novik, 2010). Dicha unión determina la entrada del genoma viral a la célula a través de su captación y tráfico intracelular y, en última instancia, la penetración al citosol. Esto ocasionará que, en algunos casos, como los virus que infectan vertebrados, puedan tener efectos citopáticos (habilidad para matar células a través de proteínas virales citotóxicas, inhibición de la síntesis de proteínas celulares, alteración del metabolismo celular o apoptosis y lisis celular) (Grove & Marsh, 2011; Wagner, 1984). En otros casos, como algunos virus que infectan vertebrados y algunos que infectan procariontes, pueden recurrir a un estado de latencia ya sea a través de la integración de su genoma al de la célula o de la formación de un episoma (Fortier & Sekulovic, 2013; Grinde, 2013; McDonnell, Sparger, & Murphy, 2013). Sin embargo, los hospederos celulares también responderán antagónicamente ante una infección viral a través de las vías de señalización de su sistema de defensa. Desde el sistema CRISPR-Cas y restricción-modificación en procariontes (Barrangou, 2015; Murray, 2002), mecanismos del silenciamiento del RNA en plantas e invertebrados (Csorba, Pantaleo, & Burgyán, 2009), hasta la modulación del sistema inmunológico innato y adaptativo en vertebrados (Flint et al., 2015).

A nivel ecológico se puede apreciar la gran diversidad de ecosistemas en los cuales los virus pueden subsistir, pero sin dejar de considerar su especificidad a ciertos hospederos. Desde la atmósfera (Reche, D'Orta, Mladenov, Winget, & Suttle, 2018), los desiertos (Zablocki, Adriaenssens, & Cowan, 2015), los océanos (Suttle, 2005) y los suelos (Williamson, Radosevich, & Wommack, 2005), donde abundan los virus de dsDNA que infectan a cepas únicas de bacterias y arqueas; pasando por los ecosistemas vegetales (Roossinck, 2012), donde están presentes principalmente virus de RNA y de ssDNA que infectan a plantas e insectos; hasta la microbiota y diversos tejidos de animales donde se encuentran virus de DNA

y RNA (Andrewes, 1963; Cadwell, 2015; Ryabov, 2017). Las relaciones simbióticas entre estos holobiontes abarcan diversos estilos de vida en los que se incluye una relación parasítica (el virus se beneficia a expensas del hospedero), comensal (el virus se beneficia sin afectar al hospedero) o mutualista (ambos se benefician uno del otro) (Roossinck & Bazán, 2017). La interacción íntima entre los virus y sus hospederos son las que han puesto a los virus entre las ramas del árbol de la vida como participantes activos en la evolución de los seres vivos.

## 1.4 Características evolutivas

Como se dijo anteriormente, los virus pueden ser considerados oportunistas, pero también son agentes indispensables en la evolución de sus hospederos a través de la transferencia horizontal de genes. Una elevada tasa de mutación [de  $1.5 \times 10^{-3}$  mutaciones por nucleótido por replicación ( $m/n/r$ ) en virus de (+)ssRNA (J. W. Drake, 1993) a  $1.8 \times 10^{-8}$   $m/n/r$  en virus de dsDNA (John W. Drake & Hwang, 2005)] combinada con procesos de selección natural, deriva génica, epistasis, recombinación y rearrreglo genético han permitido que los virus puedan adaptarse a los cambios que naturalmente sufre el hospedero (Flint et al., 2015; Gibbs & Calisher, 2005; Holmes, 2009). Desde un enfoque evolutivo, los virus en realidad son poblaciones que están en un equilibrio dinámico de replicones similares entre sí llamados cuasiespecies (Eigen, McCaskill, & Schuster, 1988). Éstas se caracterizan por presentar una gama diversa de genotipos y fenotipos que les permiten sobrevivir a eventos de selección y heredar las mutaciones seleccionadas a la progenie. Sin embargo, a pesar de la alta tasa de mutación entre dichas variantes, existen elementos genéticos de tipo cis y trans que se conservan y que intervienen en procesos de replicación y empaquetamiento del genoma y síntesis del RNA mensajero (Flint et al., 2015). Estas secuencias son bastante estables para ser utilizadas como marcadores filogenéticos. Se ha reportado que estos fósiles moleculares conservados, como la DNA y la RNA polimerasa viral, pueden trazar las relaciones filogenéticas profundas sobre su propio origen (Černý, Černá Bolfíková, de A Zanotto, Grubhoffer, & Růžek, 2015; Jácome, Becerra, Ponce de León, & Lazcano, 2015), aunque también pueden servir como marcadores filogenéticos para explicar el origen mismo y la evolución temprana de cada una de las familias virales.

### 1.4.1 Origen de los virus

La genómica comparada ha permitido trazar la historia evolutiva de todos los organismos e inferir la existencia hipotética del LCA a través de marcadores filogenéticos como los genes ribosomales y proteínas que intervienen en el procesamiento del RNA, transcripción y traducción (Becerra, Delaye, Islas, & Lazcano, 2007; Doolittle, 2000). En contraste, aún existe una fuerte discusión y un claro desconocimiento sobre el origen y evolución temprana de los virus debido a la dificultad para inferir sus relaciones filogenéticas dada su naturaleza divergente (origen polifilético) y la inexistencia de su registro fósil (Holmes, 2009). A pesar de estos problemas, se tienen cinco hipótesis sobre el origen de los virus que no necesariamente son independientes o mutuamente excluyentes. La mayoría de estas hipótesis se basa en el tamaño y naturaleza química del genoma y la conservación de algunos marcadores filogenéticos de familias virales. Con base en la bibliografía examinada, podemos clasificar a las hipótesis sobre el origen de los virus de acuerdo a si éstos aparecieron antes, durante o después de las primeras entidades celulares.

#### 1.4.1.1 Hipótesis del virocentrismo (origen precelular)

De acuerdo a algunos autores, los virus son entidades biológicas que se originaron en un periodo pre-celular y, subsecuentemente, proporcionaron la materia prima para el origen de las primeras células. Los virus de RNA son los descendientes directos del Mundo del RNA.

Para 1917, Felix D'Herelle y Frederick Twort ya habían descubierto a los virus a los que designaron como "formas de vida primordiales en el origen de la vida" (d'Herelle & Smith, 1926). Ellos partían del siguiente silogismo: los virus son pequeños y si son pequeños son simples y si todo indica que los primeros organismos debieron ser simples, por lo tanto los virus debieron haber surgido primero (Beutner, 1938; Podolsky, 1996). Actualmente, esta hipótesis no tan solo se basa en la naturaleza y el tamaño de los genomas virales, sino también, en la supuesta existencia de genes distintivos (*hallmark genes*) de origen viral, es decir, no tienen homólogos celulares como son las proteínas de cápside con dominio *jelly-roll*, la helicasa de la superfamilia 3, la DNA primasa, la ATPasa, la transcriptasa reversa, la RNA polimerasa dependiente de RNA, entre otras (Koonin & Dolja, 2006; Koonin, Senkevich, & Dolja, 2006). Koonin y otros investigadores han reformado esta hipótesis y han propuesto que el origen de los virus se llevó a cabo en diferentes etapas primordiales de la vida: Mundo del RNA [viroides



(Flores, Gago-Zachert, Serra, Sanjuán, & Elena, 2014)], Mundo del RNA/proteínas [virus de (+)RNA y de dsRNA)], Mundo de RNA-DNA (virus parecidos a los retrovirus), Mundo del DNA (virus de dsDNA) y, finalmente, una etapa post-celular temprana (fagos) y tardía (virus que infectan a eucariontes) (Koonin et al., 2006).

#### 1.4.1.2 Hipótesis de la regresión celular (origen post-celular temprano)

Otros autores sostienen que los virus se originaron por evolución regresiva de microorganismos a través de la pérdida de genes y, por lo tanto, se han convertido en parásitos intracelulares obligados en la actualidad.

En 1935, los virólogos Robert Green y Sir Patrick Laidlaw sostuvieron que los virus se originaron a partir de células pequeñas que fueron perdiendo genes y, por lo tanto, algunas funciones a través del tiempo. Ellos consideraban como evidencia a las *Rickettsia* y las *Chlamydia* que son parásitos intracelulares obligados de eucariontes (Podolsky, 1996). Actualmente, existen varios científicos que apoyan esta versión sobre el origen de los virus. Uno de ellos, Patrick Forterre (2006) menciona que los virus se originaron en un mundo de RNA-proteínas cuando las ribocélulas, con ribosomas primitivos, ya habían emergido. Estas células primordiales perdieron su maquinaria de traducción hasta convertirse en parásitos obligados. Otros científicos, como los grupos de Didier Raoult y de Jean Claverie, sostienen que los megavirus son el resultado de la pérdida de genes de un cuarto dominio de la vida (Boyer et al., 2009; Colson, de Lamballerie, Fournous, & Raoult, 2012; Colson, Gimenez, Boyer, Fournous, & Raoult, 2011). Estos virus pueden ser tan grandes como los de las bacterias más pequeñas y codificar genes para DNA polimerasas, helicasas y ribonucleótido reductasas, factor de transcripción eIF4E, aminoácido-tRNA ligasas, tRNA-aminoacil transferasas, y enzimas modificadores de tRNAs, todas ellas importantes en la replicación, transcripción y traducción del genoma (Philippe et al., 2013). Por otro lado, Nasir & Caetano-Anollés (2015) y Philippe et al (2013) también proponen que los virus modernos redujeron su genoma a partir de múltiples linajes celulares ancestrales que tenían RNA genómico y que coexistían con las células modernas.

#### 1.4.1.3 Hipótesis del escape (origen post-celular tardío)

Algunos autores aseveran que los virus son partes de genomas escapados a partir de entidades celulares de RNA o de DNA para convertirse en entidades replicativas autónomas.

Para 1944, Frank MacFarlane Burnet propuso que los virus son “fragmentos errantes de material genético de origen celular” (Antonio Lazcano, 2010). Actualmente esta hipótesis ha sido retomada por algunos virólogos como Patrick Forterre (2006) (el mismo que apoya la hipótesis de la regresión celular) quien propone que algunas moléculas de RNA se escaparon de ribocélulas. Estos virus de RNA tuvieron un origen anterior al LCA y la prueba, según él, es que no se han encontrado proteínas homólogas entre los virus y los descendientes de este ancestro de los seres vivos. Además, Forterre subraya que dada la naturaleza simple de los mecanismos de transcripción/traducción de las ribocélulas, era más fácil que algunos elementos genéticos se volvieran autónomos. Aparentemente, los RNA mensajeros de las células pudieron ser independientes gracias a su capacidad de autoreplicación y de protección por una cápside. Esta misma hipótesis sostiene que tanto los virus de RNA como los de DNA surgieron después del origen celular (Holmes, 2009).

#### 1.4.1.4 Hipótesis de la coevolutiva a largo plazo (origen simultáneo o precelular)

Otros autores mencionan que los virus tienen un origen inmediatamente anterior o simultáneo al de las células y que ambas entidades comparten módulos funcionales.

De acuerdo al virólogo Esteban Domingo (2015); a principios de este siglo, los estudios independientes de Bushman (2002), Mount (2004) y Hacker & Dobrindt (2006) sostienen dicha hipótesis gracias a la información genómica masiva con la cual se han identificado secuencias regulatorias y codificantes exclusivas de los virus (sin homólogos celulares). Existen dos módulos de proteínas exclusivamente virales: las propias (*self*) de la especie viral que son innatas y conservadas como las de la cápside y ATPasas de empaquetamiento del genoma y las no propias (*non-self*) que provienen de otros virus por transferencia horizontal como las de la replicación del genoma y las de lisis celular (Krupovič & Bamford, 2007). Ambos módulos son funcionalmente esenciales y se comparten entre todos los virus. Es el segundo módulo el que ha contribuido a la coevolución de células y de estos replicones autónomos a través de transferencias horizontales (Domingo, 2015).

#### 1.4.1.5 Hipótesis de las vesículas (origen simultáneo y posterior)

Unos autores comparten la idea de que los ancestros virales, “protovirus”, se originaron en vesículas primitivas en coevolución con “protocélulas” del Mundo del RNA.

Jalasvuori & Bamford (2008) mencionan que la mayoría de las vesículas formadas abióticamente y que contenían moléculas autoreplicativas pudieron haber sido seleccionadas

positivamente. Estos protovirus coexistían y dispersaban genes horizontalmente a vesículas mayores llamadas protocélulas. Las protocélulas sobrevivieron al Mundo del RNA gracias a la retroalimentación continua con estos protovirus y comenzaron a ser más independientes. En una etapa posterior, estos protovirus continuaron coevolucionando con protocélulas favoreciendo su selección a través de la expresión de peptidoglicano de la pared celular y de receptores membranales para la formación de “células verdaderas”. Las células se volvieron completamente autónomas mientras que los virus solo aprovechaban los recursos enzimáticos de éstas. Así, cuando emergió la población que hoy denominamos LCA, éstas células originaron mecanismos de defensa promoviendo la emergencia de los virus modernos.

## **1.4.2 Estrategias metodológicas para abordar el problema sobre su origen**

Un árbol filogenético es una representación gráfica sobre las relaciones evolutivas entre los taxa y se construye a partir de secuencias homólogas (ortólogos) (Fitch, 2000). El estudio del origen de la virósfera, como un todo, se complica porque no parece existir un marcador filogenético universal entre los linajes virales lo que demuestra, en principio, su origen polifilético. Sin embargo, existen algunos genes que se comparten en cada una de las familias virales y que podrían trazar las relaciones evolutivas monofiléticas en cada linaje. Es por ello que es importante identificar la estrategia metodológica correcta para la construcción de árboles filogenéticos de secuencias virales. En general, existen dos métodos para su construcción, uno basado en secuencias primarias de proteínas y, el otro, en estructuras virales.

### **1.4.2.1 Estrategia basada en secuencia primaria de proteínas**

Estos métodos se basan en la información que puede otorgar un alineamiento de secuencias primarias de proteínas homólogas para la construcción de árboles filogenéticos. Las regiones alineadas son de interés porque reflejan su importancia evolutiva y estructural, mientras que los espacios (*gaps*) en el alineamiento representan eventos de inserción o deleción (Lam, Hon, & Tang, 2010; McCormack & Clewley, 2002; Romero, 2004). A través de estas estrategias se analizaron la DNA polimerasa y la replicasa viral como marcadores filogenéticos. Se demostró que la historia evolutiva de la DNA polimerasa es muy compleja porque existe una clara evidencia sobre la transferencia horizontal y el desplazamiento de

genes no ortólogos entre virus, células y plásmidos (Filée, Forterre, Sen-Lin, & Laurent, 2002; Le Gall et al., 2008). También se evidenció que la RNA polimerasa dependiente de RNA presenta regiones muy conservadas entre los virus de RNA como el motivo C (Gly-Asp-Asp) (Gorbalenya et al., 2002) localizado en el subdominio palma y que, al mismo tiempo, con este análisis filogenético y el de otros marcadores, se pudieron clasificar varias familias virales de (+)ssRNA en un orden más alto: los picornavirales (Le Gall et al., 2008).

#### 1.4.2.2 Estrategia basada en organización del genoma

Para incrementar la robustez de la reconstrucción filogenética se pueden utilizar estrategias basadas en el análisis de múltiples genes o, si es el caso, en el de genomas completos de manera simultánea (Rokas, Williams, King, & Carroll, 2003). Estos estudios son útiles en mayor medida para el análisis filogenético de virus de dsDNA, ya que se cuenta con suficientes patrones genómicos como los Poxviridae (McLysaght, Baldi, & Gaut, 2003). Por otro lado, los virus de RNA tienen genomas pequeños con no más de 10 a 12 genes por lo que presentan un número menor de caracteres genéticamente informativos y, además, tienen poca resolución filogenética debido a la organización genómica variable (Holmes, 2009).

#### 1.4.2.3 Estrategia basada en estructura terciaria

Estos métodos se basan en la comparación de las estructuras tridimensionales de las proteínas, debido a que éstas presentan un grado más alto de conservación y, por lo tanto, proporcionan más información sobre su historia evolutiva que la variabilidad y dinámica de una secuencia primaria de aminoácidos (Chothia, 2003; Gerstein & Hegyi, 1998). Es debido a ello que los dominios de proteínas son considerados como unidades evolutivas (Murzin, Brenner, Hubbard, & Chothia, 1995; Riley & Labedan, 1997; Wang, Yafremava, Caetano-Anollés, Mittenthal, & Caetano-Anollés, 2007) y útiles como caracteres filogenéticos para analizar relaciones evolutivas profundas (Abroi & Gough, 2011) como es el caso de la RNA polimerasa dependiente de RNA (RdRp), cuyo subdominio palma es estructuralmente homólogo al de las DNA polimerasas celulares lo que ha apoyado a la idea de que es una de las regiones más antiguas presentes en células y en virus (Jácome et al., 2015).

### 1.4.3 Estudios pangenómicos

Hasta esta parte introductoria de la tesis, se ha mostrado que para determinar la historia evolutiva de los virus se apela al análisis filogenético de marcadores altamente conservados con las diferentes estrategias anteriormente mencionadas tales como la DNA y RNA pol, ATPasa, ribonucleótido reductasa, timidilato sintasa, helicasas, tRNA sintetasa; RdRp, RT; proteínas de cápside, entre otras. Sin embargo, es una imagen parcial evolutiva, ya que existen otras proteínas que, si bien no están compartidas entre todos los individuos de un grupo viral, forman parte de la filogenia completa y permiten comprender los procesos que generan la diversidad genética y la variación fenotípica de un clado. ¿Cuántos genomas se necesitan para definir filogenéticamente a una familia viral y de esta manera complementar el estudio de su origen y evolución temprana? Para ello es importante considerar los estudios pangenómicos que se han hecho en microorganismos y en plantas y que han ayudado a determinar en parte su dinámica evolutiva (Contreras-Moreira et al., 2017; Kaas, Friis, Ussery, & Aarestrup, 2012).

Un pangenoma se define como el repertorio genético de todos los individuos de un clado (Vernikos, Medini, Riley, & Hervé, 2015). Es decir, un pangenoma incluye 1) a todos los genes altamente conservados y que se encuentran distribuidos en todas las especies del clado (*core*, núcleo pangenómico), 2) a los genes que se conservan en algunas especies de ese clado, pero que son funcionalmente indispensables (*shell*, cubierta pangenómica) y 3) a aquellos genes que son únicos y específicos de una sola especie (*cloud*, nube pangenómica) (Medini, Donati, Tettelin, Massignani, & Rappuoli, 2005).

Han sido solo tres estudios realizados sobre pangenómica viral desde el 2013 y sólo se han hecho en virus de dsDNA. En un trabajo sobre pangenómica de fagos se determinó que su grupo de genes ortólogos continúa creciendo y que existen muchos genes únicos sin homólogos procariontes (Kristensen et al., 2013). En un estudio sobre pangenómica de un baculovirus (que infecta a insectos) se reportó que el 90% de los genes del núcleo genómico son hipotéticos y que existen muchas alteraciones fenotípicas por pérdida o ganancia de genes y sustituciones de nucleótidos (Brito et al., 2015). En otra investigación sobre pangenómica de un clado de mimivirus de Brasil (que infecta a amebas) se encontró que estos virus son pangenómicamente similares, muy probablemente debido a que se distribuyen en la misma zona geográfica (Assis et al., 2015).

## 1.5 El último ancestro común de los seres vivos y los virus

El último ancestro común (LCA) de todos los seres vivos es una población de organismos hipotética reciente de la cual todos los seres vivos (Bacteria, Archaea y Eukarya) descendieron. Su existencia se infiere a partir de análisis filogenéticos basados en secuencias de RNA ribosomal (Woese & Fox, 1977). La consistencia de los análisis filogenéticos ha caracterizado al LCA como un conjunto de organismos unicelulares parecidos a las bacterias y con un código genético basado en DNA como el actual (Becerra et al., 2007). Además, el LCA también se caracteriza por tener un repertorio genético de secuencias universalmente conservadas que intervienen en procesos de replicación y reparación del DNA; traducción y transcripción; procesamiento del RNA; síntesis de nucleótidos, aminoácidos y azúcares, y producción de energía mediada por ATPasas membranales (Becerra et al., 2007).

Estas características describen al LCA como un conjunto de individuos totalmente autónomos con la capacidad de replicarse, automantenerse y evolucionar. Sin embargo, éstas son las mismas características que excluyen a los virus de las ramas del árbol de la vida. Por un lado, a los virus no se les considera organismos vivos debido a que dependen totalmente de la maquinaria enzimática celular para replicarse y evolucionar. Además, tienen un origen polifilético, no tienen linajes ancestrales y todo indica que la mayoría de sus genes informacionales y metabólicos se originaron en genomas celulares (Moreira & López-García, 2009).

## 1.6 Los megavirus y el LCA

Existe un grupo de virus que se caracteriza por su gran tamaño genómico (hasta 2.5 millones de pares de bases) (Philippe et al., 2013) y morfológico (una cápside de hasta 1500 x 500 nm) (Legendre et al., 2014). A este hipotético clado supuestamente monofilético se le ha conocido como virus nucleocitoplásmicos de DNA de gran tamaño (NCLDV o megavirus) e incluye a siete familias: Ascoviridae, Asfarviridae, Iridoviridae, Marseilleviridae, Mimiviridae, Phycodnaviridae y Poxviridae (Lakshminarayan M. Iyer, Balaji, Koonin, & Aravind, 2006). Una de las características peculiares de los megavirus que los distingue del resto de la virosfera es que tienen genes involucrados en la replicación y reparación del DNA, transcripción y traducción como lo son la DNA polimerasa de la familia B, la topoisomerasa II A, la

endonucleasa FLAP, el antígeno nuclear de células en proliferación (PCNA), RNA polimerasa dependiente de DNA tipo II, y el factor de transcripción II B y varias tRNA sintetasas (L. M. Iyer, Aravind, & Koonin, 2001; Yutin, Wolf, Raoult, & Koonin, 2009). Aparentemente estos genes conservados indican que los megavirus tienen un ancestro común con dicho repertorio complejo (Koonin & Yutin, 2010).

La conclusión, a partir de las premisas anteriores, por parte de algunos grupos de Virología, es que este grupo viral desafió a la definición de vida y que, por lo tanto, debería ser considerado como una rama más, un cuarto dominio, derivado del LCA (Boyer, Madoui, Gimenez, La Scola, & Raoult, 2010; Nasir, Kim, & Caetano-Anolles, 2012; Didier Raoult & Forterre, 2008; Wu et al., 2011). Sin embargo, otros han reportado que exclusivamente esas secuencias altamente conservadas tienen un origen eucarionte y que, por lo tanto, no hay evidencia que sostenga la idea de otro dominio de la vida (Schulz et al., 2017; Yutin, Wolf, & Koonin, 2014). Aparentemente, el origen de los megavirus, de acuerdo a algunos autores, los ancestros de los megavirus provienen de los “polintovirus” (transposones de DNA de eucariontes capaces de formar viriones) que, a su vez, éstos evolucionaron de fagos (Koonin, Krupovic, & Yutin, 2015).

Es por ello que en la presente tesis, se pretende dilucidar el origen y evolución temprana de los virus a través de 1) un análisis general sobre el tamaño de su genoma y la distribución taxonómica en sus hospederos procariontes y eucariontes; 2) un estudio pangénómico y filogenético del repertorio proteico del núcleo, cubierta y nube de los megavirus y su relación con el último ancestro común de los seres vivos; y 3) un análisis evolutivo muy preliminar a través de la comparación de las estructuras cristalográficas de las polimerasas de RNA de virus de RNA. Cabe destacar que la idea global de esta línea de investigación es realizar un análisis pangénómico y filogenético de secuencias y estructuras terciarias para cada una de las más de 100 familias de virus de RNA y de DNA. A través de la consiliencia y discordancia de todos los análisis de datos biológicos, ecológicos, pangénómicos y filogenéticos generados para cada una de las familias virales, se espera apoyar a alguna o algunas de las hipótesis sobre el origen y evolución temprana de los virus, es decir, si éstos tienen un origen primordial, si tienen un ancestro común celular, o si son fragmentos escapados de genomas celulares.

## II. MATERIALES Y MÉTODOS

### 2.1 Construcción de la base de datos con información biológica y ecológica de los virus

Para agrupar toda la información biológica y ecológica de los virus se construyó una base de datos a partir de los registros del GenBank (<https://www.ncbi.nlm.nih.gov/genome/viruses/>), del 9o Reporte del Comité Internacional de Taxonomía de Virus (ICTV, por sus siglas en inglés) (King, Adams, & Lefkowitz, 2011), del ViralZone (<http://viralzone.expasy.org/>) y de publicaciones relevantes al mes de diciembre de 2014. Dicha información se clasificó de acuerdo al tipo de especies, tipo y tamaño del genoma, segmentación y tipo de hospedero de las más de 100 familias de virus de RNA y de DNA. Se recopilaron datos biológicos y ecológicos de 4183 especies virales de referencia, así como de 215 virus satélite y 44 viroides. De acuerdo a la clasificación de Baltimore, se obtuvieron registros de 1926 virus de dsDNA; 701, de ssDNA; 205, de dsRNA; 966, de ssRNA(+); 253, de ssRNA(-); 70, de dsDNA-RT; y 62, de ssRNA-RT. De acuerdo a la clasificación por hospedero, se obtuvieron registros de 1438 virus que infectan a Bacteria; 69, a Archaea; 74, a protistas; 1273, a plantas; 82, a hongos; 58, a plantas e invertebrados; 260, a invertebrados; 123, a invertebrados y vertebrados; y, finalmente, 1064 virus que infectan exclusivamente a vertebrados. De acuerdo a su tipo de genoma, se encontró que existen registros de 1485 virus que se clasifican en las 55 familias virales de RNA y 2697 virus, en 43 familias de DNA. De acuerdo a su nivel de segmentación, se obtuvieron registros de 3682 virus que tienen un solo segmento y solo 501 que tienen más de dos o más segmentos. Aquellos virus que no tenían un hospedero identificado en el *GenBank* (n=31) se excluyeron. La base de datos se puede verificar en este link: <https://www.frontiersin.org/articles/10.3389/fevo.2015.00143/full#h8>.

### 2.2 Análisis de los datos biológicos y ecológicos de los virus

Para determinar la distribución de los virus de acuerdo a la antigüedad de los dominios en que se encuentran clasificados los hospederos, se agruparon de manera distinta las diferentes características biológicas de éstos tal como su composición química, tamaño y segmentación del genoma. Por un lado, el promedio del tamaño del genoma de los virus



agrupados de acuerdo a la Clasificación de Baltimore, al tipo de hospedero y a la segmentación fue calculado. Para esto, las gráficas sobre el tamaño del genoma viral fueron hechas logarítmicamente con base 10. Asimismo, el porcentaje de las familias de virus de RNA y de DNA por cada hospedero fue estimado. Para ello, cada familia viral fue contada doble si ella infecta a más de un hospedero por lo cual se estimó que el 15 familias virales infectan al Dominio Bacteria [Proteobacteria ( $n = 8$  familias), otras phyla ( $n = 7$ )], 13 familias infectan al Dominio Archaea [Crenarchaeota ( $n = 9$ ) y Euryarchaeota ( $n = 4$ )], y 83 familias infectan al Dominio Eukarya [protistas y algas ( $n = 7$ ), plantas ( $n = 21$ ), hongos ( $n = 15$ ), y animales ( $n = 50$ )]. Dicha distribución de las familias virales en los tres dominios fue utilizada para adornar la filogenia preestablecida (con algunas modificaciones para este estudio) en la plataforma interactiva del árbol de la vida (IToL, por sus siglas en inglés) (Letunic & Bork, 2016).

### 2.3 Construcción de la base de datos pangenómica de los virus

Para agrupar las proteínas virales de cada familia de acuerdo a su pangenoma, se construyeron dos bases de datos: primero una proteómica y, después, una pangenómica.

La base de datos proteómica viral fue construida a partir de los proteomas de referencia (sin redundancia, completos, con secuencias codificantes descritas y validadas) por cada una de las 98 familias virales en el *GenBank* (<https://www.ncbi.nlm.nih.gov/genome/viruses/> en junio de 2016). Nosotros consideramos que la familia representa una unidad evolutiva puesto que contiene un conjunto de especies virales que comparten un ancestro común de acuerdo al ICTV. Para descargar los archivos de dichos proteomas (en formato *GenBank* que contiene toda la información de la anotación y la secuencia), se utilizó la siguiente fórmula booleana (utilizando como ejemplo a la familia Mimiviridae) en el buscador de la base de datos de nucleótidos del Centro Nacional para la Información Biotecnológica (NCBI, por sus siglas en inglés):

**Mimiviridae[Organism] AND srcdb\_refseq[PROP] NOT wgs[prop] NOT cellular organisms[ORGN] NOT AC\_000001:AC\_999999[pacc]**

Una vez descargados, los archivos de los proteomas de las especies virales se agruparon manualmente por carpetas, las cuales, representaban a cada una de las 98 familias. Con la ayuda de un *script* en *Perl*, cada proteoma viral de una sola carpeta (familia) se extrajo en formato FASTA y se guardó en un archivo de texto independiente. Se contaron todas las

proteínas de cada proteoma viral y si un proteoma se alejaba del promedio del número de proteínas ( $\pm 35\%$ ), se descartaba de la muestra para los siguientes análisis. Si bien se tienen todos los proteomas de referencia de cada una de las 121 familias actualmente (mayo de 2018), sólo se utilizaron las siete familias de megavirus (Asfarviridae, Iridoviridae, Ascoviridae, Poxviridae, Phycodnaviridae, Marseilleviridae y Mimiviridae) para este estudio debido a su implicación actual sobre el origen de los virus y el último ancestro común de los seres vivos.

La base de datos pangenómica de los megavirus fue construida usando el *software GET\_HOMOLOGUES* (Contreras-Moreira & Vinuesa, 2013). Para este estudio nosotros adecuamos los conceptos pangenómicos (Medini et al., 2005) en función de la naturaleza divergente de los virus, es decir, al núcleo lo definimos empíricamente como un repertorio de proteínas homólogas presentes en al menos un 95% de todas las especies de una misma familia viral (*core + softcore*); a la cubierta (*shell*), como un conjunto de proteínas homólogas accesorias presentes en más de dos especies de una misma familia viral; y a la nube (*cloud*), como un grupo de proteínas específicas presentes en al menos dos especies o menos de una familia viral. Para la agrupación (*clustering*) de todas las proteínas ortólogas de todos los proteomas de cada familia viral, se utilizó al proteoma más pequeño como de referencia (*query*) y se realizó una búsqueda pareada con la combinación de los algoritmos *BLASTP* (Altschul, Gish, Miller, Myers, & Lipman, 1990), *Hmmer3* (Sean R. Eddy, 2009) y *COGtriangles* (Kristensen et al., 2010) con una cobertura (C) del alineamiento del 75% y un valor esperado (E)  $<10E-05$ . Dichos ortólogos virales se buscaron en Pfam (Finn et al., 2008) (actualización del 2015) para poder identificar sus correspondientes dominios conservados.

## 2.4 Análisis pangenómicos de los megavirus

Para estimar el tamaño de la composición pangenómica (núcleo, cubierta y nube), los datos se ajustaron al modelo de crecimiento exponencial de Tettelin (Tettelin et al., 2005). Una vez generada la matriz con las presencias y ausencias de las proteínas en el pangenoma de cada familia viral (y siempre y cuando la muestra sea mayor a tres proteomas), se estiman, se extraen y se grafican los diferentes compartimentos de dicho pangenoma a través del lenguaje *R* (<https://www.r-project.org>) con la función *Circle* que el mismo programa *GET\_HOMOLOGUES* tiene. Para visualizar mejor los datos de presencia y ausencia de la matriz, el archivo se transformó a un formato de tabla (.csv).

Cabe aclarar que los proteomas de las otras 91 familias virales están en una fase preliminar de su análisis pangénomico. Hasta ahora (mayo de 2018), ya se cuenta con la matriz de los compartimentos pangénomicos para todas las familias virales.

## 2.5 Clasificación funcional de los grupos de homólogos del pangenoma de megavirus

Para complementar la información de la matriz con los grupos de ortólogos virales de cada núcleo, cubierta y nube, se utilizó no tan solo el reporte generado de cada uno de los dominios *Pfam*, sino también la información que se encuentra en el *GenBank*, *Uniprot*, *Gene Ontology*, *KEGG*, *SMART*, *PDB* y *PROSITE*. Para esto, se utilizaron los identificadores del *GenBank* y del *Pfam*, una vez determinados por *GET\_HOMOLOGUES* y se pegaron en forma de columna cada uno por separado en las plataformas de la Red de Base de Datos Biológicas (*bioDBnet*, <https://biodbnet-abcc.ncifcrf.gov/>) y de *Uniprot* (<http://www.uniprot.org/uploadlists/>). En estas plataformas se puede seleccionar la base de datos deseada para completar la información biológica como la función genética, celular y metabólica; la estructura terciaria, si es el caso; entre otras. Esta información complementaria se anexó a la tabla .csv. Asimismo, para organizar cada una de las funciones de cada grupo de ortólogos en los compartimentos pangénomicos, se recurrió a la clasificación de los grupos de ortólogos (COG, <https://ftp.ncbi.nih.gov/pub/wolf/COGs/COG0303/fun.txt>) del NCBI (Tatusov, 1997). Se identificaron las funciones a mano (a través de la plataforma de *Pfam* y de la literatura científica) de cada uno de los grupos de ortólogos que tenían un identificador *Pfam* o *GenBank*. Aquellos grupos de ortólogos sin identificadores se clasificaron como grupos con función desconocida. De la misma plataforma de *Pfam* (<https://pfam.xfam.org/search#tabview=tab1>), se extrajo la información de la distribución de cada uno de los dominios en la sección de especies (*Species*) y también se anexó a la tabla .csv. Para determinar las funciones específicas o generales de cada uno de los grupos de ortólogos, se realizaron distintas combinaciones entre los tres compartimentos (núcleo, cubierta y nube), su distribución entre los dominios (B, Bacteria; A, Archaea; E, Eukarya) y otros grupos virales (V), su función específica o general de los COG (Tabla 1).

Debido a que los valores generados por estas combinaciones son diferentes entre sí en varios órdenes de magnitud, se normalizaron logarítmicamente en el lenguaje R a través de la fórmula:

$$y_i = \log(x_i)$$

Donde  $y_i$  es la variable que representa a la versión de  $x_i$  transformada en el logaritmo de base 10 y  $x_i$  representa a los valores de la cuantificación de cada COG por compartimento pangenómico. Una vez hechas las transformaciones logarítmicas, se utilizó la función de *pheatmap* en *R* (después de instalar las bibliotecas *gplots*, *DT* y *RColorBrewer*) para visualizar la matriz.

Asimismo se utilizaron valores absolutos para resaltar los grupos de ortólogos de la nube que tienen una función desconocida. Dichos valores se visualizaron con *Circos* (Krzywinski et al., 2009) en la terminal usando solo los parámetros para ordenar los datos por columnas y filas y con sus respectivos colores (*col with row order*, *row with col order*, *col with row color*, *row with col color*, *hide relative tick marks*).

**Tabla 1. Código de letras y funciones de los Grupos de Ortólogos (COGs)**

<b>Información genética</b>	A	Procesamiento y modificación del RNA
	B	Estructura y dinámica de la cromatina
	J	Traducción, estructura y síntesis del ribosoma
	K	Transcripción
	L	Replicación, recombinación y reparación del DNA
<b>Procesos celulares</b>	D	Control del ciclo celular, división celular y división cromosómica
	M	Biogénesis de la pared celular y de la membrana
	O	Modificación postraduccional, balance entre síntesis y degradación de proteínas, chaperonas
	T	Mecanismos de transducción de señales
	U	Tráfico intracelular, secreción y transporte vesicular
	V	Mecanismos de defensa
	W	Estructuras extracelulares
	Y	Estructura nuclear
	Z	Citoesqueleto
<b>Metabolismo</b>	C	Producción y conversión energética
	E	Transporte y metabolismo de aminoácidos
	F	Transporte y metabolismo de nucleótidos
	G	Transporte y metabolismo de carbohidratos
	H	Transporte y metabolismo de coenzimas
	I	Transporte y metabolismo de lípidos
	P	Transporte y metabolismo de iones inorgánicos
	Q	Transporte y metabolismo de metabolitos secundarios
R	Función hipotética	
S	Función desconocida	
Vc*	Cápside/envoltura	
X*	Varias funciones	

\*Categorías nuevas asignadas para este estudio

## 2.6 Búsqueda de homólogos en bases de datos celulares y virales

Se seleccionaron todos aquellos grupos de ortólogos de los megavirus con las secuencias que tenían un identificador Pfam. Aquéllos que no tenían este identificador se clasificaron en los COGs R y S (funciones pobremente caracterizadas). Para identificar a los homólogos remotos de Bacteria, Archaea, Eukarya y otros virus en la base de datos *KEGG* (actualización del 2011) se utilizaron los programas de *Hmmer* (S. R. Eddy, 1998) con un valor  $E < 10E-3$  y *Psi-Blast* (Altschul et al., 1997) con un valor  $E < 10E-3$  y un valor  $C = 75\%$ . Una vez obtenidas las secuencias homólogas celulares y de otros virus para cada grupo de ortólogos de megavirus, se eliminaron aquellas redundantes con el mismo umbral de similitud mayor al 80% con *CD-HIT* (Fu, Niu, Zhu, Wu, & Li, 2012). Los títulos (*headers*) de cada secuencia *FASTA* fueron editados con *Bash* para que fuesen más cortos. Las secuencias ortólogas celulares y virales obtenidas se contabilizaron para cada uno de los grupos taxonómicos (desde familias virales hasta los phyla y dominios celulares) usando *Bash* y *Awk*.

## 2.7 Análisis filogenéticos basados en estructura primaria del repertorio del pangenoma de megavirus

Para alinear todas las secuencias homólogas celulares y virales obtenidas por perfiles *HMM* y *PSSM* de cada grupo de ortólogos de los megavirus se usó el programa de alineamiento múltiple *MAFFT* (parámetros por *default*) (Kato, Misawa, Kuma, & Miyata, 2002). Para remover las secuencias espurias y las regiones pobremente alineadas del alineamiento múltiple se utilizó *Trimal* (parámetros por *default*) (Capella-Gutiérrez, Silla-Martínez, & Gabaldón, 2009). Para construir el árbol filogenético con máxima parsimonia para cada uno de los grupos con las secuencias homólogas celulares y virales se empleó *IQ-TREE* (parámetros: *ProTest* y modelos de selección *WAG*, *LG*, *JTT*; *Ultrafast Bootstrap* con 1000 repeticiones) (Nguyen, Schmidt, von Haeseler, & Minh, 2015). Para visualizar y manipular a los árboles evolutivos generados se utilizó la plataforma de *iToL* (Letunic & Bork, 2016) en el que el archivo de entrada estaba en formato *Newick*. Al mismo tiempo se editó en *Bash* un archivo de texto con los nombres de cada unidad taxonómica operativa (OTU, por sus siglas en inglés) para darle un formato de colores cuyos códigos se encuentran en <https://htmlcolorcodes.com/es/> [Bacteria, morado (#c9a1f4); Archaea, verde (#97eda0); Eukarya, azul (#8ce2f2) y Virus, rojo

(#f8a593)]. Este archivo de texto con los códigos de colores se anexó a cada árbol filogenético en *iToL*. Cada árbol filogenético de cada grupo de ortólogos de megavirus fue circular, sin mostrar la longitud de sus ramas y eliminando aquellas ramas con un valor de *Ultrabootstrap* <85%. Las imágenes de cada árbol se exportaron en formato *.svg* o *.png*.

## 2.8 Construcción de la base de datos de estructuras terciarias

En el Laboratorio de Origen de la Vida de la Facultad de Ciencias, se ha implementado una nueva estrategia metodológica para la construcción de árboles filogenéticos y para la reconstrucción de estados ancestrales de proteínas a partir de las arquitecturas de las estructuras terciarias disponibles (Alvarez-Carreño, Alva, Becerra, & Lazcano, 2018; Jácome et al., 2015). Dicho procedimiento se utilizó para el análisis filogenético de los dominios estructurales y catalítico de la polimerasa de RNA (NS5B) del virus de la hepatitis C (VHC) y la transcriptasa reversa (RT) del virus de la inmunodeficiencia humana (VIH).

Para la construcción de la base de datos de las estructuras cristalográficas de NS5B de VHC y de RT de VIH, se buscaron los términos “Hepatitis C virus AND polymerase” y “Human immunodeficiency virus AND reverse transcriptase” en el Banco de Datos de Proteínas de la Investigación Colaborativa para la Bioinformática Estructural (*RCSB PDB*, por sus siglas en inglés) y el de Europa (*PDBe*). Se seleccionaron y descargaron aquellas estructuras de referencia para ciertos genotipos y subtipos virales. Asimismo, se descartaron aquellas estructuras redundantes >90% de similitud. Dichas estructuras seleccionadas se agruparon de acuerdo a la presencia o ausencia de ligandos acoplados, a la similitud de dichos ligandos (naturaleza química del ácido nucleico, antivirales y moléculas unidas a sitios alostéricos) y a la resolución de las mismas. Además, también se descargó una tabla personalizada con toda la información biológica de dichas estructuras para un análisis posterior.

## 2.9 Análisis filogenéticos basados en estructuras terciarias

Para la comparación de las estructuras cristalográficas seleccionadas de NS5B y RT se utilizó el programa *PDBeFold* que se basa en el alineamiento pareado de las estructuras secundarias (SSM, por sus siglas en inglés) a través de la plataforma en línea del PDBe (<http://www.ebi.ac.uk/msd-srv/ssm/>) con parámetros por *default* (Krissinel & Henrick, 2004). Para cada una de las comparaciones se calculó el siguiente valor de alineamiento estructural (SAS):

$$\text{SAS} = \text{RMSD} \times 100/n_{\text{al}}$$

Este valor representa a la desviación de la media cuadrática mínima (RMSD) que existe entre las distancias de los carbonos alfa por el número de residuos alineados ( $n_{\text{al}}$ ) de las dos proteínas superpuestas (Subbiah, Laurents, & Levitt, 1993). Se elaboró una matriz de distancia con todos los valores SAS para cada uno de los grupos de estructuras cristalográficas con o sin ligandos y/o sustratos. Se construyeron los dendogramas basados en matrices cuyas distancias se esperan igualar a la suma de la longitud de las ramas entre los OTUs con el algoritmo *FITCH* (paquetería de *PHYLP* versión 3.695). Se visualizaron y se editaron los dendogramas con FigTree.

Para la visualización interactiva y representación de las estructuras cristalográficas se utilizó el *software* Chimera versión 1.11 (Pettersen et al., 2004; Subbiah et al., 1993), el cual, a la vez, contiene programas para hacer la superposición pareada de estructuras de proteínas con su respectivo alineamiento pareado de sus secuencias (*MatchMaker*) o una superposición y alineamiento múltiple de las mismas (*Match/Align*).



## III. RESULTADOS

### 3.1 Bases de datos biológicos y ecológicos

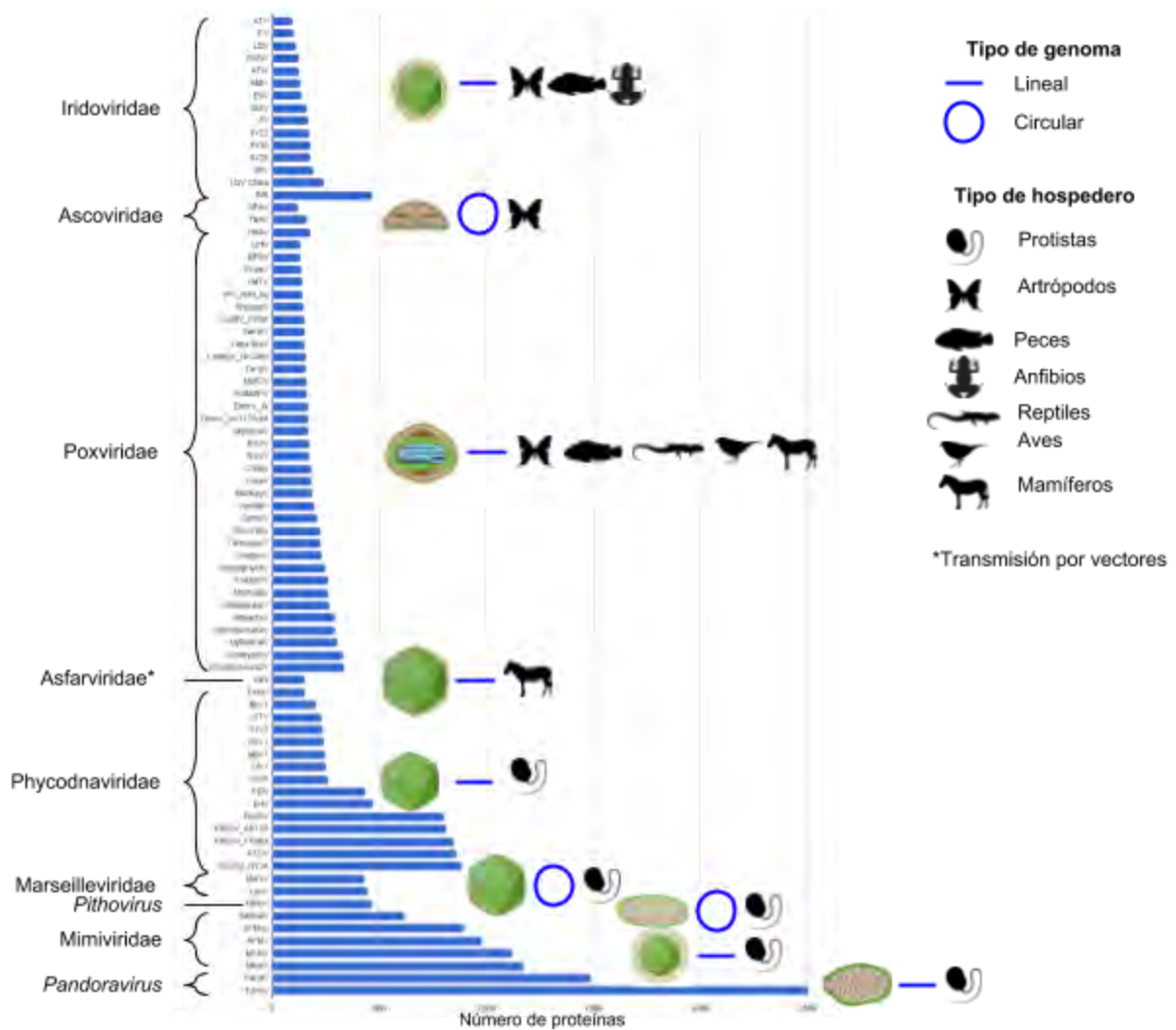
[Véase la sección ANEXO I (artículo publicado) para verificar los resultados].

### 3.2 Bases de datos de proteomas virales

Se obtuvieron registros proteómicos de 79 especies de las siete familias (Ascoviridae, Asfarviridae, Iridoviridae, Marseilleviridae, Mimiviridae, Poxviridae, Phycodnaviridae) y dos especies virales (*Pithovirus* y *Pandoravirus*). En la Figura 1 se aprecia que los Iridoviridae, Poxviridae y Phycodnaviridae tienen el mayor registro de proteomas en el *GenBank*. Los virus que infectan a invertebrados y vertebrados tienen los proteomas más pequeños (~200 proteínas), mientras que aquéllos que infectan a protistas, poseen los proteomas de hasta más de 2,000 proteínas. El tipo de genoma no depende del tamaño del proteoma. Las cápsides son complejas y tienen una membrana lipídica interna.

### 3.3 Pagenoma viral de los megavirus

Para el análisis proteómico posterior se descartaron tres proteomas de virus que no pertenecían a una familia y al único virus de la familia Asfarviridae. También se excluyeron aquéllos proteomas que resultaron alejarse de la media de la suma del número de proteínas por cada familia viral. Finalmente se utilizaron 64 proteomas de referencia para dicho análisis con un total de más de 18 mil proteínas (Tabla 1). La granularidad fina de los grupos de homólogos se definió a través de la combinación de *Blastp* y *Hmmer* para la búsqueda robusta de ortólogos y parálogos basada en secuencias a través de Pfam (2015) gracias a la paquetería de software *GET\_HOMOLOGUES* (Contreras-Moreira & Vinuesa, 2013). Para estimar el número de genes conservados en la familia viral (núcleo genómico), el número de genes compartidos encontrado en la adición secuencial de cada nuevo genoma viral fue extrapolado por el ajuste de una función de decaimiento exponencial de los datos con el modelo de Tettelin (Fig. 2).

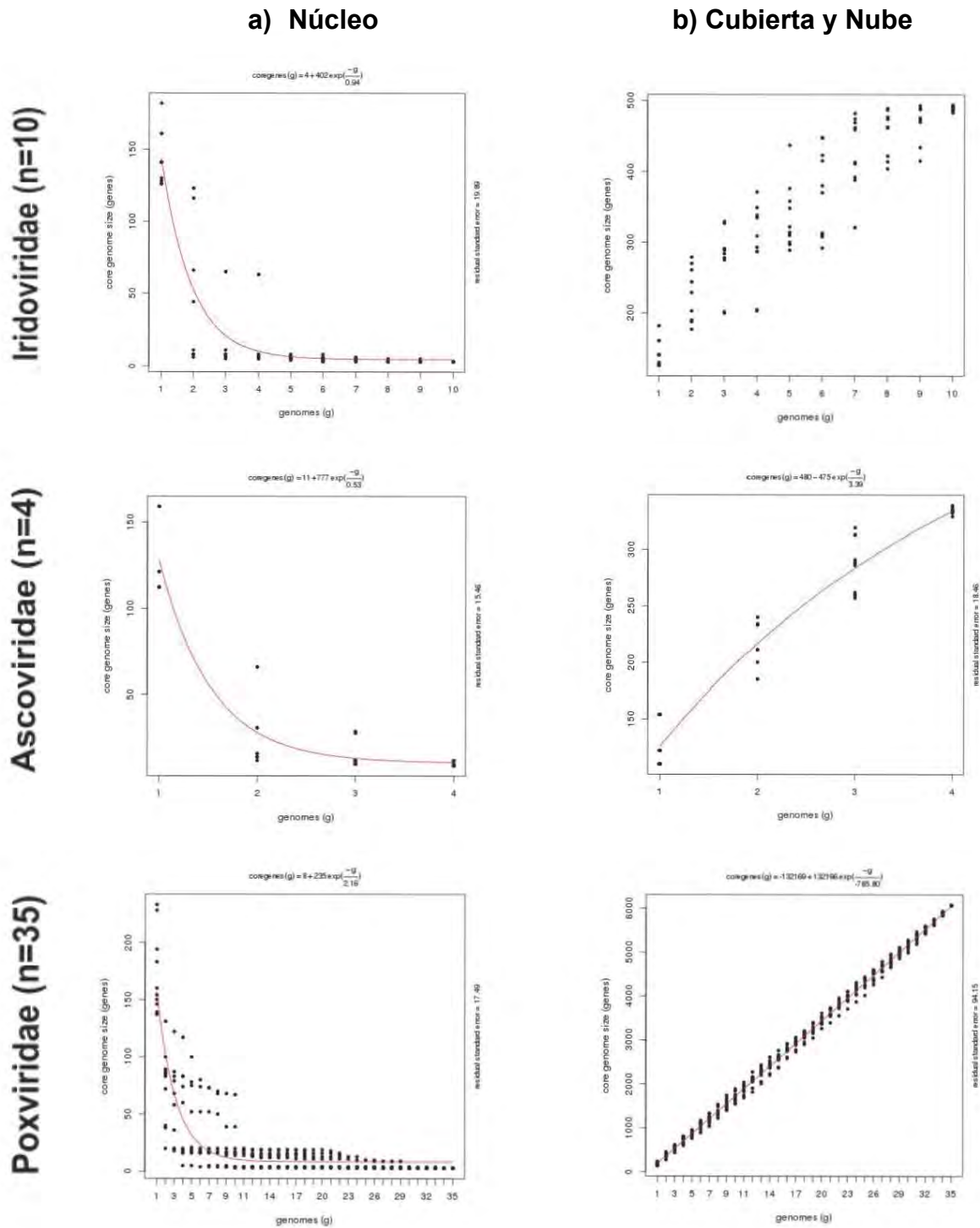


**Fig. 1** Número de proteínas por genoma de cada una de las familias de los megavirus. Los virus con proteomas más pequeños se encuentran distribuidos, principalmente, en virus que infectan invertebrados y vertebrados (entre 100 a 300 proteínas). Los proteomas virales más grandes se encuentran distribuidos en familias que infectan protistas (entre 200 a 2,500 proteínas).

El resultado de todas las permutaciones muestra que el número de genes compartidos en el núcleo pangénomico disminuye con la adición de un nuevo genoma en cada familia viral (Fig. 2a). No obstante, la extrapolación de cada una de las curvas indica que el número de proteínas del núcleo puede llegar a mantenerse relativamente constante a pesar de la adición de nuevos genomas. El caso anterior es más evidente con Poxviridae: el núcleo genómico mínimo de sus 35 genomas alcanza una curva asintótica de cuatro grupos de homólogos (Fig. 2a y Tabla 1). Para estimar el número de genes esenciales (cubierta) y únicos (nube) se ajustaron también por el modelo de decaimiento exponencial (Fig. 2b). Se observó que en todos los genomas de los megavirus son abiertos y su tamaño puede incrementarse con el número de nuevos genomas virales agregados. Además, los genes parálogos son una evidencia de que la cubierta y, principalmente, la nube, están en constante crecimiento debido al origen de nuevos genes para el repertorio genético de cada familia viral (Tabla 1). La división de los grupos de homólogos de la matriz de COGTriangles (núcleo, cubierta y nube) para cada familia viral se aprecia en la Figura 3. Los núcleos más pequeños resultaron ser para aquellas familias virales que infectan a algas, invertebrados y vertebrados; mientras que los más grandes se encontraron en familias virales que infectan a amebas, lo que es esperable que ya que su genoma es de 5 a 10 veces más grande que el resto de los megavirus. Debido a que el número de genomas es pequeño, Ascoviridae y Mimiviridae no tienen cubierta. La nube contiene genes que son únicos a especies virales y destaca el hecho que Phycodnaviridae es la que presenta la mayor diversidad genética, mientras que Marseilleviridae presenta un poco más de 600 grupos ortólogos, quizás porque la familia no es tan divergente.

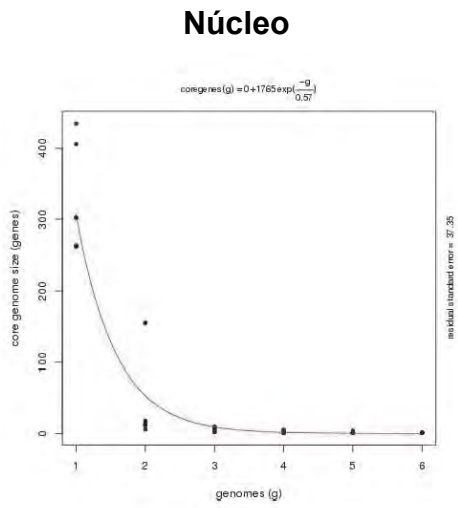
**Tabla 1 Grupos de ortólogos del pangénoma de los megavirus**

Familia viral	Número de genomas	Número de proteínas	Número de grupos					
			Núcleo		Cubierta		Nube	
			Ortólogos	Parálogos	Ortólogos	Parálogos	Ortólogos	Parálogos
Iridoviridae	10	1568	4	0	143	6	565	24
Ascoviridae	4	586	22	2	0	0	535	12
Poxviridae	35	7284	4	0	235	17	1343	77
Phycodnaviridae	6	2327	5	0	22	1	1717	92
Marseilleviridae	5	2236	250	21	100	2	615	8
Mimiviridae	4	4169	541	6	0	0	1486	36

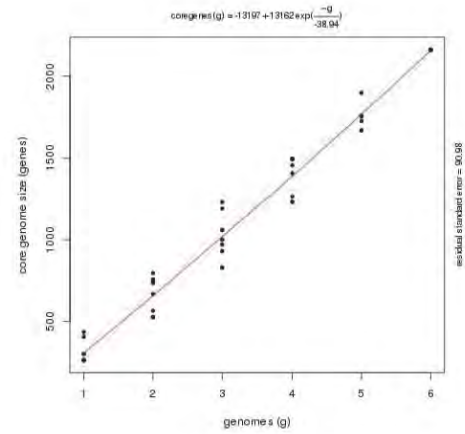


**Fig.2** Estimación estadística del núcleo (a) y el resto del pangenoma (b) de los megavirus. El número de genes compartidos, esenciales y únicos de cada familia viral se grafica como una función del número de genomas secuencialmente agregados en la clusterización.

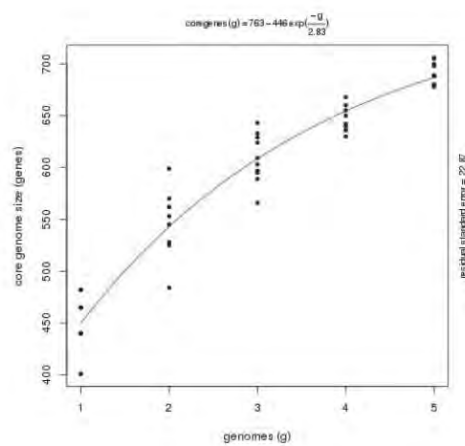
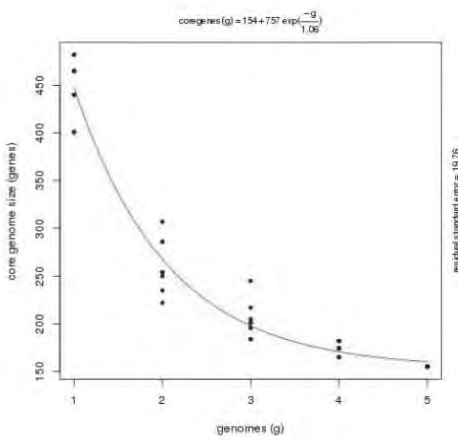
**Phycodnaviridae (n=6)**



**Cubierta y Nube**



**Marseilleviridae (n=5)**



**Mimiviridae (n=4)**

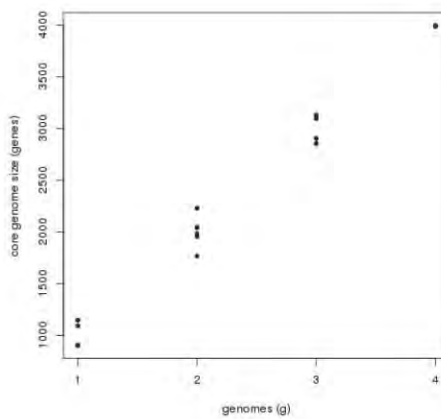
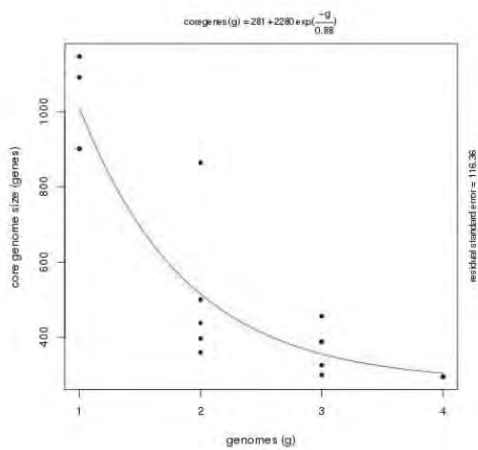
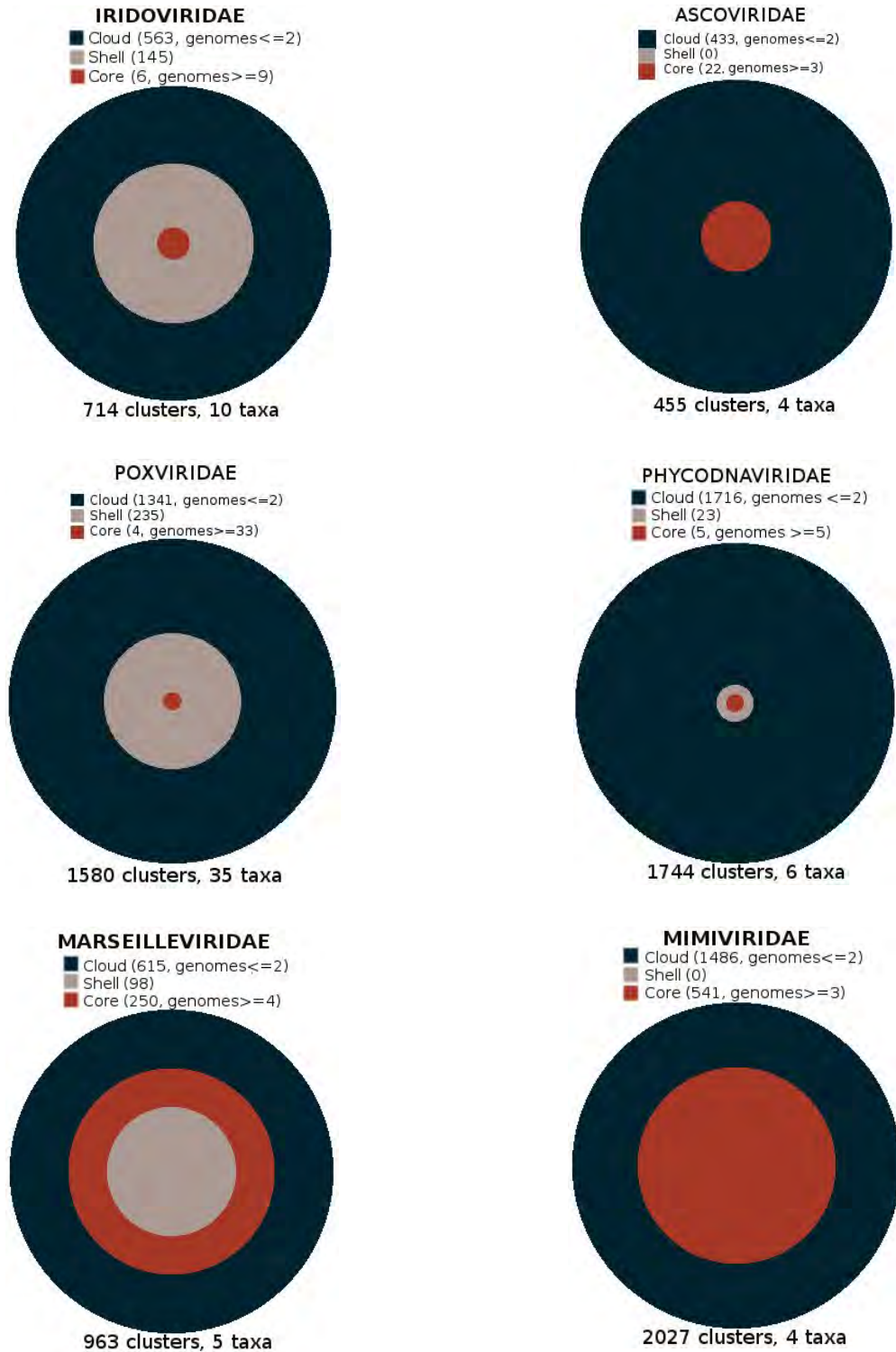


Fig.2 (Continuación...)



**Fig. 3** Análisis gráfico de del pangenoma de los megavirus. Los grupos de genes ortólogos se muestran compartimentalizados en frecuencias relativas (números de grupos) contenidos en el núcleo (*core* y *softcore*), cubierta (*shell*) y nube (*cloud*).

### 3.4 Composición funcional del pangenoma de los megavirus

Para comprender los roles funcionales de los grupos de ortólogos que constituyen el núcleo, la cubierta y la nube, se utilizó la clasificación funcional de los COGs (*Cluster of Orthologous Groups*) (Tatusov, 1997). Este sistema de clasificación se basa en la relación de homología entre proteínas. Se utilizó este sistema para poder clasificar los grupos de ortólogos virales de acuerdo a su función obtenida a través de la base de datos Pfam o de referencias citadas. Dichas categorías, identificadas con una letra, pertenecen a cuatro funciones generales: i) procesamiento y almacenamiento de la información genética (A, B, J, K, L), ii) procesos celulares y de señalización (D, M, N, O, T, U, V, W, Y, Z), iii) metabolismo (C, E, F, G, H, I, P, Q), y iv) funciones desconocidas (R, S). Se agregaron las categorías X y Vc para incluir grupos de ortólogos que intervienen en varias funciones (genéticas, celulares y metabólicas) y en el procesamiento de la cápside, respectivamente (X). Al mismo tiempo, los COGs se clasificaron de acuerdo a su presencia en el núcleo, en la cubierta y en la nube del pangenoma viral (Fig. 4).

La Figura 4 muestra un *heatmap* de las frecuencias relativas para cada una de las funciones de los grupos de ortólogos de cada una de las familias de los megavirus por núcleo, cubierta y nube. De acuerdo al carácter predominante de las funciones de los grupos ortólogos, el *heatmap* se puede dividir en tres agrupaciones. En primer lugar, las categorías funcionales con mayor densidad de grupos de proteínas (parte inferior del *heatmap*) son aquellas que se concentran, como era esperado, en el pangenoma (cubierta y nube) y, las cuales, no se les ha asignado una función específica. El núcleo pangenómico presenta principalmente este tipo de ortólogos sin función y, se observa sobre todo, en los megavirus con genomas más grandes (Marseilleviridae y Mimiviridae, ambos infectan a amebas). Estos grupos de ortólogos con función indeterminada abarcan hasta un 70% de todo el pangenoma de los megavirus como se ha determinado en otros estudios (Boyer, Gimenez, Suzan-Monti, & Raoult, 2010; D. Raoult, 2004). Algunos de estos ortólogos con función desconocida podrían tener funciones en la formación de la cápside (Sobhy, Scola, Pagnier, Raoult, & Colson, 2015) y, en estudios preliminares nuestros y de *GenBank* y de *Uniprot*, se ha encontrado que la mayoría presenta firmas moleculares referentes a funciones de tipo informacional (como factores de transcripción y traducción) y celular (como transducción de señales y apoptosis). En esta misma agrupación, predominan aquellos grupos de ortólogos que intervienen en procesos celulares como la

apoptosis, el plegamiento y balance entre síntesis y degradación de proteínas (O); y, además, en procesos informacionales como la transcripción (K), la replicación y la reparación del DNA (L).

En el segundo lugar (parte media del *heatmap*), se encuentra el grupo de ortólogos con funciones celulares como mecanismos de defensa (V, p. ej. quimiocinas), como transducción de señales (T, p. ej. cinasas de tirosina/serina, por ejemplo) y como estructuras extracelulares (W, p. ej. colágena); con funciones metabólicas como la biosíntesis de aminoácidos (E, p. ej. transferasa de glutamina), nucleótidos (F, p. ej. reductasa de dihidrofolato), carbohidratos (G, p. ej. transferasa de glicosilo) y lípidos (I, lipasa de triacilglicerol), y como la producción energética (C, p. ej. citocromo P450); y con funciones informacionales como el procesamiento del RNA (A, algunas helicasas y exonucleasas), como la traducción (J, sintetisas de tRNA). Las funciones anteriores se conservan principalmente en los Mimiviridae.

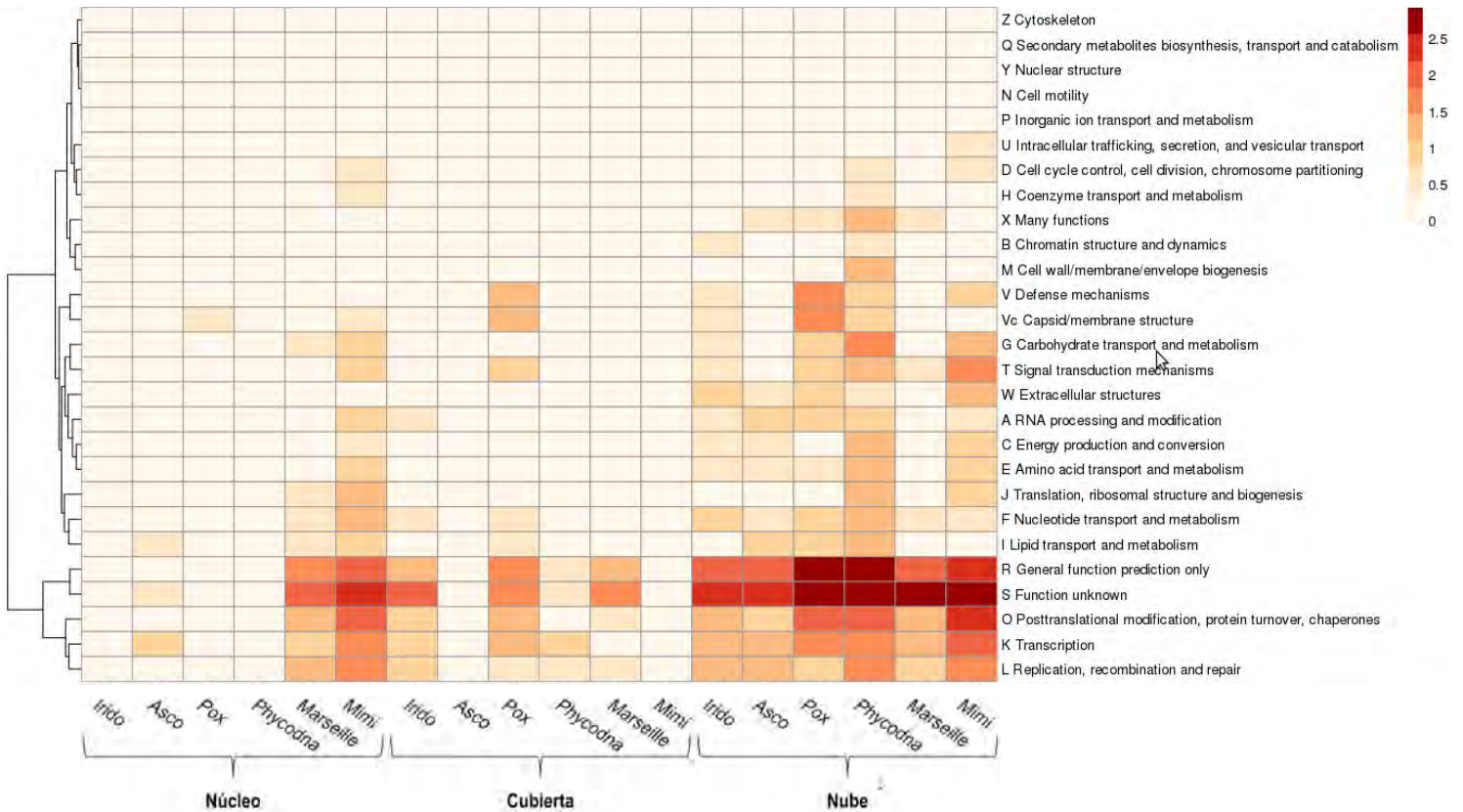
Finalmente, son nulos o pocos los grupos de ortólogos identificados que intervienen en la dinámica del ciclo celular (D, ciclinas), en procesos metabólicos que intervienen en el transporte de coenzimas (H, p. ej. transferasa de N-acetilo), en mecanismos celulares que intervienen en la membrana (M, transportadores ABC) y en procesos informacionales (B, histonas). Otras funciones, como las repeticiones de ankirina, repeticiones de leucina, ATPasas, están principalmente sobrerrepresentadas en las familias con los genomas más grandes (Poxviridae, Phycodnaviridae, Marseilleviridae y Mimiviridae).

El 30 % de los grupos de ortólogos con una función conocida están presentes en algunos o en los tres dominios de la vida Archaea (A), Bacteria (B) y Eukarya (E) y en otros grupos virales (V); mientras que el resto son proteínas huérfanas con función desconocida (S) o hipotética (R) distribuidos sobretudo en proteomas virales (Fig. 5a). Como se dijo anteriormente, estos grupos de ortólogos tienen firmas moleculares que se distribuyen en otros virus o en los dominios celulares. Estos grupos con función desconocida se distribuyen principalmente en el pangénoma (cubierta y nube) de los megavirus y en el núcleo genómico. Sin embargo, en éste último, sólo están presentes en los Marseilleviridae y Mimiviridae hasta en un 98% (466/473) (Fig. 5b).

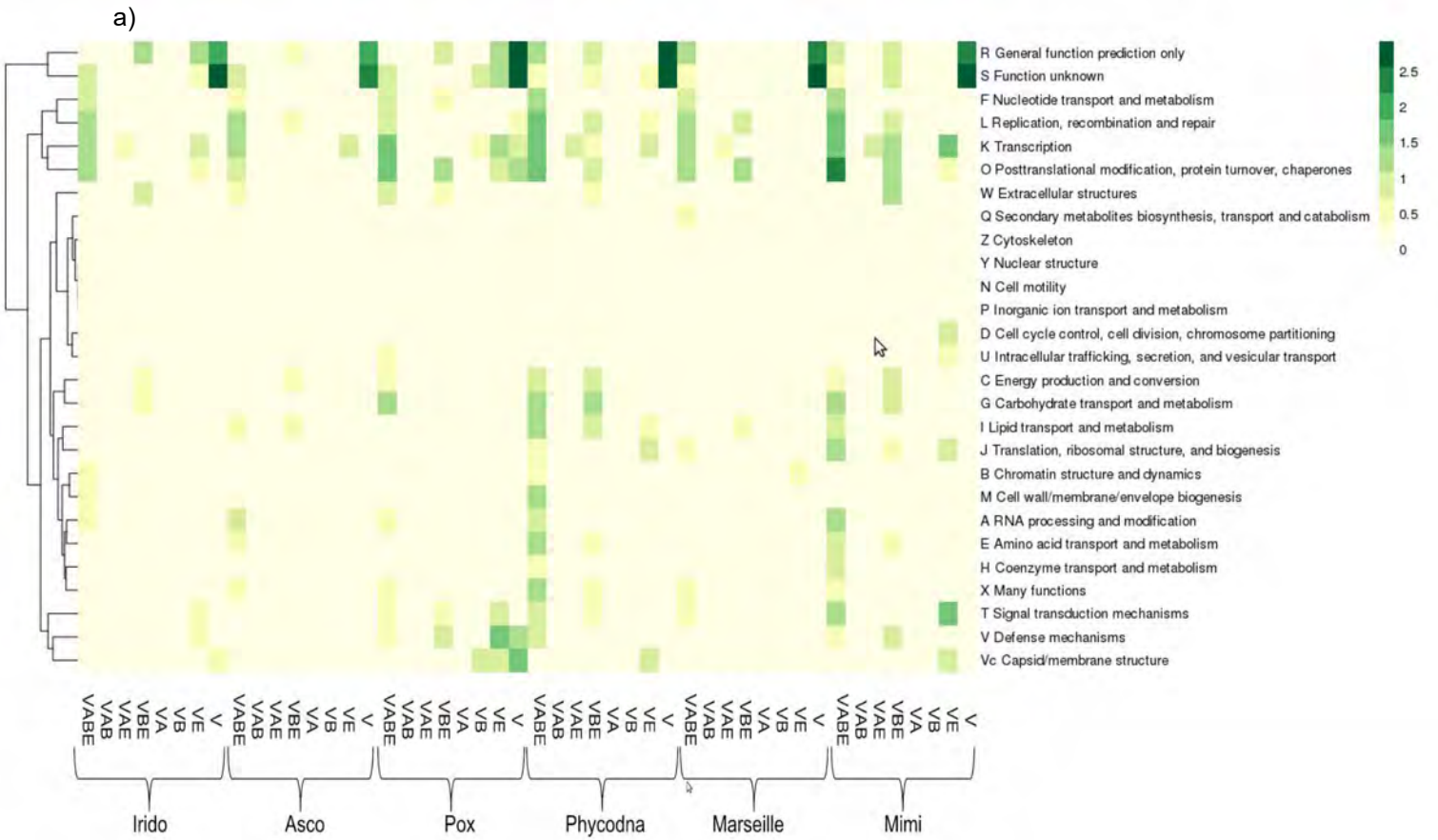
El 17% de los grupos de ortólogos se distribuye en los tres dominios y otros virus (VABE), y principalmente participan en procesos postraduccionales, en el plegamiento de proteínas, en la apoptosis y en la transducción de señales; así como, en la transcripción y procesamiento del DNA y RNA; y en el metabolismo de nucleótidos (parte superior del *heatmap*). Los Phycodnaviridae y los Mimiviridae también presentan grupos de ortólogos



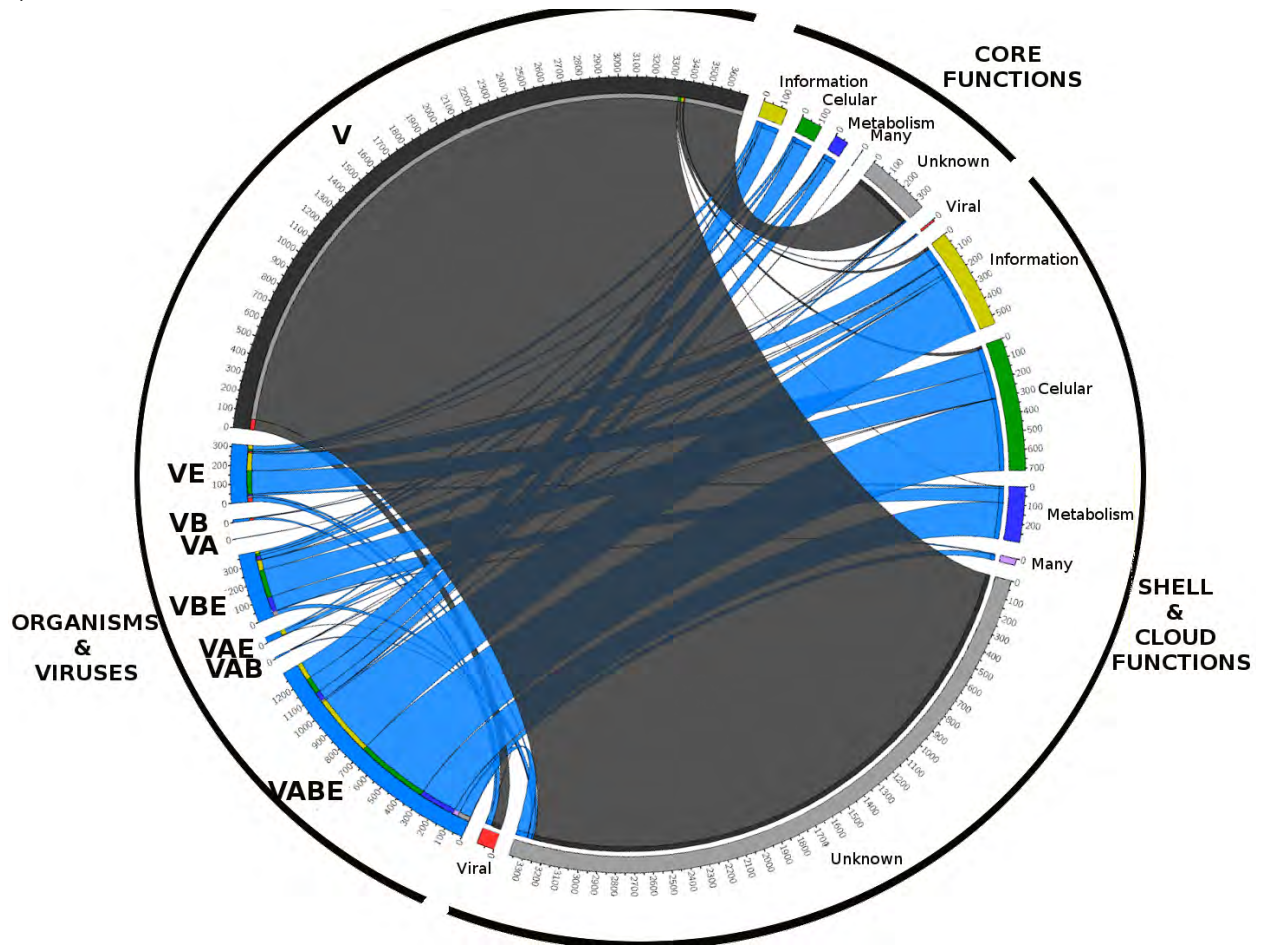
distribuidos en VABE y que tienen funciones relacionadas a la traducción y procesamiento del RNA; a mecanismos de defensa; y al metabolismo de lípidos, aminoácidos y coenzimas (parte inferior del *heatmap*). La mayoría de los ortólogos que tienen que ver con funciones de transcripción y replicación y reparación del DNA y procesos postraduccionales y apoptosis tienen una distribución en los tres dominios de la vida en otros virus (Fig. 5b).



**Fig.4** Comparación funcional y pangenómica de los grupos de ortólogos de los megavirus. En el eje horizontal, las funciones se dividen de acuerdo a los grupos de ortólogos (COGs) descritos por (Tatusov, 1997) con valores normalizados logarítmicamente en base 10. Estas funciones intervienen en el procesamiento y almacenamiento de la información genética (A, B, J, K, L), procesos celulares y de señalización (D, M, N, O, T, U, V, W, Y, Z), metabolismo (C, E, F, G, H, I, P, Q), funciones desconocidas (R, S). Otras categorías fueron agregadas para este trabajo: varias funciones (X) y relacionadas a la cápside (Vc). En el eje vertical, las funciones se agrupan de acuerdo a la frecuencia de los grupos de ortólogos presentes en el núcleo genómico, en la cubierta y en la nube y por cada una de las familias virales: Iridoviridae, Ascoviridae, Poxviridae, Phycodnaviridae, Marseilleviridae y Mimiviridae. Se observan tres grupos de ortólogos de acuerdo a su frecuencia: poco o nada abundantes (principalmente procesos celulares como citoesqueleto o biosíntesis de coenzimas), medianamente abundantes (principalmente en procesos virales como la formación de cápside, en transducción de señales como receptores de tirosina cinasa y en biosíntesis de nucleótidos como la ribonucleótido reductasa) y muy abundantes (funciones desconocidas, de transcripción como la RNAPol y factores de transcripción y de apoptosis como los dominios repetidos de ankirina y de replicación como la DNAPol). Ascoviridae y Mimiviridae no tienen cubierta porque presentan solo cuatro genomas.



b)



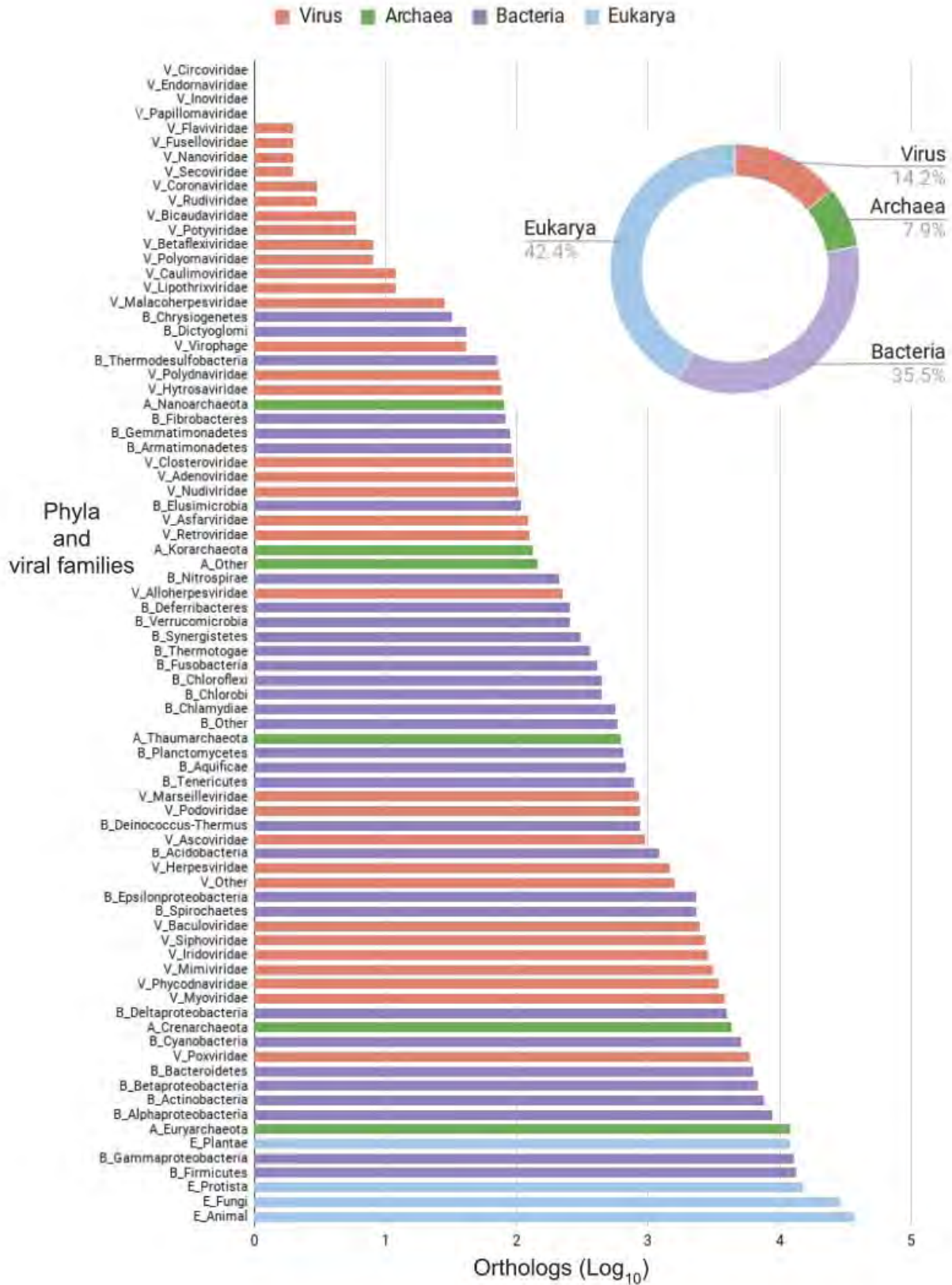
**Fig.5** Comparación funcional y de la distribución por dominios de los grupos de ortólogos de los megavirus obtenida de la base de datos Pfam. a) En el eje horizontal se presentan nuevamente los COGs de acuerdo a la figura 6. En el eje vertical, la frecuencia de los grupos ortólogos se presenta por familias virales (V) y por su distribución en otros virus (V), en Archaea (A), en Bacteria (B) y en Eukarya (E). b) La visualización en *Circos* (Krzywinski et al., 2009) presenta la frecuencia absoluta de los grupos de ortólogos (ancho de la banda) agrupados de acuerdo al pangenoma (Core, para el núcleo genómico y Shell and Cloud, para la cobertura y a la nube) y a las categorías funcionales generales (Inf, procesamiento y almacenamiento de la información genética en amarillo; Cel, procesos celulares y de señalización en verde; Met, metabolismo en azul rey; Many, varias funciones en morado; *Unknown*, funciones desconocidas en gris y *Viral*, funciones virales en rojo). Asimismo, estos grupos de ortólogos se clasifican de acuerdo a su distribución en uno o más dominios de la vida (ABE) y en virus (V) en bandas en azul cielo. El 70% de los grupos ortólogos con función desconocida se distribuye aparentemente en la cobertura y nube de los megavirus.

### 3.5 Análisis filogenéticos basados en la estructura primaria

Para determinar el posible origen de cada uno de los grupos de ortólogos del pangenoma completo (núcleo, cubierta y nube) de los megavirus, se seleccionaron a aquéllos

que tienen un dominio Pfam (Iridoviridae, 213/714; Ascoviridae, 131/455; Poxviridae, 603/1580; Phycodnaviridae, 486/1744; Marseilleviridae, 184/963; y Mimiviridae, 743/2027). Los ortólogos con función desconocida (R y S) no fueron considerados en este estudio debido a que no presentan homólogos celulares (secuencias huérfanas). Usando la base de datos *KEGG*, se procedió a hacer la búsqueda de homólogos lejanos de células y de otros virus con perfiles de modelos ocultos de Markov (HMM) y con matrices de sustitución específica por iteraciones (PSI-BLAST). Al realizar un conteo de todas las secuencias encontradas por similitud por cada uno de los grupos de ortólogos de los megavirus, se identificó que la mayoría de ellas pertenece a eucariontes, ya sea a sus propios hospederos (algas, protistas, artrópodos, anfibios, aves y mamíferos) o a organismos filogenéticamente relacionados. También se encontraron secuencias pertenecientes a Bacteria como las proteobacteria (principalmente Gammaproteobacteria) y Firmicutes y a Archaea como Euryarchaeota y Crenarchaeota. Por otro lado, el 14% de las secuencias ortólogas de otros virus pertenecieron, principalmente, a las de fagos (Myoviridae y Siphoviridae), a las de virus que infectan insectos y a decápodos (Baculoviridae) y a las de virus que infectan a animales (Herpesviridae). Existen secuencias ortólogas, como las de cápside y algunos factores de transcripción, que solo tienen homólogos entre los mismos megavirales (Poxviridae, Mimiviridae y Phycodnaviridae, principalmente) (Fig. 6).



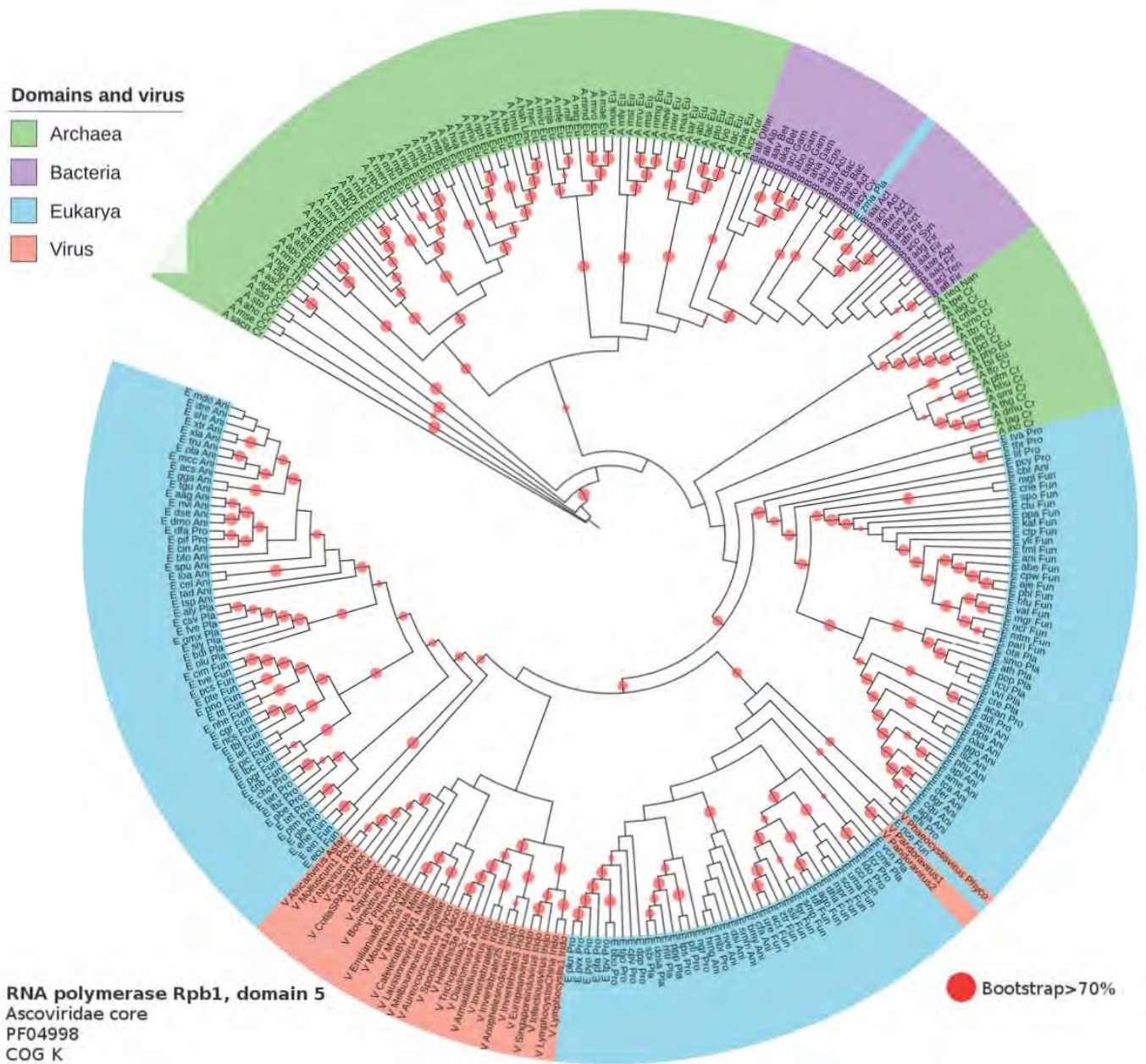


**Fig.6** Distribución taxonómica de las secuencias de los grupos homólogos entre megavirus, otros virus y células de la base de datos *KEGG* obtenidas a partir de la comparación de perfiles HMM y de iteraciones

por PSI-BLAST. Se observa que los grupos ortólogos de los megavirus tienen homólogos en el 40% de los casos para eucariontes, principalmente con sus propios hospederos u hospederos relacionados filogenéticamente (animales, hongos, protistas y plantas).

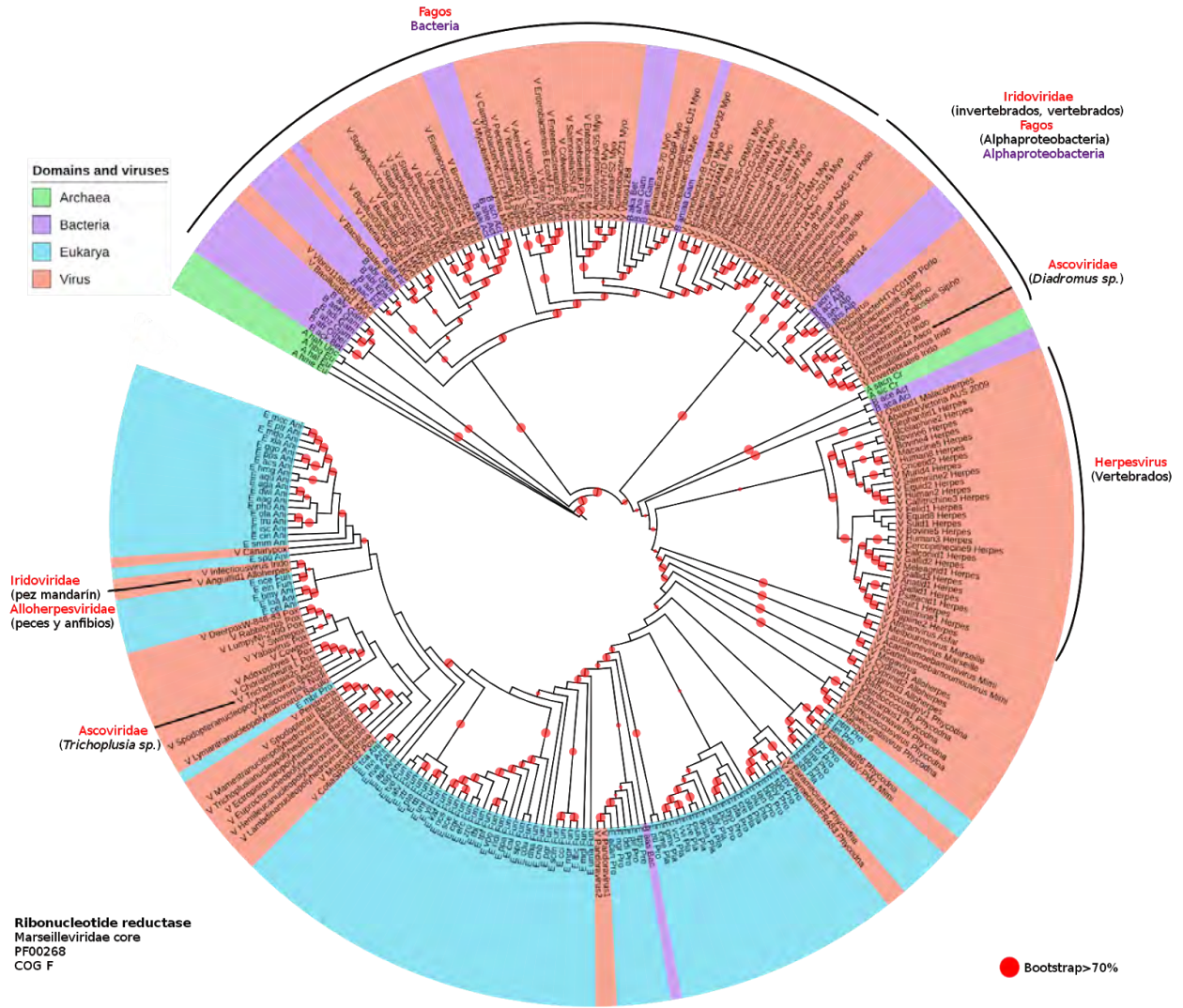
Se construyeron más de 2,000 árboles filogenéticos basados en máxima verosimilitud. Por cuestiones de espacio, sólo se mostrarán y discutirán las filogenias de algunas proteínas. La mayoría de las filogenias de los grupos de ortólogos se pueden ver en la página de iTOL (<https://itol.embl.de/>, en la sección *Sharing data* con el ID *clon666*). Varias de las proteínas clave para la reconstrucción filogenética de los megavirus en otros estudios es la RNA polimerasa dependiente de DNA tipo II (RNAPol II) que es un enzima ubicua que interviene en procesos de transcripción de la información genética (Yutin & Koonin, 2012). El grupo de ortólogos de RNAPol II, en particular, el subdominio 1, se encontró en varios núcleos pangenómicos de los megavirus. Este subdominio 1 de la RNAPol II forma dos clados: por una parte, forma una rama en la que por un lado se encuentran la mayor parte de las familias megavirales y otra rama en la que se encuentran protistas, plantas, hongos y animales (Fig. 7). El otro clado agrupa exclusivamente a las especies de *Pandoravirus* (los megavirus con el genoma más grande), a *Phaeocystis globosa virus* (Phycodnaviridae) y a otros protistas, plantas, hongos y animales. Los *Pandoravirus* quedan junto al *Nosema ceranae* (un microsporidia parásito de abejas), mientras que el phycodnavirus se agrupa con *Entamoeba histolytica* (un protista parásito de vertebrados).

Otra proteína del repertorio conservado de los megavirus en dichos estudios anteriores es la ribonucleótido reductasa (RnR) tipo Ia. Esta proteína está involucrada en la biosíntesis aeróbica de los desoxirribonucleótidos en condiciones aeróbicas. Se ha sugerido que la subunidad pequeña de la RnR (sRnR) tiene una filogenia compleja basada en una transferencia horizontal múltiple (Yutin & Koonin, 2012). El grupo ortólogo de sRnR se encontró en los núcleos pangenómicos de megavirus de nuestra base de datos. Al hacer el análisis filogenético de este grupo, encontramos que, a diferencia de la RNAPol II, se encontró distribuido en varios grupos virales cada uno distribuido con su respectivo hospedero eucarionte (Fig. 8). Es decir, los virus filogenéticamente más cercanos, los Iridoviridae y Ascoviridae, quedan juntos con sus respectivos hospederos (invertebrados y vertebrados, principalmente). Los Herpesviridae quedan junto a eucariontes y los fagos quedan junto a varios phyla bacterianos. Sin embargo, es interesante que el grupo de ortólogos de sRnR de Iridoviridae y Ascoviridae también forma otro clado agrupado con Alphaproteobacteria.



**Fig. 7** Árbol filogenético ML del dominio 5 de la subunidad 1 de la RNAPol II (grupo ortólogo 70 del núcleo pangenómico de Ascoviridae). Los megavirus forman un solo clado junto al clado de protistas, plantas, hongos y animales, mientras que otros megavirus, como *Pandoravirus* y un phycodnavirus, forman otro clado junto a otros eucariontes, en particular, con organismos parásitos como *Microsporidia sp.* y *Entamoeba sp.*

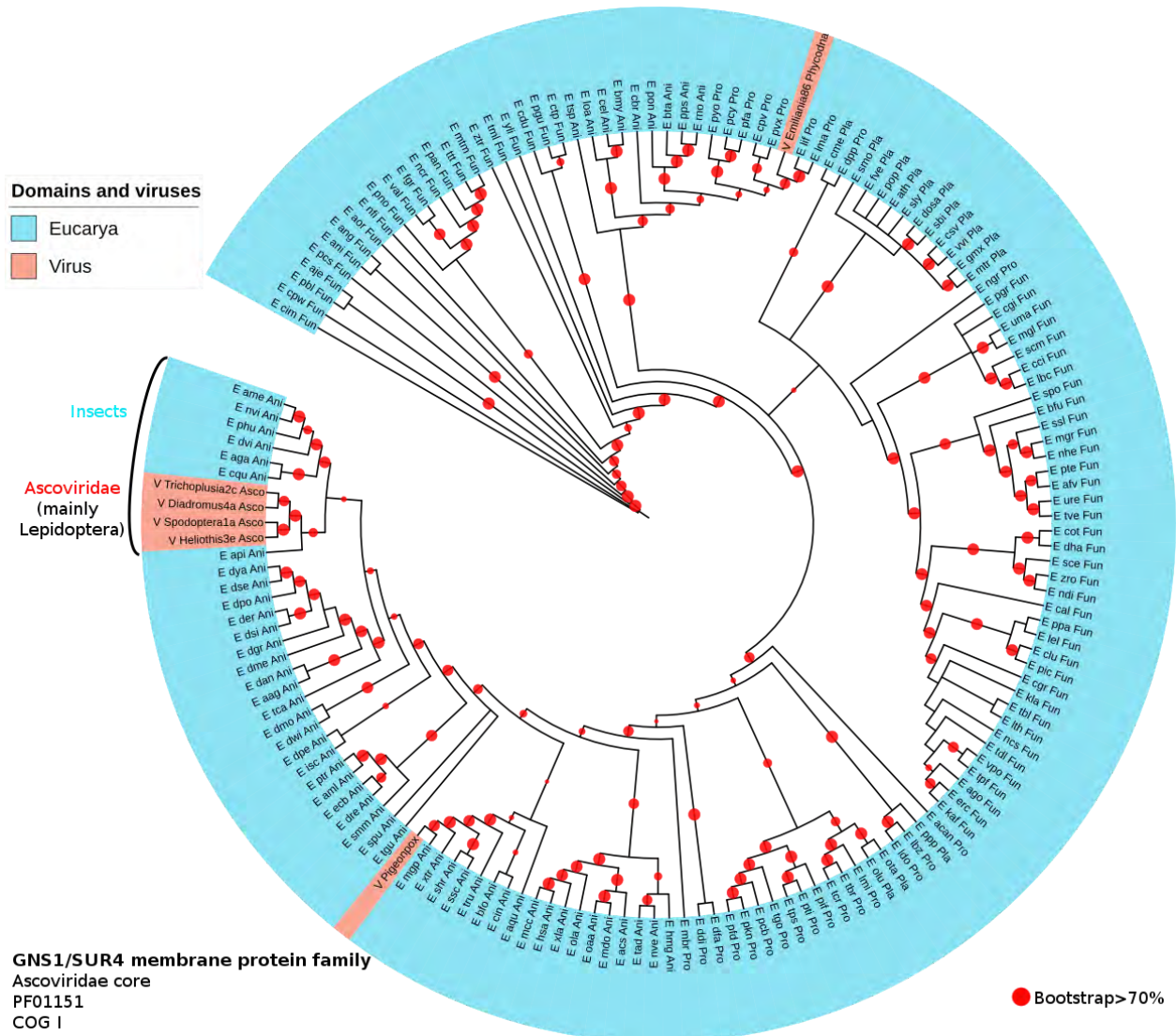




**Fig. 8** Árbol filogenético ML de la subunidad pequeña de la ribonucleótido reductasa IA (grupo ortólogo 1528 del núcleo pangénomico de Marseilleviridae). Los grupos taxonómicos están escritos de acuerdo al color de sus dominios correspondientes. Los hospederos están escritos en paréntesis y en negro. Cada OTU está representado con una letra al principio de acuerdo al dominio (B, Bacteria; A, Archaea y E, Eukarya) o a virus (V) seguido por el código de tres letras del KEGG.



Otra de las proteínas que se encontraron en el núcleo de los megavirus fue la proteína de elongación de cadenas largas de ácidos grasos, la elongasa GNS1/SUR4. Los ortólogos celulares de esta proteína encontrada en estos megavirus formaron tres grupos: 1) los Ascoviridae quedaron junto a invertebrados como los insectos, 2) un poxviridae quedó junto a su hospedero aviar y 3) un phycodnavirus se agrupó con los protistas patógenos de *Leishmania* sp. (Fig. 9). Este grupo de ortólogos megavirales sólo se encontró en el núcleo, en la cubierta y en la nube de Ascoviridae, Poxviridae y Phycodnaviridae, respectivamente.

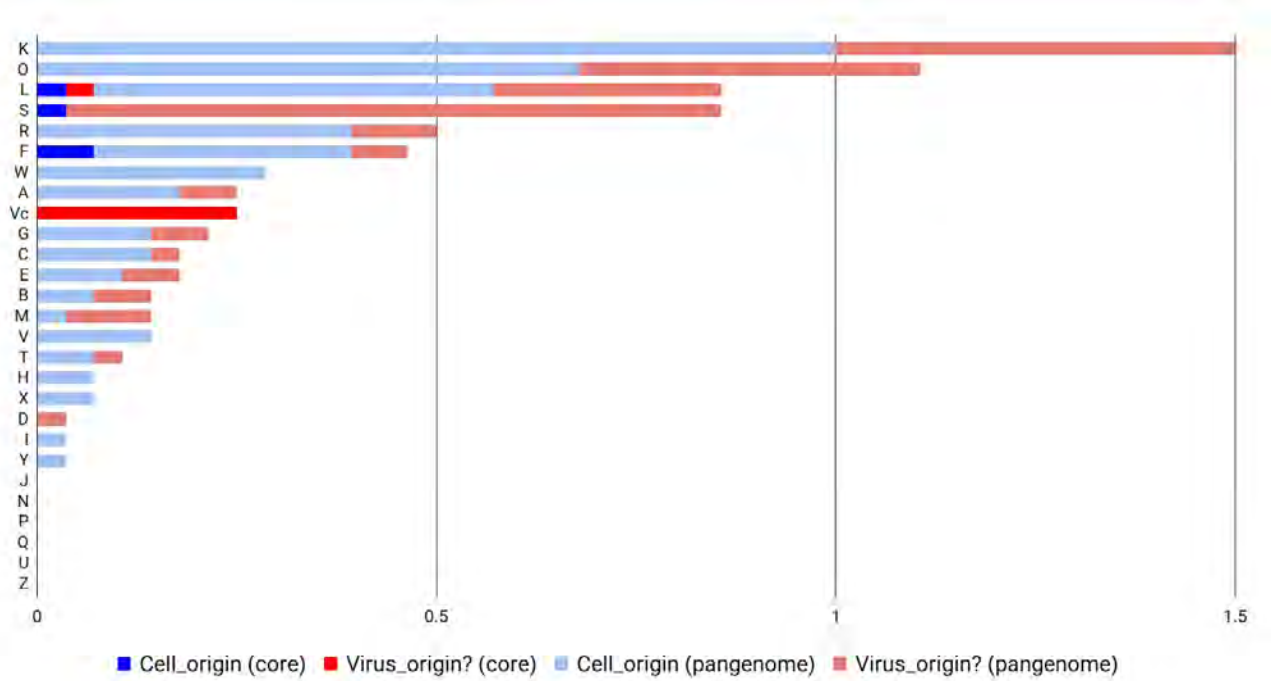


**Fig. 9** Árbol filogenético ML del dominio de elongasa (grupo ortólogo 15 del núcleo pangenómico de Ascoviridae). Los grupos taxonómicos están escritos de acuerdo al color de sus dominios correspondientes. Los hospederos están escritos en paréntesis y en negro. Cada OTU está representado con una letra al principio: E, Eukarya o V, virus seguido por el código de tres letras del *KEGG* (para células) o *GenBank* (para virus).

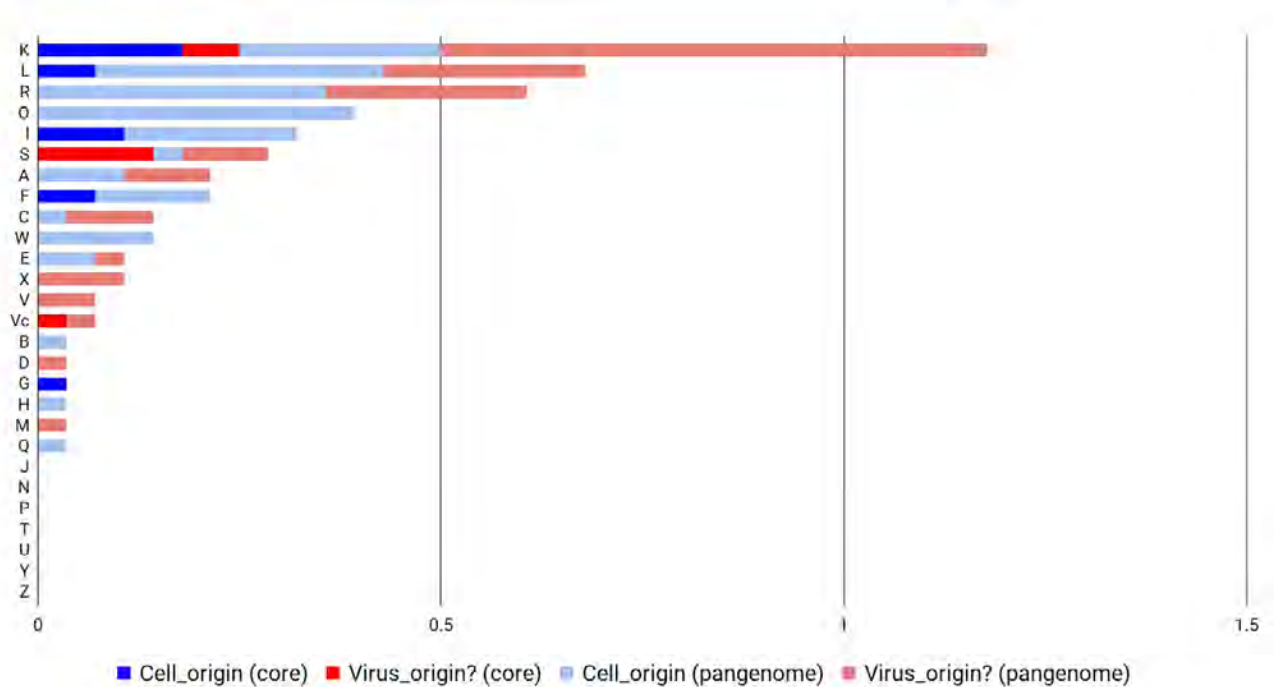
El análisis detallado de la historia evolutiva de cada uno de los grupos de ortólogos de todas las familias de los megavirus, como en el caso anterior, está en proceso actualmente. Por ahora se cuenta con un análisis parcial del origen celular y viral de cada uno de los COGs distribuidos en el núcleo, la cubierta o la nube. Por ejemplo, en la Figura 10, se presenta el análisis poco detallado sobre el origen celular o viral de los grupos de ortólogos por COGs de Iridoviridae y Ascoviridae. Los grupos de ortólogos más sobrerrepresentados para ambas familias intervienen en procesos de transmisión de la información genética (COGs K y L), de postraducción y de apoptosis (COG O) o, por el contrario, tienen una función desconocida (COGs R y S). El 67% de todos los grupos de ortólogos con función de transcripción (COG K) es de origen celular en Iridoviridae (Fig. 10a), mientras que el 64% de estos grupos con esta función es de origen viral en Ascoviridae (Fig. 10b). El 63% de los grupos de ortólogos con funciones de reparación y recombinación del DNA (COG L) son de origen celular para ambas familias. El 61% y 100% de los grupos de ortólogos con funciones de postraducción y apoptosis (COG O) también son de origen celular para Iridoviridae y Ascoviridae, respectivamente.

Entre un 88-96% de los genes de ortólogos con funciones desconocidas (COG S) para ambas familias tiene un posible origen viral. Por el contrario, entre un 60-80% de los grupos de ortólogos con funciones hipotéticas tiene un posible origen celular para ambas familias. Dichas proteínas con una función probable aún mantienen una firma molecular a través de la conservación de dominios y motivos que pudieron haber sido genes escapados de genomas celulares, pero dada la tasa de divergencia en los virus, la proteína completa ha perdido su señal filogenética. Esto podría significar que un estudio más detallado sobre su historia evolutiva a través de la comparación de las estructuras terciarias (aun sin caracterizarse) de estas proteínas con funciones desconocidas con respecto a las de células con los mismos dominios o motivos, se podría determinar su posible origen celular. Aproximadamente el 85% de las proteínas metabólicas de Iridoviridae y el 100% de las de Ascoviridae tienen un origen esencialmente celular.

a) Iridoviridae



b) Ascoviridae



**Fig. 10** Origen celular o viral de los grupos de ortólogos a través de su análisis filogenético de secuencias de a) Iridoviridae y b) Ascoviridae. Los números normalizados del eje de las x representan al número de ortólogos de origen celular (tonalidad azul) o viral (tonalidad roja) y si pertenecen al núcleo genómico (*core*, en color fuerte) o a la cubierta y nube (*pangenome*, en color suave) y a qué categoría funcional representados por letras (COGs) en Tabla 1. Todos los valores se normalizaron logarítmicamente en base 10.

Por otro lado, el grupo de ortólogos de la cápside (COG Vc) se presenta tanto en el núcleo como en la cubierta pangenómica de Iridoviridae y Ascoviridae y tiene aparentemente un origen viral (Fig. 10). El grupo de ortólogos de esta proteína viral presenta cierto grado de similitud para los propios megavirus, es decir, no se encontraron ortólogos de otras familias de virus de dsDNA (Fig. 11). Estas secuencias homólogas de Ascoviridae e Iridoviridae forman un clado con Marseilleviridae, las de Phycodnaviridae forman un clado propio y las de Mimiviridae forman otro junto a secuencias de eucariontes. El par de ortólogos celulares encontrados junto a los mimivirus pertenecen a un protista (amoeba) y a un animal (hidra), ambos eucariontes de agua dulce; y cuya anotación para ambos se refiere a secuencias parecidas a la cápside (*capsid-like*). Las secuencias ortólogas de la cápside de megavirus de este análisis pertenecen al dominio 1 (D1), el cual, tiene a su vez, está compuesto por el dominio *jelly-roll* (plegamiento que consiste en cuatro pares de hojas beta antiparalelas). El alineamiento múltiple de la Figura 11 muestra varios fragmentos de secuencias con aminoácidos conservados. En estas cinco regiones de la D1 se encuentran en, primer lugar, aminoácidos que rompen la estructura secundaria (glicina y prolina); luego, aminoácidos polares con carga neutra (glutamina, serina, treonina y asparagina) y aminoácidos con carga negativa (ácido aspártico y ácido glutámico) y, finalmente, aminoácidos hidrofóbicos (alanina, leucina, isoleucina, valina, metionina, fenilalanina y triptófano). Cabe señalar que estas zonas, dentro de la estructura tridimensional de la cápside, se encuentran en las conexiones del *jelly-roll* (asas).



horizontalmente a eucariontes (E, en azul) como a amebas (hospedero de *Marseilleviridae* y *Mimiviridae*) y a cnidarios. El código de tres letras de los eucariontes pertenece al *KEGG*. Ani, animales; Pro, protistas. Uno de los dos dominios (D1) de la cápside se representa aquí con una línea azul horizontal. Las cinco regiones marcadas en recuadros azules pertenecen a las conexiones entre las hojas beta del *jelly-roll*. Los aminoácidos están coloreados de acuerdo a su función. Aminoácidos polares sin carga, en verde; aminoácidos hidrofóbicos, en morado; aminoácidos con carga negativa, en rosa; aminoácidos con carga positiva, en rojo; tirosina y histidina, en turquesa; glicina, en naranja; prolina, en amarillo; cisteína, en salmón.

### **3.6 Análisis filogenéticos basados en la estructura terciaria**

Cabe destacar que el análisis filogenético basado en secuencias expuesto con anterioridad, tiene la limitante de tener poca información evolutiva, principalmente, para homólogos distantes como se pudo ver en las filogenias con la RNAPol dependiente de DNA de virus y la D1 de la cápside de virus de dsDNA. Es por ello que nuestro laboratorio ha estandarizado una técnica complementaria al análisis filogenético basado en secuencias y que se basa en la comparación de estructuras cristalográficas para determinar la relación evolutiva de ortólogos distantes como pueden ser todas las RNA polimerasas (Jácome et al, 2015). En el ANEXO II se muestran los resultados y la discusión del estudio comparativo de las estructuras tridimensionales de la RNA polimerasas dependientes de RNA (RdRp) del virus de la hepatitis C (VHC) y del virus de inmunodeficiencia humana (VIH). Ambas estructuras cuentan con dominios catalíticos y estructurales similares (dominios palma, dedos y pulgar) para interactuar con el RNA viral.

Por otro lado, nuestro grupo de trabajo está analizando la historia evolutiva del *jelly-roll* de las cápsides icosaédricas de virus de RNA. La comparación basada en las secuencias de la D1 no mostró homólogos celulares. Sin embargo, el análisis comparativo de estructuras cristalográficas de dicha proteína en virus de RNA ha mostrado que existen ortólogos celulares distantes, principalmente, de proteínas membranales que están glicosiladas (resultados preliminares y no mostrados aquí).

Para esta sección, véase el ANEXO II (artículo por publicarse) para verificar los resultados de la comparación de las RdRp de virus de RNA.



## IV. DISCUSIÓN

El descubrimiento y el estudio de nuevos virus, la curación de base de datos sobre sus datos biológicos y ecológicos, el número creciente de proteomas de referencia virales y el diseño de algoritmos más sofisticados que requieren menos tiempo y memoria computacional han abierto la oportunidad para re-evaluar las estrategias para analizar y comprender no tan solo las diferencias funcionales y filogenéticas entre los virus, sino también su mismo origen y evolución temprana. No se puede analizar la historia evolutiva de los virus sin contrastarla con la filogenia de sus hospederos. Asimismo, muchos de los estudios se han concentrado en marcadores filogenéticos que intervienen en procesos de síntesis del DNA, transcripción y la formación de la cápside en general, como en el caso particular de los megavirus (Lakshminarayan M. Iyer et al., 2006; Koonin et al., 2006; Kristensen et al., 2013). Esto ha dado lugar a que se analice sólo una parte de la historia evolutiva de un grupo de genes ortólogos, y que ésta se extrapole a la historia evolutiva de toda una entidad biológica como puede ser un grupo viral sin considerar tiempos de divergencia que pueden ser cortos y su polimorfismo genético (Pamilo & Nei, 1988). Por otro lado, la dificultad para interpretar y extrapolar los análisis filogenéticos basados en secuencias es una limitante para el dilucidar el origen y la evolución temprana de los virus. El número creciente de estructuras cristalográficas virales y nuevas metodologías para el estudio de su evolución permitirán hacer estudios de filogenias profundas de los virus. Es por ello que en el presente trabajo se realizaron tres estrategias para tratar de dilucidar el origen y evolución temprana de los virus.

En primer lugar, el análisis de los datos biológicos y ecológicos de los virus permitieron concluir que la “simplicidad” de los virus de RNA no está correlacionada con la antigüedad de sus hospederos. Además, las más de 50 familias de virus de RNA sólo infectan principalmente a vertebrados, invertebrados, plantas y hongos. Existen dos familias de virus de RNA que sólo infectan a procariontes. Los Cystoviridae (dsRNA) y los Leviviridae (+ssRNA) infectan a proteobacterias como *Salmonella*, *Escherichia* y *Pseudomonas*, entre otras. Dichas bacterias están muy relacionadas a la microbiota de los animales. Ambas familias de virus de RNA pudieron haber evolucionado a partir de virus de RNA de eucariontes (Reoviridae) y adaptarse, de manera secundaria, a dichos procariontes patógenos de animales. Además, si bien puede ser un sesgo en las bases de datos actuales, no existen virus de RNA que infecten a Archaea. Por lo tanto, los virus de RNA tienen una estrecha relación evolutiva con los eucariontes, es

decir, los virus de RNA tienen un origen muy reciente que contradice la hipótesis virocéntrica que los coloca en etapas más tempranas de la vida, inclusive, en el mundo del RNA/proteínas (Koonin et al, 2006). Para una mayor discusión, véase ANEXO I.

En segundo lugar, el estudio de la pangenómica viral puede ayudar a interpretar cómo se van originando de manera individual los grupos de ortólogos que se van agrupando en la núcleo, la cubierta y la nube pangenómica. En el primer nivel del pangenoma, el núcleo, están las proteínas más conservadas entre la mayoría de los individuos virales de una misma familia. Aquí se encuentran todos los marcadores filogenéticos que forman parte de un virus, como las polimerasas y las proteínas que permiten la formación de la cápside. En el segundo nivel, la cubierta, están aquellas proteínas que si bien son esenciales, no se comparten entre todas las especies de una misma familia viral. Finalmente, en el tercer nivel, en la nube, están todas las proteínas que son únicas para cada especie viral y que no se comparten entre toda la familia. Estas proteínas son, por lo regular, de origen reciente y han divergido tan rápidamente que no presentan homólogos celulares a nivel de estructura primaria. Es por ello que en una primera etapa se estudiaron los genomas de los megavirus como un caso particular. Se empezó por este grupo dada la gran controversia sobre su origen, ya que algunos autores los posicionan en el origen mismo del último ancestro común de los seres vivos como un cuarto dominio de la vida (Boyer et al., 2010). Las siete proteínas que están muy conservadas en este grupo y que han puesto a los megavirus como un clado monofilético en el árbol de la vida son la helicasa de la superfamilia II (NCVOG0076), el factor de transcripción A2L-like (NCVOG0262), la subunidad alfa de la RNA polimerasa (NCVOG0271), la enzima capping del RNA mensajero, la ATPasa A32 (NCVOG0249), la subunidad pequeña de la ribonucleótido reductasa (NCVOG0276), la proteína de envoltura miristilada (NCVOG0211), la helicasa-primasa (NCVOG0023), y la DNA polimerasa (NCVOG0038) (Yutin et al., 2009).

En nuestro estudio pangenómico encontramos que la mayoría de las proteínas anteriores se encuentra en el núcleo y en la cubierta de cada una de las familias de megavirus. Una de las proteínas que encontramos en nuestro repertorio pangenómico y que ha servido para reconstruir filogenias de megavirus en los mencionados estudios previos fue la RNA polimerasa dependiente de DNA tipo II (RNAPol II) que es la que se encarga de transcribir el DNA en RNA mensajero en eucariontes (Young, 1991). Particularmente, la subunidad grande de la RNAPol II, la RPB1, se encontró en la mayoría de los núcleos pangenómicos de los megavirus. La RPB1 posee actividad catalítica y contiene el dominio carboxilo terminal (CTD) compuesto por repeticiones en heptapéptidos en tándem, el cual, tiene la particularidad de



estar conservado en los linajes más recientes de los eucariontes (desde hongos a mamíferos) (Hsin & Manley, 2012). De acuerdo a la Figura 7, la RBP1 ya estaba en el ancestro común de todos los megavirus y muy probablemente tiene un origen muy reciente en el ancestro común de los eucariontes, mientras que la RBP1 de los *Pandoravirus* y *Phaeocystis globosa virus* (Phycodnaviridae) tiene su origen mucho más reciente en los *Microsporidia* (hongos) y *Entamoeba histolytica* (un protista), respectivamente, ambos patógenos intracelulares de animales.

Otra de las proteínas que encontramos en varios núcleos pangénomicos de los megavirus y con la que se han construido filogenias para determinar que estos virus gigantes forman un cuarto dominio de la vida fue la ribonucleótido reductasa (RnR). Esta enzima permite la reducción de precursores de ribonucleótidos a desoxirribonucleótidos por lo que se ha sugerido que ésta y la timidilato sintasa (produce desoxitimidina 5' monofosfato a partir de desoxiuridina 5' monofosfato) han sido colocadas en la etapa de transición del mundo del RNA al de DNA (P. Forterre, 2002). Sin embargo, nuestros resultados mostraron que la historia evolutiva de la subunidad pequeña de la RnR (sRnR) de Poxviridae, Phycodnaviridae, Marseilleviridae y Mimiviridae ha coevolucionado con la de sus hospederos eucariontes (animales, algas y otros protistas). Además, nuestros resultados indicaron que la filogenia de la sRnR de Iridoviridae y Ascoviridae, familias que infectan a invertebrados y vertebrados de linajes antiguos, tiene dos historias independientes (Fig. 8). Por un lado, la sRnR de los ascovirus que infectan a polillas de la familia Noctuidae tiene un origen en insectos, mientras que la sRnR de los iridovirus que infectan al pez mandarín de agua dulce (*Siniperca chuatsi*) tiene una historia compartida con la de los alloverpesvirus (que infectan a peces y a anfibios) y probablemente tiene un origen en los animales acuáticos como los nemátodos y Fungi. Por otro lado, la sRnR de los ascovirus que infectan a avispas parasitoides (*Diadromus spp.*) tiene un origen en los iridovirus que infectan a insectos y vertebrados. Estos últimos tienen una sRnR ligada a la historia evolutiva de fagos que infectan proteobacterias y a la de las Alphaproteobacteria mismas. Por lo tanto, esta sRnR podría tener un origen bacteriano. Esta misma subunidad pequeña, en particular de la RnR tipo IA, es el sitio catalítico de toda la proteína y contiene un centro diférrico que lo caracteriza para unirse al oxígeno e iniciar la síntesis de desoxirribonucleótidos (Torrents, Aloy, Gibert, & Rodríguez-Trelles, 2002), es decir, la historia evolutiva de la proteína está relacionada con la gran oxigenación terrestre hace 2,500 millones de años. Por lo tanto, esta proteína puede ser un indicador geológico que posiciona a todos los organismos que la contienen

(Proteobacteria y eucariontes) alrededor de ese evento, por lo que los megavirus no pudieron haber surgido antes de dicho evento biogeoquímico. Lo anterior es una evidencia complementaria para descartar la idea de que existe un cuarto dominio de la vida derivado del último ancestro común de los seres vivos hace más de 3,500 millones de años (Colson et al., 2011). Por otro lado, y de acuerdo a nuestro análisis pangenómico, tanto el grupo de ortólogos de RnR como los de la deoxinucleósido cinasa, entre otras proteínas que intervienen en procesos metabólicos (COG F), esencialmente tienen un origen celular (85% para Iridoviridae y 100% para Ascoviridae en la Fig. 10). Sin embargo, no es muy claro si el grupo de ortólogos 652 de Iridoviridae que representa a la deaminasa de citidina (PF00383), una enzima que cataliza la hidrólisis de citidina en uridina y amoníaco (A. Lazcano, Guerrero, Margulis, & Oró, 1988), tiene su origen en una Betaproteobacteria porque la filogenia sólo presenta cuatro ramas (ver filogenia en el la liga de iTOL). De todas formas, ningún megavirus hasta ahora analizado tiene grupos de ortólogos que constituyan alguna ruta metabólica completa. Sólo son dominios que seguramente fueron adquiridos por transferencia horizontal de sus hospederos de los cuales no intervienen negativamente en su adecuación biológica.

También se encontraron proteínas estructurales en la base de datos pangenómica que se comparten entre los megavirus. Por ejemplo, en los Iridoviridae y en los Ascoviridae se encontró que el dominio de la familia Erv1/Alr (COG C) se encuentra distribuido principalmente en la nube de ambas familias. Esta familia proteica, junto con la oxidoreductasa Mia40, interviene en el proceso de formación de uniones disulfuro para la importación de proteínas en el espacio intermembranal de la mitocondria (Fischer & Riemer, 2013) y, por lo tanto, está distribuida solamente en eucariontes. Otro de los grupos de ortólogos que intervienen en procesos celulares para la formación de membrana fue la familia GNS1/SUR4 (COG I) (Fig. 9). Este tipo de proteínas membranales intervienen en procesos de formación de cadenas largas de ácidos grasos (hasta 26 átomos de carbono) como precursores para la biosíntesis de ceramida y esfingolípidos exclusivamente de membranas eucariontes (Tvrdik et al., 2000). Nuestros resultados mostraron que el grupo de ortólogos de la GNS1/SUR4 de Ascoviridae, *Pigeonpoxvirus* y de un phycodnavirus tiene su origen en sus respectivos hospederos (insectos, aves y protistas).

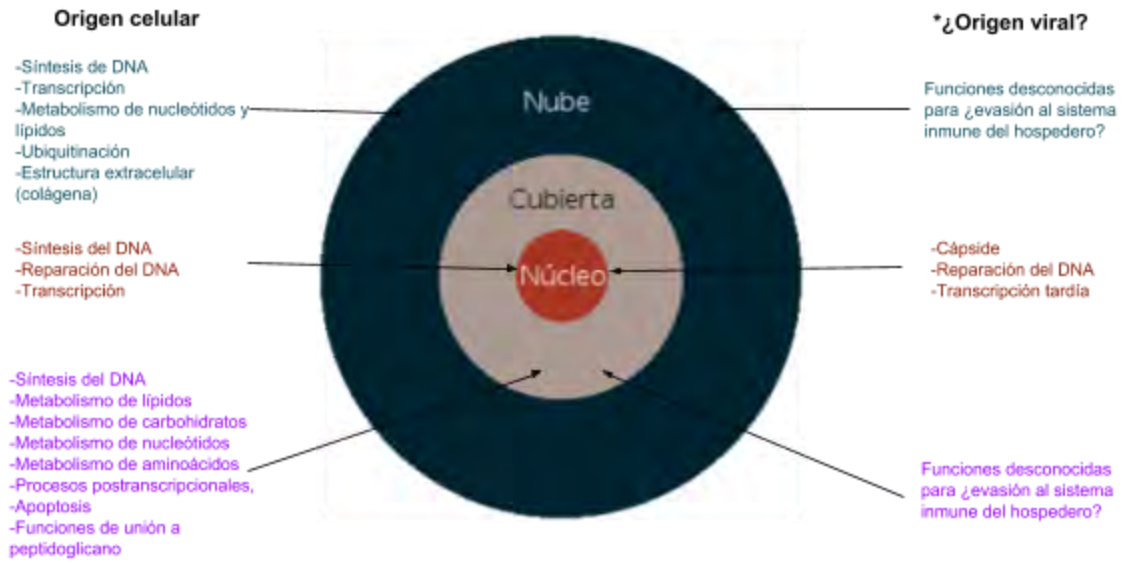
Finalmente, en tercer lugar, estrategias como los estudios filogenéticos basados en estructuras terciarias han permitido hacer un análisis más profundo sobre el origen mismo de los virus como ha sido el caso de las RNA polimerasas de virus (Jácome et al., 2015), en particular, del VHC y del VIH (véase discusión del artículo sobre el tema en el ANEXO II). Otra

proteína estructural que se encontró principalmente en los núcleos pangenómicos de los megavirus fue la de la cápside. Reconstruir la filogenia de la cápside basada en secuencias provee poca información evolutiva debido a que éstas son extremadamente divergentes y tienen múltiples parálogos (Fig 11). La proteína de cápside de megavirus sólo tiene homólogos dentro del mismo clado de los virus gigantes y no en otros grupos virales de dsDNA. Existen ortólogos en cnidarios y protistas, organismos que comparten el ambiente con su hospedero ameboide, pero evidentemente fueron transferidos horizontalmente de mimivirus a estos eucariontes. No es de extrañarse que nuestros resultados (Fig. 11), como en diversos trabajos donde se analiza la estructura primaria, hayan mostrado que la proteína de cápside (Vc) tenga un origen viral. De hecho, se ha mencionado que las proteínas de cápside son gen únicos de virus (*hallmark genes*) que no tienen ortólogos celulares (Koonin et al., 2006). Sin embargo, en estudios previos (Krupovic & Koonin, 2017) y en resultados preliminares de nuestro laboratorio sobre el estudio del dominio *jelly-roll* de cápsides icosaédricas de virus de RNA, han demostrado que a través del análisis comparativo de estructuras terciarias es posible que dicha proteína tenga un origen celular. Su origen podría estar relacionado con proteínas celulares que presentan el dominio *jelly-roll* y que se ubican en espacios transmembranales y que tienen funciones de unión a carbohidratos y de regulación del sistema inmunitario como el factor de necrosis tumoral (Bodmer, Schneider, & Tschopp, 2002). Es decir, es muy probable que a través de la comparación de las estructuras cristalográficas de las proteínas virales que han sido catalogadas como únicas de los virus (*hallmark genes* como la RNA polimerasa dependiente de RNA, helicasas de la superfamilia 3, endonucleasas, proteínas empacadoras del DNA, entre otras; véase Koonin et al, 2006) o aquéllas con función desconocida (COG S y R), podrían ser ortólogos distantes de células y que se han originado a partir de los genomas de sus hospederos.

En la Fig. 12 podemos ver un modelo sobre el origen de algunas funciones en proteomas de Iridoviridae y Ascoviridae. Al menos para estas dos familias de megavirus, nuestro trabajo indica que la mayoría de las funciones que intervienen en procesos centrales y que se encuentran en el núcleo genómico provienen de genomas de los hospederos correspondientes y, por lo tanto, son más antiguas con respecto a las de un posible origen viral. Por otro lado, aquellas funciones que intervienen en procesos de evasión del sistema inmune innato o que son putativas o desconocidas se encuentran principalmente en la nube y podrían tener un origen viral reciente. Sin embargo, existen proteínas que intervienen en la

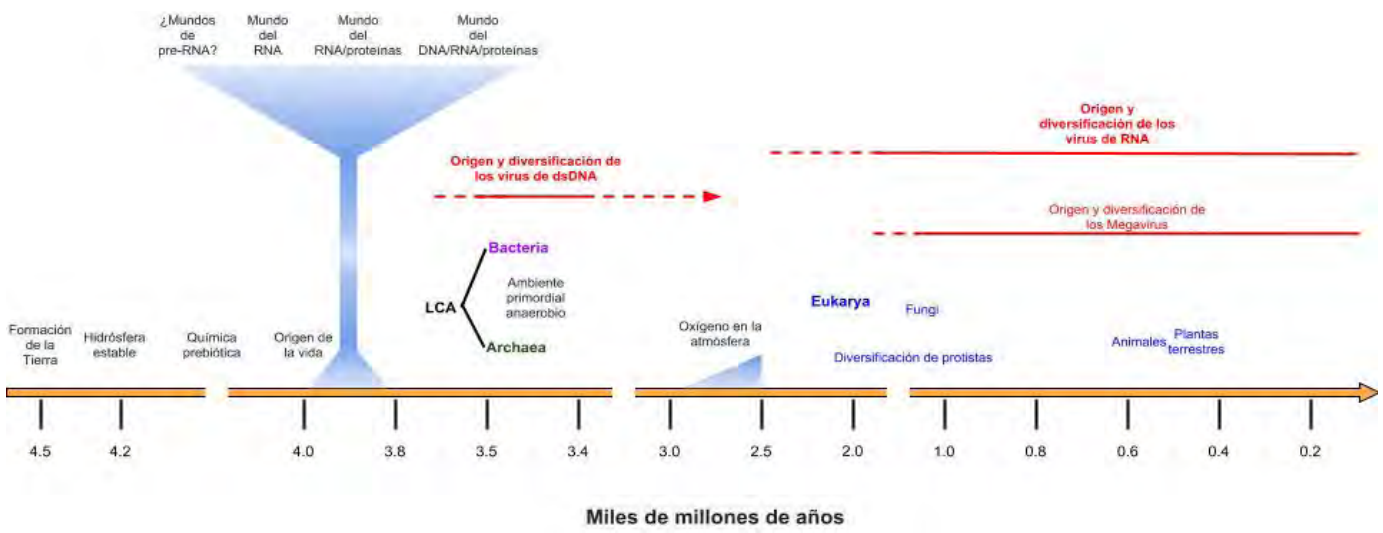
morfogénesis de la cápside y que se encuentra también en el núcleo genómico que podrían explicar su origen celular a través del estudio de las estructuras las estructuras terciarias.

De lo anteriormente expuesto sobre los datos biológicos analizados de todos los virus registrados en el GenBank en la actualidad, los estudios pangenómicos de algunas familias de virus de dsDNA (megavirus) y algunos resultados preliminares sobre la comparación de la estructura terciaria de la RNA polimerasa y el dominio jelly-roll de las cápsides de virus de RNA, se podría inferir que es muy posible que los virus surgieron mucho después del Mundo del RNA. También se podría concluir que los primeros virus de DNA, como pueden ser los fagos, surgieron durante o después del origen del último ancestro común de los seres vivos (LCA). Por otro lado, los virus de RNA surgieron a partir de genomas de sus hospederos eucariontes y los virus de RNA que infectan a Proteobacteria surgieron probablemente a partir de virus eucariontes que infectan la microbiota de animales. Finalmente, los megavirus surgieron probablemente a partir de los eucariontes de linajes antiguos como pueden ser los protistas como las algas (Fig. 13).



\* Análisis con estructuras terciarias han demostrado que dominios como el jelly roll de virus icosaédricos aparentemente tiene un origen celular (Krupovic y Koonin, 2017; resultados preliminares de nuestro laboratorio).

**Fig. 12** Modelo sobre el origen de los grupos de ortólogos del pangenoma de los Iridoviridae y Ascoviridae. Al menos para estas dos familias de megavirus (dsDNA) se aprecia que la mayoría de las funciones que intervienen en procesos para la transmisión de la información genética y para la síntesis de compuestos orgánicos (carbohidratos, lípidos y aminoácidos), se encuentran principalmente en el núcleo y tienen un origen celular. Por lo tanto, estas secuencias son más antiguas que aquellas que se encuentran en la nube. Por otro lado, la mayoría de las funciones que pueden evadir al sistema inmune innato de invertebrados y vertebrados son de origen viral, de novo y de un periodo reciente.



**Fig. 13** Modelo sobre el origen de los primeros virus de dsDNA, los megavirus y los virus de RNA.

## V. CONCLUSIONES

- El tamaño del genoma viral no está correlacionado con la historia evolutiva de los hospederos correspondientes. Es decir, genomas pequeños y “simples” como los de RNA virales no se encuentran distribuidos ampliamente en las ramas más antiguas del árbol de la vida, mientras que los genomas grandes y “complejos” de dsDNA sí lo están.
- No se conocen virus de RNA en Archaea, y las únicas dos familias virales con un genoma de RNA que infectan procariontes, Cystoviridae y Leviviridae, infectan a Proteobacteria que forman parte de la microbiota animal.
- Los virus de RNA parecen tener un origen relacionado a la historia evolutiva de los eucariontes.
- No todos los ortólogos de los megavirus que se encuentran en el núcleo pangenómico muestran que sean un solo clado.
- Se confirma lo que en otros dos estudios se ha concluido (Schulz et al., 2017; Yutin et al., 2014): los megavirus no forman un cuarto dominio de la vida derivado del último ancestro común de los seres vivos.
- El pangenoma de los virus es abierto y nuevos genes se están originando en la cubierta y en la nube, pero también, en el caso particular de Marseilleviridae y Mimiviridae, el núcleo presenta proteínas de novo que intervienen probablemente en la morfogénesis de la cápside.
- La subunidad pequeña de la ribonucleótido reductasa tipo la tiene un origen eucarionte (para los Poxviridae, Phycodnaviridae, Marseilleviridae y Mimiviridae) y un origen celular (para los Iridoviridae y Ascoviridae) ligado al evento geológico de la oxigenación terrestre.
- Las mismas filogenias de los grupos ortólogos del núcleo y del resto del pangenoma permitieron demostrar que las proteínas conservadas de los fagos quedan en un clado monofilético con las de procariontes.
- La implementación de una estrategia basada en estructuras terciarias permitirá tener un análisis filogenético más fino de las proteínas virales como el caso de la RNA polimerasa dependiente de RNA y la transcriptasa reversa.
- Estos análisis preliminares permiten concluir que el escenario más probable sobre el origen y evolución temprana de los virus es a través del escape de genes de sus hospederos correspondientes.

## VI. PERSPECTIVAS

Es importante recalcar que este trabajo, el origen y evolución temprana de los virus, tendrá continuidad ya que se ha establecido como una línea de investigación formal en el Laboratorio de Origen de la Vida de la Facultad de Ciencias, UNAM. Es por ello que a continuación se enuncian las siguientes perspectivas:

- Hacer un análisis pangenómico del resto de las 114 familias de RNA y de DNA restantes para determinar la historia evolutiva del repertorio proteómico de su propio núcleo, cubierta y nube.
- Analizar las secuencias homólogas del mobiloma (plásmidos, transposones, retrotransposones, virofagos) debido a que podrían explicar el mecanismo por el cual los virus se originan (Yutin, Shevchenko, Kapitonov, Krupovic, & Koonin, 2015).
- Construir una base de datos de todas las estructuras cristalográficas, al menos, de cada núcleo pangenómico para cada una de las familias y determinar su historia evolutiva a través de su comparación tridimensional.
- En este estudio no se analizaron los parálogos que se encontraron en varios genomas de megavirus. Muchos de ellos intervienen en la formación de cápsides. Sería interesante determinar si son parálogos que se duplicaron antes o después de la “especiación” de los virus de cada familia.
- Aquellas proteínas que tienen una función desconocida es más complicado determinar su filogenia. Sin embargo, el análisis de firmas moleculares en motivos y dominios estructurales de estas mismas, podría ser un indicador de su posible origen. Se espera que la mayoría de estas proteínas en las que no se ha identificado su función estén relacionadas principalmente a la transcripción, recombinación, reparación del DNA, empaquetamiento y morfogénesis de la cápside y a la evasión del sistema de defensa del hospedero.

## VII. REFERENCIAS

- Abroi, A., & Gough, J. (2011). Are viruses a source of new protein folds for organisms? - Virosphere structure space and evolution. *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology*, 33(8), 626–635.
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215(3), 403–410.
- Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17), 3389–3402.
- Alvarez-Carreño, C., Alva, V., Becerra, A., & Lazcano, A. (2018). Structure, function and evolution of the hemerythrin-like domain superfamily. *Protein Science: A Publication of the Protein Society*, 27(4), 848–860.
- Andrewes, C. H. (1963). Classification of Viruses of Vertebrates. In *Advances in Virus Research* (pp. 271–296).
- Assis, F. L., Bajrai, L., Abrahao, J. S., Kroon, E. G., Dornas, F. P., Andrade, K. R., ... Colson, P. (2015). Pan-Genome Analysis of Brazilian Lineage A Amoebal Mimiviruses. *Viruses*, 7(7), 3483–3499.
- Baltimore, D. (1971). Expression of animal virus genomes. *Bacteriological Reviews*, 35(3), 235–241.
- Barrangou, R. (2015). The roles of CRISPR–Cas systems in adaptive immunity and beyond. *Current Opinion in Immunology*, 32, 36–41.
- Becerra, A., 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(1), 361–379.
- Belyi, V. A., Levine, A. J., & Skalka, A. M. (2010). Sequences from ancestral single-stranded DNA viruses in vertebrate genomes: the parvoviridae and circoviridae are more than 40 to 50 million years old. *Journal of Virology*, 84(23), 12458–12462.
- Beutner, R. (1938). *Life's Beginning on the Earth*.
- Bodmer, J.-L., Schneider, P., & Tschopp, J. (2002). The molecular architecture of the TNF superfamily. *Trends in Biochemical Sciences*, 27(1), 19–26.
- Boyer, M., Gimenez, G., Suzan-Monti, M., & Raoult, D. (2010). Classification and determination of possible origins of ORFans through analysis of nucleocytoplasmic large DNA viruses. *Intervirology*, 53(5), 310–320.
- Boyer, M., Madoui, M.-A., Gimenez, G., La Scola, B., & Raoult, D. (2010). Phylogenetic and phyletic studies of informational genes in genomes highlight existence of a 4 domain of life including giant viruses. *PLoS One*, 5(12), e15530.
- Boyer, M., Yutin, N., Pagnier, I., Barrassi, L., Fournous, G., Espinosa, L., ... Raoult, D. (2009). Giant Marseillevirus highlights the role of amoebae as a melting pot in emergence of chimeric microorganisms. *Proceedings of the National Academy of Sciences of the United States of America*, 106(51), 21848–21853.
- Brandes, N., & Linial, M. (2016). Gene overlapping and size constraints in the viral world. *Biology Direct*, 11, 26.
- Breitbart, M., & Rohwer, F. (2005). Here a virus, there a virus, everywhere the same virus? *Trends in Microbiology*, 13(6), 278–284.
- Brito, A. F. de, Braconi, C. T., Weidmann, M., Dilcher, M., Alves, J. M. P., Gruber, A., & Zanotto, P. M. de A. (2015). The Pangenome of the *Anticarsia gemmatalis* Multiple Nucleopolyhedrovirus (AgMNPV). *Genome Biology and Evolution*, 8(1), 94–108.

- Bushman, F. (2002). *Lateral DNA Transfer: Mechanisms and Consequences*. CSHL Press.
- Cadwell, K. (2015). Expanding the role of the virome: commensalism in the gut. *Journal of Virology*, 89(4), 1951–1953.
- Campillo-Balderas, J. A., Lazcano, A., & Becerra, A. (2015). Viral Genome Size Distribution Does not Correlate with the Antiquity of the Host Lineages. *Frontiers in Ecology and Evolution*, 3. <https://doi.org/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.
- Černý, J., Černá Bolfíková, B., de A Zanotto, P. M., Grubhoffer, L., & Růžek, D. (2015). A deep phylogeny of viral and cellular right-hand polymerases. *Infection, Genetics and Evolution: Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases*, 36, 275–286.
- Chothia, C. (2003). Evolution of the Protein Repertoire. *Science*, 300(5626), 1701–1703.
- Colson, P., de Lamballerie, X., Fournous, G., & Raoult, D. (2012). Reclassification of giant viruses composing a fourth domain of life in the new order Megavirales. *Intervirology*, 55(5), 321–332.
- Colson, P., Gimenez, G., Boyer, M., Fournous, G., & Raoult, D. (2011). The giant Cafeteria roenbergensis virus that infects a widespread marine phagocytic protist is a new member of the fourth domain of Life. *PloS One*, 6(4), e18935.
- Contreras-Moreira, B., Cantalapiedra, C. P., García-Pereira, M. J., Gordon, S. P., Vogel, J. P., Igartua, E., ... Vinuesa, P. (2017). Analysis of Plant Pan-Genomes and Transcriptomes with GET\_HOMOLOGUES-EST, a Clustering Solution for Sequences of the Same Species. *Frontiers in Plant Science*, 8, 184.

- Contreras-Moreira, B., & Vinuesa, P. (2013). GET\_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis. *Applied and Environmental Microbiology*, 79(24), 7696–7701.
- Csorba, T., Pantaleo, V., & Burgyán, J. (2009). RNA Silencing: An Antiviral Mechanism. In *Natural and Engineered Resistance to Plant Viruses, Part I* (Vol. 75, pp. 35–230). Elsevier.
- d'Herrelle, F., & Smith, G. H. (1926). The Bacteriophage and Its Behavior. *The Science News-Letter*, 9(274), 10.
- Domingo, E. (2015). *Virus as Populations: Composition, Complexity, Dynamics, and Biological Implications*. Academic Press.
- Doolittle, W. F. (2000). The nature of the universal ancestor and the evolution of the proteome. *Current Opinion in Structural Biology*, 10(3), 355–358.
- Drake, J. W. (1993). Rates of spontaneous mutation among RNA viruses. *Proceedings of the National Academy of Sciences*, 90(9), 4171–4175.
- Drake, J. W., & Hwang, C. B. C. (2005). On the mutation rate of herpes simplex virus type 1. *Genetics*, 170(2), 969–970.
- Duffy, S., & Holmes, E. C. (2009). Validation of high rates of nucleotide substitution in geminiviruses: phylogenetic evidence from East African cassava mosaic viruses. *The Journal of General Virology*, 90(Pt 6), 1539–1547.
- Eddy, S. R. (1998). Profile hidden Markov models. *Bioinformatics*, 14(9), 755–763.
- Eddy, S. R. (2009). A new generation of homology search tools based on probabilistic inference. *Genome Informatics. International Conference on Genome Informatics*, 23(1), 205–211.
- Edwards, R. A., & Rohwer, F. (2005). Viral metagenomics. *Nature Reviews. Microbiology*, 3(6), 504–510.

- Eigen, M., McCaskill, J., & Schuster, P. (1988). Molecular quasi-species. *The Journal of Physical Chemistry*, 92(24), 6881–6891.
- Filée, J., Forterre, P., Sen-Lin, T., & Laurent, J. (2002). Evolution of DNA polymerase families: evidences for multiple gene exchange between cellular and viral proteins. *Journal of Molecular Evolution*, 54(6), 763–773.
- Finn, R. D., Tate, J., Mistry, J., Coggill, P. C., Sammut, S. J., Hotz, H.-R., ... Bateman, A. (2008). The Pfam protein families database. *Nucleic Acids Research*, 36(Database issue), D281–D288.
- Fischer, M., & Riemer, J. (2013). The mitochondrial disulfide relay system: roles in oxidative protein folding and beyond. *International Journal of Cell Biology*, 2013, 742923.
- Fitch, W. M. (2000). Homology a personal view on some of the problems. *Trends in Genetics: TIG*, 16(5), 227–231.
- Flint, J., Rall, G. F., Racaniello, V. R., & Skalka, A. M. (2015). *Principles of Virology, Volume II: Pathogenesis & Control*.
- Flores, R., Gago-Zachert, S., Serra, P., Sanjuán, R., & Elena, S. F. (2014). Viroids: survivors from the RNA world? *Annual Review of Microbiology*, 68, 395–414.
- Forterre, P. (2002). The origin of DNA genomes and DNA replication proteins. *Current Opinion in Microbiology*, 5(5), 525–532.
- Forterre, P. (2006). The origin of viruses and their possible roles in major evolutionary transitions. *Virus Research*, 117(1), 5–16.
- Fortier, L.-C., & Sekulovic, O. (2013). Importance of prophages to evolution and virulence of bacterial pathogens. *Virulence*, 4(5), 354–365.
- Fu, L., Niu, B., Zhu, Z., Wu, S., & Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, 28(23), 3150–3152.

- Gerstein, M., & Hegyi, H. (1998). *Comparing Genomes in Terms of Protein Structure: Surveys of a Finite Parts List*. <https://doi.org/10.21236/ada472206>
- Gibbs, A. J., & Calisher, C. H. (2005). *Molecular Basis of Virus Evolution*. Cambridge University Press.
- Gorbalenya, A. E., Pringle, F. M., Zeddani, J.-L., Luke, B. T., Cameron, C. E., Kalkmakoff, J., ... Ward, V. K. (2002). The palm subdomain-based active site is internally permuted in viral RNA-dependent RNA polymerases of an ancient lineage. *Journal of Molecular Biology*, *324*(1), 47–62.
- Grinde, B. (2013). Herpesviruses: latency and reactivation - viral strategies and host response. *Journal of Oral Microbiology*, *5*. <https://doi.org/10.3402/jom.v5i0.22766>
- Grove, J., & Marsh, M. (2011). The cell biology of receptor-mediated virus entry. *The Journal of Cell Biology*, *195*(7), 1071–1082.
- Hacker, J., & Dobrindt, U. (2006). *Pathogenomics: Genome Analysis of Pathogenic Microbes*. John Wiley & Sons.
- Holmes, E. C. (2009). *The Evolution and Emergence of RNA Viruses*. Oxford University Press.
- Hsin, J.-P., & Manley, J. L. (2012). The RNA polymerase II CTD coordinates transcription and RNA processing. *Genes & Development*, *26*(19), 2119–2137.
- Iyer, L. M., Aravind, L., & Koonin, E. V. (2001). Common origin of four diverse families of large eukaryotic DNA viruses. *Journal of Virology*, *75*(23), 11720–11734.
- Iyer, L. M., Balaji, S., Koonin, E. V., & Aravind, L. (2006). Evolutionary genomics of nucleo-cytoplasmic large DNA viruses. *Virus Research*, *117*(1), 156–184.
- Jácome, R., Becerra, A., Ponce de León, S., & Lazcano, A. (2015). Structural Analysis of Monomeric RNA-Dependent Polymerases: Evolutionary and Therapeutic Implications. *PLoS One*, *10*(9), e0139001.

- Jalasvuori, M., & Bamford, J. K. H. (2008). Structural co-evolution of viruses and cells in the primordial world. *Origins of Life and Evolution of the Biosphere: The Journal of the International Society for the Study of the Origin of Life*, 38(2), 165–181.
- Kaas, R. S., Friis, C., Ussery, D. W., & Aarestrup, F. M. (2012). Estimating variation within the genes and inferring the phylogeny of 186 sequenced diverse *Escherichia coli* genomes. *BMC Genomics*, 13, 577.
- Katoh, K., Misawa, K., Kuma, K.-I., & Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, 30(14), 3059–3066.
- King, A. M. Q., Adams, M. J., & Lefkowitz, E. J. (2011). *Virus Taxonomy: Classification and Nomenclature of Viruses : Ninth Report of the International Committee on Taxonomy of Viruses*. Elsevier.
- Koonin, E. V., & Dolja, V. V. (2006). Evolution of complexity in the viral world: The dawn of a new vision. *Virus Research*, 117(1), 1–4.
- 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, 10–24.
- Koonin, E. V., Senkevich, T. G., & Dolja, V. V. (2006). The ancient Virus World and evolution of cells. *Biology Direct*, 1, 29.
- Koonin, E. V., & Starokadomskyy, P. (2016). Are viruses alive? The replicator paradigm sheds decisive light on an old but misguided question. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 59, 125–134.
- Koonin, E. V., & Yutin, N. (2010). Origin and evolution of eukaryotic large nucleo-cytoplasmic

- DNA viruses. *Intervirology*, 53(5), 284–292.
- Krissinel, E., & Henrick, K. (2004). Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallographica. Section D, Biological Crystallography*, 60(Pt 12 Pt 1), 2256–2268.
- Kristensen, D. M., Kannan, L., Coleman, M. K., Wolf, Y. I., Sorokin, A., Koonin, E. V., & Mushegian, A. (2010). A low-polynomial algorithm for assembling clusters of orthologous groups from intergenomic symmetric best matches. *Bioinformatics*, 26(12), 1481–1487.
- Kristensen, D. M., Waller, A. S., Yamada, T., Bork, P., Mushegian, A. R., & Koonin, E. V. (2013). Orthologous gene clusters and taxon signature genes for viruses of prokaryotes. *Journal of Bacteriology*, 195(5), 941–950.
- Krupovič, M., & Bamford, D. H. (2007). Putative prophages related to lytic tailless marine dsDNA phage PM2 are widespread in the genomes of aquatic bacteria. *BMC Genomics*, 8(1), 236.
- Krupovic, M., & Koonin, E. V. (2017). Multiple origins of viral capsid proteins from cellular ancestors. *Proceedings of the National Academy of Sciences of the United States of America*, 114(12), E2401–E2410.
- Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., ... Marra, M. A. (2009). Circos: an information aesthetic for comparative genomics. *Genome Research*, 19(9), 1639–1645.
- Lam, T. T.-Y., Hon, C.-C., & Tang, J. W. (2010). Use of phylogenetics in the molecular epidemiology and evolutionary studies of viral infections. *Critical Reviews in Clinical Laboratory Sciences*, 47(1), 5–49.
- Lazcano, A. (2010). Origen y evolución de los virus: ¿genes errantes o parásitos primitivos? *Mensaje Bioquímico*, XXXIV, 73–84.
- Lazcano, A., Guerrero, R., Margulis, L., & Oró, J. (1988). The evolutionary transition from RNA

- to DNA in early cells. *Journal of Molecular Evolution*, 27(4), 283–290.
- Le Gall, O., Christian, P., Fauquet, C. M., King, A. M. Q., Knowles, N. J., Nakashima, N., ... Gorbalenya, A. E. (2008). Picornavirales, a proposed order of positive-sense single-stranded RNA viruses with a pseudo-T = 3 virion architecture. *Archives of Virology*, 153(4), 715–727.
- Legendre, M., Bartoli, J., Shmakova, L., Jeudy, S., Labadie, K., Adrait, A., ... Claverie, J.-M. (2014). Thirty-thousand-year-old distant relative of giant icosahedral DNA viruses with a pandoravirus morphology. *Proceedings of the National Academy of Sciences of the United States of America*, 111(11), 4274–4279.
- Letunic, I., & Bork, P. (2016). Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Research*, 44(W1), W242–W245.
- McCormack, G. P., & Clewley, J. P. (2002). The application of molecular phylogenetics to the analysis of viral genome diversity and evolution. *Reviews in Medical Virology*, 12(4), 221–238.
- McDonnell, S. J., Sparger, E. E., & Murphy, B. G. (2013). Feline immunodeficiency virus latency. *Retrovirology*, 10, 69.
- McLysaght, A., Baldi, P. F., & Gaut, B. S. (2003). Extensive gene gain associated with adaptive evolution of poxviruses. *Proceedings of the National Academy of Sciences of the United States of America*, 100(26), 15655–15660.
- Medini, D., Donati, C., Tettelin, H., Massignani, V., & Rappuoli, R. (2005). The microbial pan-genome. *Current Opinion in Genetics & Development*, 15(6), 589–594.
- Moreira, D., & López-García, P. (2009). Ten reasons to exclude viruses from the tree of life. *Nature Reviews. Microbiology*, 7(4), 306–311.
- Mount, D. W. (2004). *Bioinformatics: Sequence and Genome Analysis*. CSHL Press.



- Murray, N. E. (2002). 2001 Fred Griffith review lecture. Immigration control of DNA in bacteria: self versus non-self. *Microbiology*, 148(Pt 1), 3–20.
- Murzin, A. G., Brenner, S. E., Hubbard, T., & Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, 247(4), 536–540.
- Nasir, A., & Caetano-Anollés, G. (2015). A phylogenomic data-driven exploration of viral origins and evolution. *Science Advances*, 1(8), e1500527.
- Nasir, A., Kim, K. M., & Caetano-Anolles, G. (2012). Giant viruses coexisted with the cellular ancestors and represent a distinct supergroup along with superkingdoms Archaea, Bacteria and Eukarya. *BMC Evolutionary Biology*, 12, 156.
- Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, 32(1), 268–274.
- Pamilo, P., & Nei, M. (1988). Relationships between gene trees and species trees. *Molecular Biology and Evolution*, 5(5), 568–583.
- Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., & Ferrin, T. E. (2004). UCSF Chimera--a visualization system for exploratory research and analysis. *Journal of Computational Chemistry*, 25(13), 1605–1612.
- Philippe, N., Legendre, M., Doutre, G., Couté, Y., Poirot, O., Lescot, M., ... Abergel, C. (2013). Pandoraviruses: amoeba viruses with genomes up to 2.5 Mb reaching that of parasitic eukaryotes. *Science*, 341(6143), 281–286.
- Podolsky, S. (1996). The role of the virus in origin-of-life theorizing. *Journal of the History of Biology*, 29(1), 79–126.
- Rakhuba, D. V., Kolomiets, E. I., Dey, E. S., & Novik, G. I. (2010). Bacteriophage receptors,

- mechanisms of phage adsorption and penetration into host cell. *Polish Journal of Microbiology / Polskie Towarzystwo Mikrobiologow = The Polish Society of Microbiologists*, 59(3), 145–155.
- Raoult, D. (2004). The 1.2-Megabase Genome Sequence of Mimivirus. *Science*, 306(5700), 1344–1350.
- Raoult, D., & Forterre, P. (2008). Redefining viruses: lessons from Mimivirus. *Nature Reviews. Microbiology*, 6(4), 315–319.
- Reche, I., D'Orta, G., Mladenov, N., Winget, D. M., & Suttle, C. A. (2018). Deposition rates of viruses and bacteria above the atmospheric boundary layer. *The ISME Journal*, 12(4), 1154–1162.
- Riley, M., & Labedan, B. (1997). Protein evolution viewed through Escherichia coli protein sequences: introducing the notion of a structural segment of homology, the module. *Journal of Molecular Biology*, 268(5), 857–868.
- Rokas, A., Williams, B. L., King, N., & Carroll, S. B. (2003). Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature*, 425(6960), 798–804.
- Romero, P. (2004). Bioinformatics: Sequence and Genome Analysis. *Briefings in Bioinformatics*, 5(4), 393–396.
- Roossinck, M. J. (2012). Plant virus metagenomics: biodiversity and ecology. *Annual Review of Genetics*, 46, 359–369.
- Roossinck, M. J., & Bazán, E. R. (2017). Symbiosis: Viruses as Intimate Partners. *Annual Review of Virology*, 4(1), 123–139.
- Ryabov, E. V. (2017). Invertebrate RNA virus diversity from a taxonomic point of view. *Journal of Invertebrate Pathology*, 147, 37–50.
- Sanjuán, R., Nebot, M. R., Chirico, N., Mansky, L. M., & Belshaw, R. (2010). Viral mutation

- rates. *Journal of Virology*, 84(19), 9733–9748.
- Schulz, F., Yutin, N., Ivanova, N. N., Ortega, D. R., Lee, T. K., Vierheilig, J., ... Woyke, T. (2017). Giant viruses with an expanded complement of translation system components. *Science*, 356(6333), 82–85.
- Sobhy, H., Scola, B. L., Pagnier, I., Raoult, D., & Colson, P. (2015). Identification of giant Mimivirus protein functions using RNA interference. *Frontiers in Microbiology*, 6, 345.
- Subbiah, S., Laurents, D. V., & Levitt, M. (1993). Structural similarity of DNA-binding domains of bacteriophage repressors and the globin core. *Current Biology: CB*, 3(3), 141–148.
- Suttle, C. A. (2005). Viruses in the sea. *Nature*, 437(7057), 356–361.
- Tatusov, R. L. (1997). A Genomic Perspective on Protein Families. *Science*, 278(5338), 631–637.
- Tettelin, H., Masignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., ... Fraser, C. M. (2005). Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial “pan-genome.” *Proceedings of the National Academy of Sciences of the United States of America*, 102(39), 13950–13955.
- Torrents, E., Aloy, P., Gibert, I., & Rodríguez-Trelles, F. (2002). Ribonucleotide reductases: divergent evolution of an ancient enzyme. *Journal of Molecular Evolution*, 55(2), 138–152.
- Tvrđik, P., Westerberg, R., Silve, S., Asadi, A., Jakobsson, A., Cannon, B., ... Jacobsson, A. (2000). Role of a new mammalian gene family in the biosynthesis of very long chain fatty acids and sphingolipids. *The Journal of Cell Biology*, 149(3), 707–718.
- Vernikos, G., Medini, D., Riley, D., & Hervé, T. (2015). Ten years of pan-genome analyses. *Current Opinion in Microbiology*, 23, 148–154.
- Wagner, R. R. (1984). Cytopathic Effects of Viruses: A General Survey. In *Viral Cytopathology* (pp. 1–63).

- Wang, M., Yafremava, L. S., Caetano-Anollés, D., Mittenthal, J. E., & Caetano-Anollés, G. (2007). Reductive evolution of architectural repertoires in proteomes and the birth of the tripartite world. *Genome Research*, *17*(11), 1572–1585.
- Williamson, K. E., Radosevich, M., & Wommack, K. E. (2005). Abundance and diversity of viruses in six Delaware soils. *Applied and Environmental Microbiology*, *71*(6), 3119–3125.
- Woese, C. R., & Fox, G. E. (1977). Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proceedings of the National Academy of Sciences of the United States of America*, *74*(11), 5088–5090.
- Wu, D., Wu, M., Halpern, A., Rusch, D. B., Yooseph, S., Frazier, M., ... Eisen, J. A. (2011). Stalking the fourth domain in metagenomic data: searching for, discovering, and interpreting novel, deep branches in marker gene phylogenetic trees. *PLoS One*, *6*(3), e18011.
- Young, R. A. (1991). RNA polymerase II. *Annual Review of Biochemistry*, *60*, 689–715.
- Yutin, N., & Koonin, E. V. (2012). Hidden evolutionary complexity of Nucleo-Cytoplasmic Large DNA viruses of eukaryotes. *Virology Journal*, *9*, 161.
- Yutin, N., Shevchenko, S., Kapitonov, V., Krupovic, M., & Koonin, E. V. (2015). A novel group of diverse Polinton-like viruses discovered by metagenome analysis. *BMC Biology*, *13*, 95.
- Yutin, N., Wolf, Y. I., & Koonin, E. V. (2014). Origin of giant viruses from smaller DNA viruses not from a fourth domain of cellular life. *Virology*, *466-467*, 38–52.
- Yutin, N., Wolf, Y. I., Raoult, D., & Koonin, E. V. (2009). Eukaryotic large nucleo-cytoplasmic DNA viruses: clusters of orthologous genes and reconstruction of viral genome evolution. *Virology Journal*, *6*, 223.
- Zablocki, O., Adriaenssens, E. M., & Cowan, D. (2015). Diversity and Ecology of Viruses in Hyperarid Desert Soils. *Applied and Environmental Microbiology*, *82*(3), 770–777.

## **VIII. ANEXO I**

(Artículo publicado)

**Viral genome size distribution does not correlate with the antiquity of the host lineages**



# Viral Genome Size Distribution Does not Correlate with the Antiquity of the Host Lineages

José A. Campillo-Balderas<sup>1</sup>, Antonio Lazcano<sup>1,2</sup> and Arturo Becerra<sup>1\*</sup>

<sup>1</sup> Evolutionary Biology, Facultad de Ciencias, Universidad Nacional Autónoma de México, Mexico City, Mexico, <sup>2</sup> Miembro de El Colegio Nacional, Mexico City, Mexico

## OPEN ACCESS

### Edited by:

Johann Peter Gogarten,  
University of Connecticut, USA

### Reviewed by:

Youn-Sig Kwak,  
Gyeongsang National University,  
South Korea  
Soo Rin Kim,  
Kyungpook National University,  
South Korea

### \*Correspondence:

Arturo Becerra  
abb@ciencias.unam.mx

### Specialty section:

This article was submitted to  
Evolutionary and Genomic  
Microbiology,  
a section of the journal  
Frontiers in Ecology and Evolution

Received: 04 July 2015

Accepted: 07 December 2015

Published: 23 December 2015

### Citation:

Campillo-Balderas JA, Lazcano A and  
Becerra A (2015) Viral Genome Size  
Distribution Does not Correlate with  
the Antiquity of the Host Lineages.  
*Front. Ecol. Evol.* 3:143.  
doi: 10.3389/fevo.2015.00143

It has been suggested that RNA viruses and other subcellular entities endowed with RNA genomes are relicts from an ancient RNA/protein World which is believed to have preceded extant DNA/RNA/protein-based cells. According to their proponents, this possibility is supported by the small-genome sizes of RNA viruses and their manifold replication strategies, which have been interpreted as the result of an evolutionary exploration of different alternative genome organizations and replication strategies during early evolutionary stages. At the other extreme are the giant DNA viruses, whose genome sizes can be as large as those of some prokaryotes, and which have been grouped by some authors into a fourth domain of life. As argued here, the comparative analysis of the chemical nature and sizes of the viral genomes reported in GenBank does not reveal any obvious correlation with the phylogenetic history of their hosts. Accordingly, it is somewhat difficult to reconcile the proposal of the putative pre-DNA antiquity of RNA viruses, with their extraordinary diversity in plant hosts and their apparent absence among the Archaea. Other issues related to the genome size of all known viruses and subviral agents and the relationship with their hosts are discussed.

**Keywords:** viral genome sizes, origin of viruses, viruses and the RNA-protein world

## INTRODUCTION

Almost as soon as they were discovered and characterized as subcellular entities, viruses were considered by some to be the first forms of life (d'Herelle, 1926). Many were convinced that the small size and apparent simplicity of modern viruses could be interpreted as evidence of their primitiveness, and that they could be considered as operational models of the processes that had led to the emergence of life (Beutner, 1938; Podolsky, 1996; Fisher, 2010).

During the past few years, this virocentric hypothesis has been resurrected based on both the nature and size genome of viruses. Since RNA viruses have small genomes and diverse strategies of replication, it has been suggested they have their roots in the early stages of evolution that preceded the appearance of cellular DNA genomes (Forterre, 2006; Koonin et al., 2006; Agol, 2010; Koonin and Dolja, 2013). On the other extreme are the so-called nucleocytoplasm large DNA viruses (NCLDV) endowed with the largest viral genomes reported so far, which in some cases may be even larger than some small prokaryotic genomes, have been considered by some as a fourth domain of life comparable to the Bacteria, the Archaea and the Eucarya (Raoult et al., 2004; Boyer et al., 2009; Nasir et al., 2012).

There are several studies trying to date the emergence of viruses. In some of them, it has been analyzed the distribution of viral-protein domains at the SCOP database and it has been concluded viruses emerged from primordial segmented RNA proto-virocells and not from modern cellular entities, and also, they suggested that eukaryotic viruses are not descendants from prokaryotic viruses (Nasir et al., 2015). In others, it has been determined the existence of replication-and-structure hallmark genes not found in cell genomes, and therefore it has been proposed an ancient virus world (Koonin et al., 2006). Moreover, in other works, it has been proposed these genes have homologs in cell genomes, and probably they could be horizontally transferred between cells and viruses, and between viruses and other viruses (Caprari et al., 2015). In other studies, it has been suggested the origin of viruses coincides with the appearance of viral capsid. The cell capsid-like genes could be considered an exaptation that emerged from horizontal gene transfer from cells to cellular parasitic templates (Jalasvuori et al., 2015).

However, the ultimate origin of viruses is still unknown and remains an open issue. In the present study, we have studied the correlation of genome size of both RNA- and DNA viruses with the antiquity of the lineages of their prokaryotic and eukaryotic hosts in order to date the possible emergence of viruses after cell origin, and gain some insights on the evolutionary aspects of their phylogenetic relationship. We have used both the genomic information of the reference species of all viral families reported in GenBank as of December 2014, and the molecular, cellular, phylogenetic, and information of their hosts distributed in the three major domains of life.

## MATERIALS AND METHODS

### Retrieval of Viral and Host Data

Biological data of RNA- and DNA viruses (species, host, group, family, taxonomic code, genome size, and number of segments) as of December 2014 were retrieved from GenBank (<http://www.ncbi.nlm.nih.gov/genomes/>) on a plain-text file (see Supplementary Material 1) and, in some cases, verified or complemented with the 9th Report of the International Committee on Taxonomy of Viruses (King et al., 2011), the viral web resource ViralZone (<http://viralzone.expasy.org/>), and from relevant publications. A total of 4182 viral reference strains were obtained from the Genbank. A total of 215 satellite- and 44 viroid reference strains were also retrieved, but treated as additional independent points for this study. Viruses whose host was not identified in the GenBank ( $n = 31$ ) were excluded.

### Classification of Virus Data

All virus reference strains of the database were classified in four categories (see Supplementary Material 1). According to the Baltimore Classification System (Baltimore, 1971), the first category included seven groups: double-stranded DNA (dsDNA,  $n = 1926$ ), single-stranded DNA (ssDNA,  $n = 701$ ), double-stranded RNA (dsRNA,  $n = 205$ ), positive-sense ssRNA [ssRNA(+),  $n = 966$ ], negative-sense ssRNA [ssRNA(-),  $n = 253$ ], reverse-transcribing dsDNA (dsDNA-RT,  $n = 70$ ), and reverse-transcribing ssRNA (ssRNA-RT,  $n = 62$ ). Depending

on the host type, the second category included viruses infecting prokaryotes: Bacteria ( $n = 1438$ ) and Archaea ( $n = 69$ ); and viruses infecting eukaryotes: diatoms ( $n = 5$ ), algae ( $n = 37$ ), protists ( $n = 32$ ), plants ( $n = 1273$ ), fungi ( $n = 82$ ), plants and invertebrates ( $n = 58$ ), invertebrates ( $n = 260$ ), invertebrates and vertebrates ( $n = 123$ ), vertebrates ( $n = 1064$ ). According to their level of segmentation, the third category included viruses from 1 to more than 105 segments divided by Baltimore groups and host types. According to the genome type, the fourth category included 55 families of RNA viruses ( $n = 1485$ ) and 43 families of DNA viruses ( $n = 2697$ ).

### Analysis of Viral Genome Size According to the Baltimore Classification, Host Type, and Their Level of Segmentation

The genome size average of each set of viruses grouped by Baltimore Classification and host type was calculated. All viruses were also classified by genome size, host type, and number of segments. The large ranges of genome sizes were graphed logarithmically with a base-10 log scale in both cases.

### Analysis of Viral Genome Size According to the Antiquity of Cell Domains

Data of the 99 families of viruses that have an identified host in the GenBank was compiled according to viral genome nature and host type. The percentage of RNA- and DNA viral families of each host type was estimated. Viral families were double-counted if they infected more than one host type. Viral families were classified according to host types as follows: proteobacteria ( $n = 8$  viral families), other bacteria ( $n = 7$ ) for the Bacteria domain; Crenarchaeota ( $n = 9$ ) and Euryarchaeota ( $n = 4$ ) for the Archaea domain; and protists and algae ( $n = 7$ ), plants ( $n = 21$ ), fungi ( $n = 15$ ), and animals ( $n = 50$ ) for the Eucarya domain.

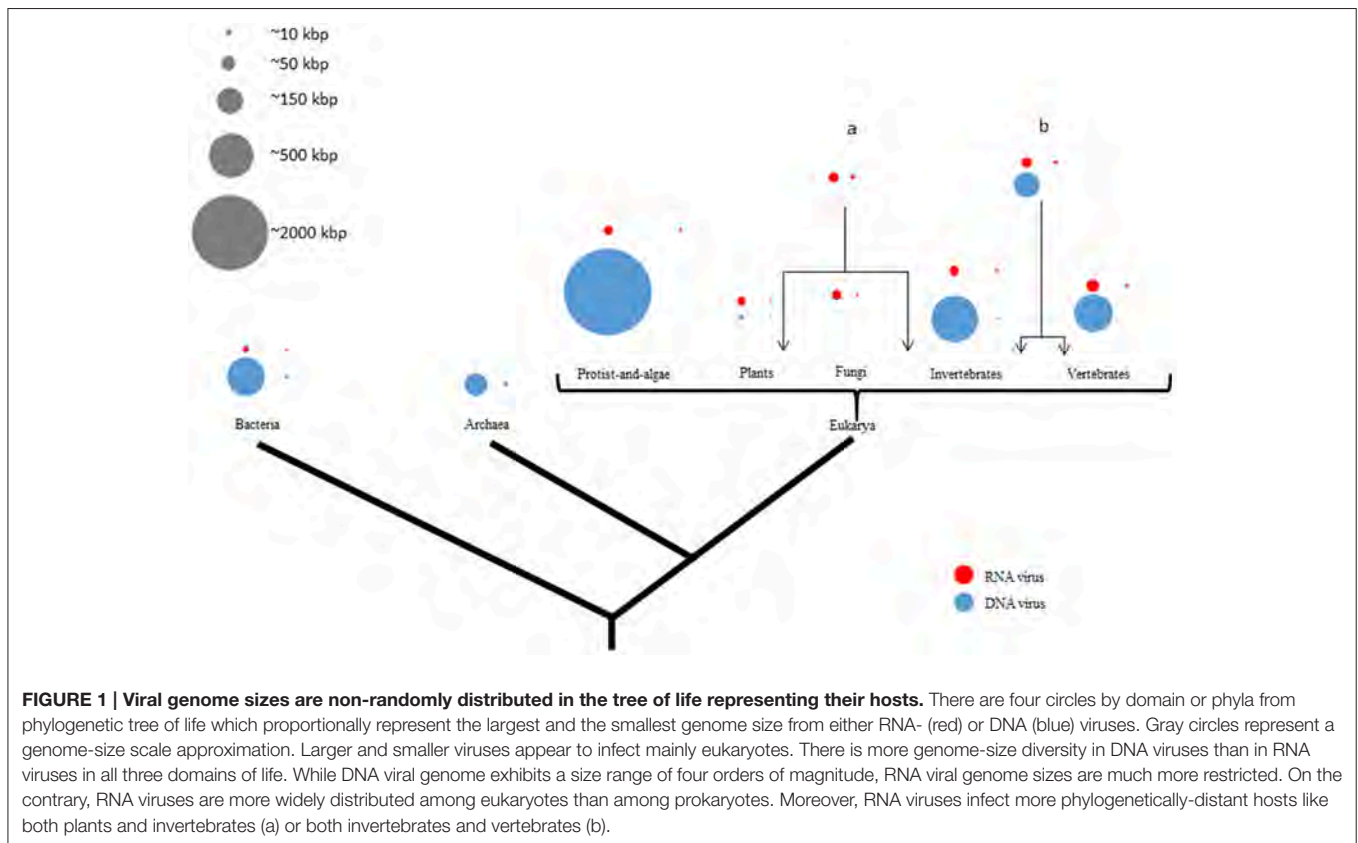
The Interactive Tree of Life (IToL) platform was used to generate a phylogenetic tree of cell domains following the online instructions based on Letunic and Bork (2007).

## RESULTS

There is a bias in our study. We have determined the number of all viral records according to the Baltimore Classification and host type. Viruses that infect bacteria (34%), plants (31%), and vertebrates (25%) are the most represented viruses in the current database due to medical, agricultural, and historical reasons (see Supplementary Material 1).

### Larger and Smaller Viruses Mainly Infect Eukaryotes

DNA viruses exhibit the most diverse-size genomes of all viruses in this study. In our sample, DNA-virus genome sizes vary by approximately four orders of magnitude, with the smallest (0.859 kbp) recorded in *Circovirus SFBeef* (ssDNA), and the largest one (2473 kbp) in *Pandoravirus salinus* (dsDNA). RNA viruses have the most-restricted size genomes of all viruses (Figure 1). Interestingly, DNA viruses which infect bacteria have genome sizes slightly larger than those which infect some



animals. RNA-virus genome sizes vary by approximately one order of magnitude, from the smallest (1.8 kbp) reported in *Saccharomyces cerevisiae killer virus M1* (dsRNA) to the largest one (33.452 kbp) in Ball python nidovirus [ssRNA(+)]. In spite of several major searches, as of June 2015, no RNA viruses infecting Archaea have been yet reported.

We have analyzed the genome size distribution of the viruses in our sample following the Baltimore Classification of the seven groups of viruses [dsDNA, ssDNA, dsRNA, ssRNA(+), ssRNA(-), ssRNA-RT, and dsDNA-RT] and host type (Bacteria, Archaea; diatoms, algae, plants, fungi, invertebrates, vertebrates, both plants and invertebrates, and both invertebrates and vertebrates).

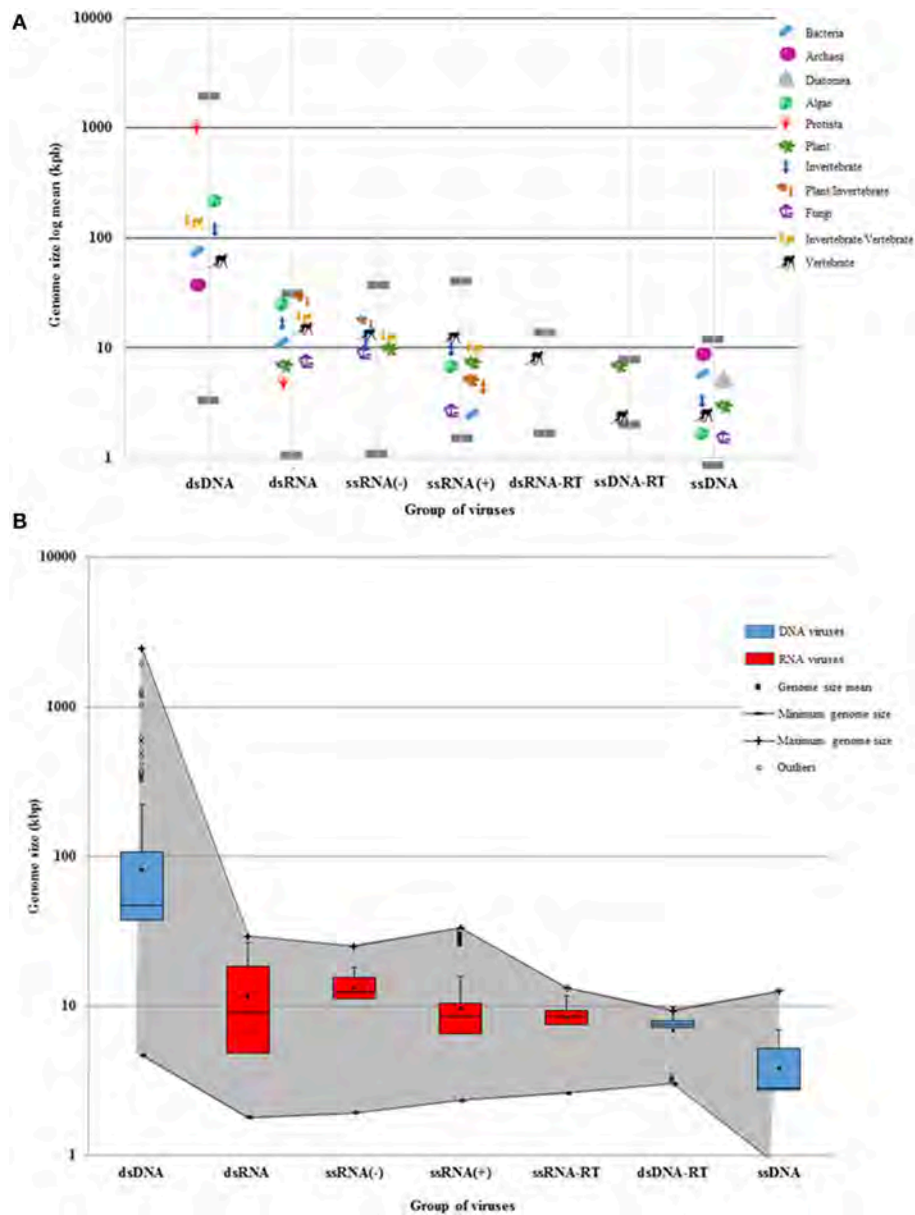
As shown in **Figures 1, 2**, the DNA- and RNA viral entities analyzed in our sample exhibit considerable diversity in their genome sizes but within well-defined limits. Moreover, viral genome sizes are not randomly distributed. Our results show that DNA-virus genomes exhibit a diverse range of sizes, with the dsDNA viruses displaying a more size-flexible genome distribution than ssDNA viruses (**Figure 2**). The genome sizes of dsDNA viruses vary by approximately two orders of magnitude, and can be divided into three well-defined genome-size groups: (a) those which infect protists (1000 kbp); (b) those which infect algae, invertebrates, and vertebrates (from 160 to 240 kbp); and (c) those which infect prokaryotes and vertebrates (from 40 to 70 kbp). There are no reports of dsDNA viruses infecting plants or fungi. Of the 33 families and 133 unclassified species of dsDNA viruses, the amoeba-infecting *P. salinus* (unclassified)

has the largest genome (2500 kbp), and the vertebrate-infecting *Bovine polyomavirus* (Polyomaviridae), the smallest one (5 kbp) (**Figure 2A** and Supplementary Material 2). The majority of the largest-genome sizes of dsDNA, ssDNA, and ssRNA(-) viruses occupies the third quartile (75%) of the data (**Figure 2B**).

Secondly, our results demonstrate that RNA viruses exhibit the most restricted-size genome distribution of all viruses (**Figure 2**). The dsRNA viral group has, on the average, the largest genomes of all RNA viruses. Of the 12 families and 19 unclassified species of dsRNA viruses, the multihost-infecting *Fiji disease virus* (Reoviridae) has the largest genome (29 kbp), while the fungi-infecting *Saccharomyces cerevisiae killer virus M1* has the smallest one (2 kbp) reported so far (**Figure 2B**). The majority of dsRNA has genome sizes that range from 4 to 9 kbp (50 and 75%, respectively) (**Figure 2B**).

The available data show that ssRNA(-) viruses have a limited range of genome sizes (from 10 to 15 kbp) independently of their eukaryotic hosts. Of the nine families and eight unclassified species of ssRNA(-) viruses, the plant-and-invertebrate-infecting *Rice grassy stunt virus* (unclassified) has the largest genome (25 kbp), while the plant-infecting *Blueberry mosaic associated virus* (Ophioviridae) has the smallest one (2 kbp) (**Figure 2A**). The majority of ssRNA(-) has genome sizes that range from 1 to 3 kbp (50 and 75%, respectively) (**Figure 2B**). On the other hand, dsRNA viruses and some ssRNA(-) that infect either plants or vertebrates (some of them via a vector) have rather large genomes (**Figure 2** and Supplementary Material 2).





**FIGURE 2 | Viral genome sizes and the Baltimore groups. (A)** The genome size average of viruses was calculated according to each Baltimore group and host type. Each host is denoted by a representative organism. Gray lines show the largest and the smallest virus of each Baltimore group. DNA viruses are grouped into two size extremes of the log graph (dsDNA and ssDNA). The dsDNA virus genomes show a more extended diversity than ssDNA genomes. Bacteria hosts can also be infected by longer dsDNA viruses, but rarely by RNA viruses. RNA and ssDNA viruses display more restricted genomes. However, RNA virus genomes can be larger than ssDNA genomes. The largest and the smallest RNA viruses are found as parasites infecting eukaryotes. Some families of RNA viruses can infect phylogenetically-distant hosts (plants and invertebrates, or invertebrates and vertebrates). Retroviruses only infect plants and vertebrates. **(B)** The maximum and minimum genome size of RNA (red) and DNA (blue) viruses are denoted by a gray shadow. While DNA viruses have a more delineated plasticity of genome size, RNA viruses have more restricted genomes. There are more ds- and ss-DNA viral species that tend to have larger genomes as shown in the third quartile. While ssDNA and ds-RT DNA viruses have as restricted genomes as RNA viruses, dsDNA viruses tend to have larger genomes as shown in the outliers. The mean, the median, and the standard deviation were calculated from the 4182 viral reference strains of GenBank.

On the average, the smallest RNA viral genomes are found in the ssRNA(+) viruses. They can be divided in two genome-size groups: (a) those which infect bacteria and fungi (4 kbp) and (b) those which infect algae, plants, invertebrates, and vertebrates (from 6 to 12 kbp) (Figure 2). Of the 33 families and 61 unclassified species of ssRNA(+)

studied here, the vertebrate-infecting *Ball python nidovirus* (Nidoviridae) has the largest genome (33 kbp), and the fungi-infecting *Ophiostoma mitovirus 6* (Narnaviridae), the smallest one (2 kbp). The majority of ssRNA(+) has genome sizes that range from 1 to 2 kbp (75 and 50%, respectively) (Figure 2B).

Thirdly, the retro-transcribing viruses (ssRNA-RT and dsDNA-RT) have the most limited habitats of all viruses. The ssRNA-RT viruses only infect vertebrates and have genome sizes of 2–13 kbp. Similarly, the dsDNA-RT viruses described so far only infect plants or vertebrates and have a genome of ~9 kbp (Figure 2). While the majority of ssRNA-RT has genome sizes that range from 1 to 0.8 kbp (75 and 50%, respectively) (Figure 2B), most of dsDNA-RT have genome sizes that range from 0.3 to 7 kbp (75 and 50%, respectively) (Figure 2B).

Quite surprisingly, the smallest viral genomes are found in the ssDNA viruses (Circoviridae, ~1 to 2 kbp). The genomes of ssDNA viruses are clearly more size-restricted than those of dsDNA viruses, and in our sample the upper-size limit of ssDNA genome sizes is 10 kbp on average (Figure 2A). The available data show that ssDNA viruses are the only ones that infect diatomea, and the smallest known genomes of ssDNA viruses (and, indeed, of all viruses) are found in algae on average. There is only one ssDNA viral species with a genome size of 2 kbp that infects fungi. In the sample reported here, of all nine families of ssDNA viruses and the 54 unclassified species of ssDNA viruses, the invertebrate-infecting *Bombyx mori bidensovirus* (Bidnaviridae) has the largest genome (12 kbp), while the invertebrate-infecting *Circoviridae SFBeef* (Circoviridae) has the smallest one (<1 kbp). The majority of ssDNA viruses has genome sizes that range from 0.8 to 2 kbp (50 and 75%, respectively) (Figure 2B).

Quite remarkably, the analysis of the distribution of viral sample studied here indicates that viruses endowed with the largest and the smallest genomes only infect eukaryotes (see Supplemented Material 2).

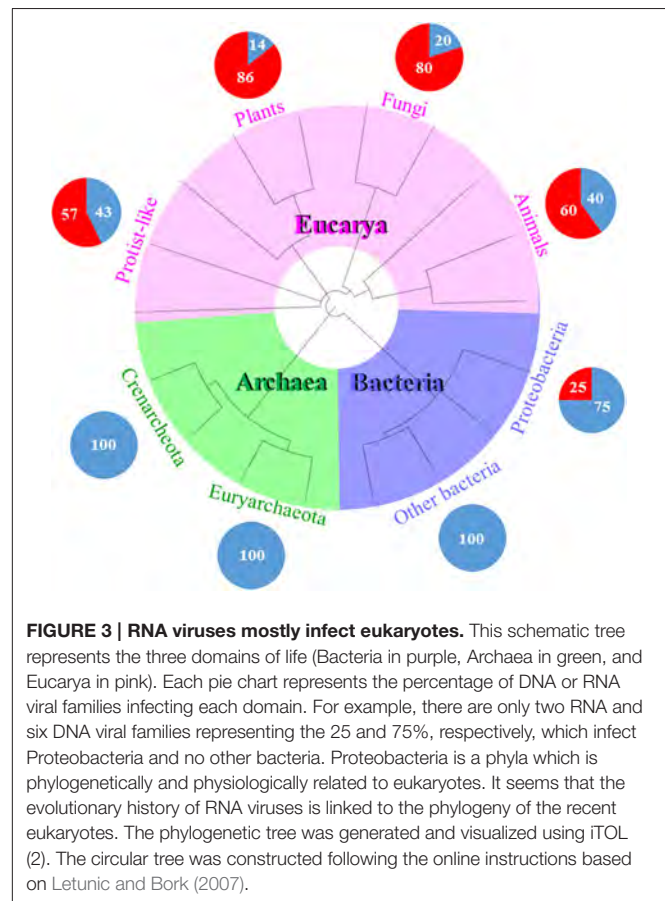
## RNA Viruses Mostly Infect Eukaryotes

The sample reported here has 44% of DNA viruses (44 families) and 56% of RNA viruses (55 families) (Supplementary Material 3). There are 18 DNA viral families that infect prokaryotes, and 26 DNA viral families that infect eukaryotes. On the other side, there are only two RNA viral families that infect prokaryotes, and 53 RNA viral families that infect eukaryotes. Hence, RNA viral families are found mostly infecting the eukaryotic domain (Figure 3).

Ten DNA- and RNA viral families have been described that infect Bacteria. There are eight DNA- and only two RNA families of bacterial viruses. Seven of the eight DNA viral families infect the Actinobacteria, Deinococcus-Thermus, Firmicutes, and Tenericutes clades. Of all the eight viral families that infect Proteobacteria, only two of them are RNA viruses.

Of the 12 DNA viral families that are known to infect Archaea, four of them infect Euryarchaeota, and nine infect Crenarchaeota, while the Fuselloviridae infect both archeal clades. The Myoviridae and the Siphoviridae have viral species that cross-infect bacteria and archaea. As of today, there is not a single record of an RNA virus infecting an archaea (Figure 3).

There are 79 DNA- and RNA viral families that infect Eucarya. Six of them infect protist-and-algae hosts, and 74 families infect fungi, plants, and animals. The Reoviridae is the only family that infects both protist-and-algae and multicellular eukaryotes. Of the 55 RNA viral families, the 96% of them belong to eukaryotic viruses, i.e., most RNA viral families are found in eukaryotes and not in prokaryotes (Figure 3).

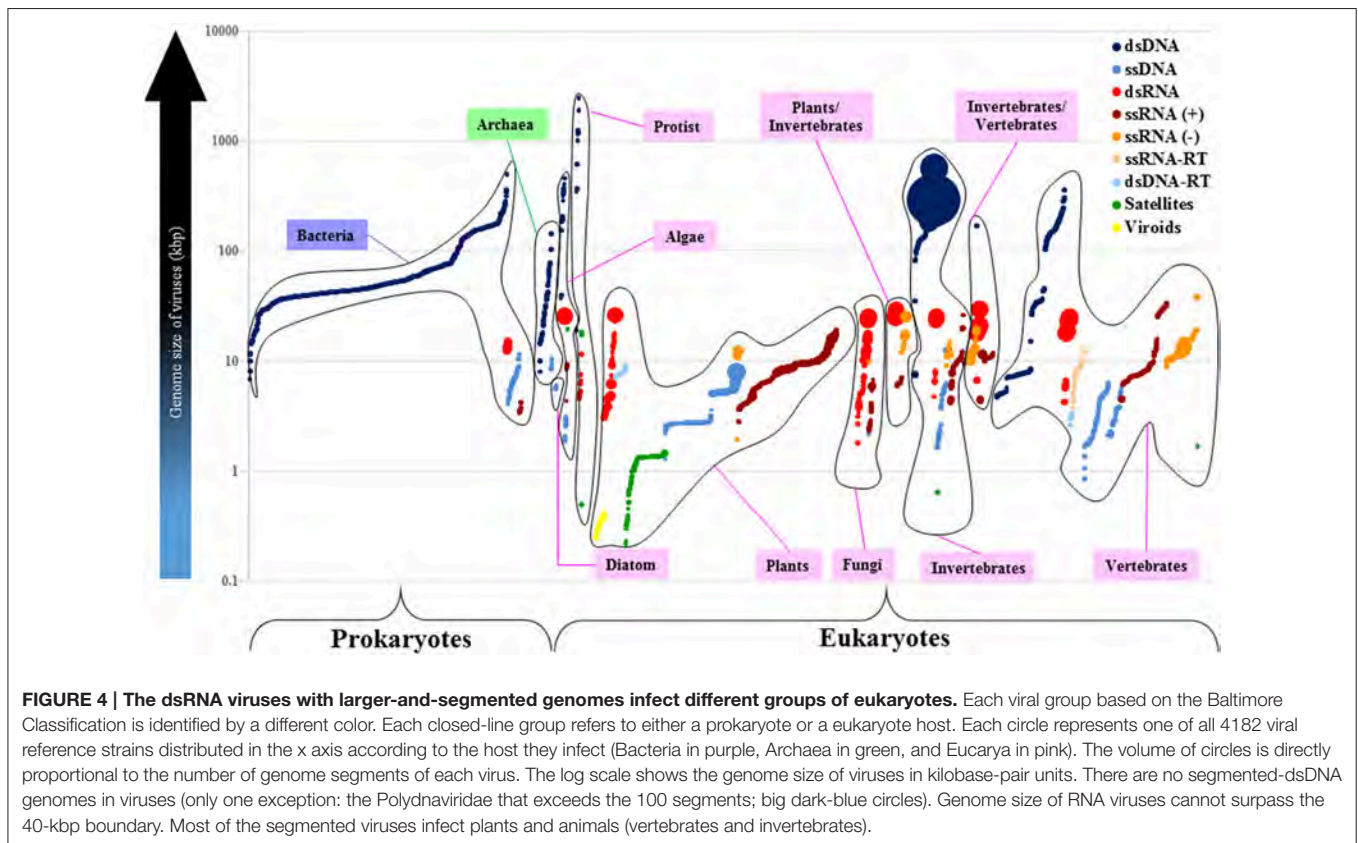


## The dsRNA Viruses with Larger and Segmented Genomes Infect a Wide Range of Phylogenetically-Distant Eukaryotes

In our sample, 89% of all known viruses are non-segmented. Of all the DNA- and RNA viruses described so far, 9% of them have two or three segments, and only 2% of them have more than four segments. There are twice as much segmented RNA viruses than segmented DNA viruses. There is a clear bias in their distribution. Of all segmented DNA- and RNA viruses, 60% infect plants, 21% infect several phylogenetically-distant eukaryotes (either plants and invertebrates, or invertebrates and vertebrates), 18% infect vertebrates, and only 1% infects bacteria.

The Polydnviridae is the only dsDNA viral group with segmented genomes, and it exhibits the largest genome size of all known invertebrate viruses. Their genomes range from 15 to 105 segments (Figure 4). The second most-segmented genomes of all known viruses are found in the plant-infecting Nanoviridae (ssDNA), and range from 6 to 14 segments. The plant-infecting Geminiviridae (two segments), and the vertebrate-infecting Birdnaviridae (two segments) are the only other ssDNA viral families endowed with segmented genomes.

On the other hand, there are 20 RNA viral families with segmented genomes. The third most-segmented genomes (2–12 segments) of all known viruses are found in the Reoviridae (dsRNA) which also exhibit, on the average, the largest genomes of all RNA viruses, but do not seem to surpass the 40-kbp size



limit. The Reoviridae is the only viral family that infects several eukaryotic hosts including algae, plants, fungi, invertebrates, and vertebrates (Figure 4 and Supplementary Material 3). There are seven dsRNA-viral families whose genomes range from 2 to 12 segments, and have sizes that range from 3 to 29 kbp. There are five ssRNA(−) viral families whose genomes range from 1 to 8 segments, and vary from 1 to 38 kbp (the Filoviridae have the largest genome of all known RNA viruses, with two segments that add up to 38 kbp). There are eight ssRNA(+) viral families whose genomes range from 1 to 4 segments, and whose sizes vary from 3 to 19 kbp.

Therefore, viruses endowed with the smallest genomes, like ssDNA and RNA viruses, appear to have the most segmented genomes (with an extreme exception in dsDNA viruses from the Polydnaviridae). Moreover, the largest RNA genomes are segmented and are found in dsRNA viruses that infect a wide range of phylogenetically-distant eukaryotes.

## DISCUSSION

In this report we have investigated the relationship between viral genome sizes and the antiquity of their host lineages, and have shown that the available data demonstrate that genome sizes do not exhibit a random distribution. In our sample, genome sizes of dsDNA viruses have the highest diversity and also surpass by far the more restricted genome sizes of RNA and ssDNA viruses. On the contrary, RNA viruses exhibit a wider range of eukaryotic hosts, but infect relatively few bacterial lineages compared to DNA viruses.

Our results show that dsDNA viral genomes display a much more diverse size range than RNA- and ssDNA viruses. This can be explained as a result of the enhanced chemical stability of the Watson-Crick helices. The available data indicate that dsDNA viral genome sizes can be divided in three major groups. The largest dsDNA genomes are those of the so-called nucleocytoplasmic large DNA viruses (NCLDV) that infect amoeba, such as *Pandoravirus* and *Pithovirus* (Raoult et al., 2004; Pennisi, 2013; Philippe et al., 2013; Legendre et al., 2014), whose genomes appear to have major contributions from other viruses as well as from prokaryotic and eukaryotic microbes, due to accretion processes in which horizontal gene transfer may have played a significant role (Filée et al., 2007; Colson and Raoult, 2010; Filée, 2013).

The second group is that of dsDNA viruses that infect algae, invertebrates and vertebrates, and that have genome sizes that range from 150 to 240 kbp. This second genome-size group includes giant viruses like the Phycodnaviridae, the Iridoviridae, and the Asfarviridae. It has been suggested that these eukaryotic viral families share a common ancestor with the largest-genome giant viruses, supporting the idea of an additional branch of life (Iyer et al., 2006; Boyer et al., 2009; Yutin and Koonin, 2012; Nasir et al., 2015). However, it has been argued that the giant viruses, like the Marseilleviridae, have in fact increased their genomes with eukaryotic sequences through horizontal gene transfer (Moreira and Brochier-Armanet, 2008; Boyer et al., 2009; Filée, 2014) and do not constitute a fourth domain of life (Yutin et al., 2014).

Finally, the third and also the smallest genome-size group includes dsDNA viruses that infect prokaryotes and vertebrates.



It is possible that the small size of prokaryotic viruses is constrained by their small-size capsids (Krupovic et al., 2011). However, the Myoviridae and the Siphoviridae, the only two families that cross-infect both Bacteria and Archaea, have genomes that range from 10 to 500 kbp. Together with the Podoviridae, these two-tailed viral families of bacteriophages appear to be an ancient and genetically connected viral group (Hendrix, 2002). There is no evidence of cross-infection between prokaryotes and eukaryotes, which may suggest a domain-specific origin of viruses. The smallest genomes of dsDNA viruses (5–7 kbp) are those of the Polyomaviridae and the Papillomaviridae, which infect mammals and birds (de Villiers et al., 2004; Crandall et al., 2006). Therefore, the largest genomes of dsDNA viruses are found in those which infect eukaryotes.

It is somewhat surprising that the smallest genomes are found not only in RNA viruses but also in ssDNA viruses. Although both viral types exhibit a difference of one magnitude in their genome sizes, the smallest viral genomes are found in ssDNA viruses. It thus appears that the genomes of ssDNA viruses are subjected to the same restrictions that hinder the size increase in RNA viral genomes, most likely due to the lack of repair mechanisms (Reanne, 1982). Both viral types exhibit comparable behavior, including high mutation rate, large population sizes, small levels of horizontal gene transfer, little gene duplication, overlapping reading frames and, often, little recombination (Duffy and Holmes, 2008; Holmes, 2009).

Unlike DNA viruses, RNA viral families infect a wide range of phylogenetically diverse eukaryotic hosts, an evolutionary dispersal that may explain why some of them have coevolved with their invertebrate vectors (Gray and Banerjee, 1999; Lobo et al., 2009; Obbard and Dudas, 2014). One of the viral families that infect multiple hosts is the Reoviridae, which also exhibit multiple segmentation and large genomes (see **Figure 3**). It has been suggested that segments of dsRNA genome of Reoviridae probably recombine through complementation when two or more viruses co-infect a single cell (Reanne, 1982; Froissart et al., 2004; Holmes, 2009). It has been argued that the Reoviridae cannot undergo major increases in the genome size, since this would require a complex molecular machinery including unwinding proteins, DNA-dependent ATPases, and nucleases which are not encoded by RNA viruses (Reanne, 1982).

It is somewhat surprising that with the exception of only two known examples, all RNA viral families appear to be restricted to eukaryotic hosts. The only two families that infect bacteria (Proteobacteria) are the Leviviridae [ssRNA(+)] and the Cystoviridae (dsRNA). It has been speculated that the latter could be derived from eukaryotic viruses (Holmes, 2009).

The wide range of RNA viral parasites infecting nucleated cells very likely explains the eukaryotic defense mechanisms that include degradation of viral RNA, presence of microRNAs,

and RNAi mechanisms against viruses (Berkhout and Haasnoot, 2006; Lodish et al., 2008; Obbard et al., 2009; Parameswaran et al., 2010; Obbard and Dudas, 2014). RNA-mediated silencing is a highly conserved mechanism that was probably present in the last common ancestor of eukaryotes (Cerutti and Casas-Mollano, 2006), which may indicate an ancient evolutionary relationship between nucleated cells and RNA viruses, whose origin could thus be placed some time near the actual emergence of eukaryotic microbes.

As reviewed above, it has been argued that viruses were the first living entities and RNA viruses or viroids may be direct descendants of the RNA World. It has also been suggested that retroviral-like elements are relics of the early evolutionary transition from an RNA/protein world into the extant DNA/RNA/protein cells, and that the ancestor of dsDNA giant viruses was an ancient cell (Podolsky, 1996; Daròs et al., 2006; Koonin et al., 2006; Flores et al., 2014). Our results suggest that these schemes may be mistaken. This alternative possibility is supported by phylogenetic analyses that indicate that all known viral monomeric RNA polymerases are derived from cellular DNA polymerases A and B (Jácome et al., 2015). Although the results presented here may be severely affected by methodological issues that include biased representations of viral diversity, our data show that in terms of their genome size and organization RNA viruses are not endowed with the simpler and smallest genomes of all known viruses as is generally believed, and in fact that they may be more closely related to the evolutionary history of their eukaryotic hosts. Our results also suggest that since retroviruses appear to be restricted to plants and vertebrates, they could not have played a role in the evolutionary transition from primitive cellular RNA genomes to the extant DNA-based genetic systems of extant cells, nor the viral reverse transcriptase can be considered an evolutionary vestige of the polymerase that played a role in this transition. The results presented here demonstrate that viral genome sizes are not randomly distributed, but do not appear to be correlated with the antiquity of their hosts. Therefore, viruses may be ancient, but not primitive.

## ACKNOWLEDGMENTS

We are indebted to Dr. León Patricio Martínez Castilla for several useful discussions. JC is supported by the Consejo Nacional de Ciencia y Tecnología (CONACyT), scholarship number 165264. The support of the Posgrado en Ciencias Biológicas, UNAM, to JC is gratefully acknowledged.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <http://journal.frontiersin.org/article/10.3389/fevo.2015.00143>

## REFERENCES

- Agol, V. I. (2010). Which came first, the virus or the cell? *Paleontol. J.* 44, 728–736. doi: 10.1134/S0031030110070038
- Baltimore, D. (1971). Expression of animal virus genomes. *Bacteriol. Rev.* 35, 235–241.

- Berkhout, B., and Haasnoot, J. (2006). The interplay between virus infection and the cellular RNA interference machinery. *FEBS Lett.* 580, 2896–2902. doi: 10.1016/j.febslet.2006.02.070
- Beutner, R. (1938). *Life's Beginning on the Earth*. Baltimore: The Williams & Wilkins Company.

- Boyer, M., Yutin, N., Pagnier, I., Barrassi, L., Fournous, G., Espinosa, L., et al. (2009). Giant Marseillevirus highlights the role of amoebae as a melting pot in emergence of chimeric microorganisms. *Proc. Natl. Acad. Sci. U.S.A.* 106, 21848–21853. doi: 10.1073/pnas.0911354106
- Caprari, S., Metzler, S., Lengauer, T., and Kalinina, O. V. (2015). Sequence and structure analysis of distantly-related viruses reveals extensive gene transfer between viruses and hosts and among viruses. *Viruses* 10, 5388–5409. doi: 10.3390/v7102882
- Cerutti, H., and Casas-Mollano, J. A. (2006). On the origin and functions of RNA-mediated silencing: from protists to man. *Curr. Genet.* 50, 81–99. doi: 10.1007/s00294-006-0078-x
- Colson, P., and Raoult, D. (2010). Gene repertoire of amoeba-associated giant viruses. *Intervirology* 53, 330–343. doi: 10.1159/000312918
- Crandall, K. A., Pérez-Losada, M., Christensen, R. G., McClellan, D. A., and Viscidi, R. P. (2006). Phylogenomics and molecular evolution of polyomaviruses. *Adv. Exp. Med. Bio.* 577, 46–59. doi: 10.1007/0-387-32957-9\_3
- Daros, J.-A., Elena, S. F., and Flores, R. (2006). Viroids: an Ariadne's thread into the RNA labyrinth. *EMBO Rep.* 7, 593–598. doi: 10.1038/sj.embor.7400706
- de Villiers, E. M., Fauquet, C., Broker, T. R., Bernard, H. U., and Zur Hausen, H. (2004). Classification of papillomaviruses. *Virology* 324, 17–27. doi: 10.1016/j.virol.2004.03.033
- d'Herelle, F. (1926). *The Bacteriophage and its Behavior*. Baltimore, MD: The Williams & Wilkins Company.
- Duffy, S., and Holmes, E. C. (2008). Phylogenetic evidence for rapid rates of molecular evolution in the single-stranded DNA begomovirus tomato yellow leaf curl virus. *J. Virol.* 82, 957–965. doi: 10.1128/JVI.01929-07
- Filee, J. (2013). Route of NCLDV evolution: the genomic accord. *Curr. Opin. Virol.* 3, 595–599. doi: 10.1016/j.coviro.2013.07.003
- Filee, J. (2014). Multiple occurrences of giant virus core genes acquired by eukaryotic genomes: the visible part of the iceberg? *Virology* 466, 53–59. doi: 10.1016/j.virol.2014.06.004
- Filée, J., Siguier, P., and Chandler, M. (2007). I am what I eat and I eat what I am: acquisition of bacterial genes by giant viruses. *Trends Genet.* 23, 10–15. doi: 10.1016/j.tig.2006.11.002
- Fisher, S. (2010). Are RNA viruses vestiges of an RNA world? *J. Gen. Philos. Sci.* 41, 121–141. doi: 10.1007/s10838-010-9119-8
- Flores, R., Gago-Zachert, S., Serra, P., Sanjuán, R., and Elena, S. F. (2014). Viroids: Survivors from the RNA World? *Annu. Rev. Microbiol.* 68, 395–414. doi: 10.1146/annurev-micro-091313-103416
- Forterre, P. (2006). The origin of viruses and their possible roles in major evolutionary transitions. *Virus Res.* 117, 5–16. doi: 10.1016/j.virusres.2006.01.010
- Froissart, R., Wilke, C. O., Montville, R., Remold, S. K., Chao, L., and Turner, P. E. (2004). Co-infection weakens selection against epistatic mutations in RNA viruses. *Genetics* 168, 9–19. doi: 10.1534/genetics.104.030205
- Gray, S. M., and Banerjee, N. (1999). Mechanisms of arthropod transmission of plant and animal viruses. *Microbiol. Mol. Biol. Rev.* 63, 128–148.
- Hendrix, R. W. (2002). Bacteriophages: evolution of the majority. *Theor. Popul. Biol.* 61, 471–480. doi: 10.1006/tpbi.2002.1590
- Holmes, E. (2009). *The Evolution and Emergence of RNA Viruses*. New York, NY: Oxford University Press Inc.
- Iyer, L. A., Balaji, S., Koonin, E. V., and Aravind, L. (2006). Evolutionary genomics of nucleocytoplasmic large DNA viruses. *Virus Res.* 117, 156–184. doi: 10.1016/j.virusres.2006.01.009
- Jácome, R., Becerra, A., Ponce De León, S., and Lazcano, A. (2015). Structural analysis of monomeric RNA-dependent polymerases: evolutionary and therapeutic implications. *PLoS ONE* 10:e0139001. doi: 10.1371/journal.pone.0139001
- Jalasuuri, M., Mattila, S., and Hoikkala, V. (2015). Chasing the origin of viruses: capsid-forming genes as a life-saving preadaptation within a community of early replicators. *PLoS ONE* 10:e0126094. doi: 10.1371/journal.pone.0126094
- King, A. M. Q., Lefkowitz, E., Adams, M. J., and Carstens, E. B. (2011). *Virus Taxonomy: Classification and Nomenclature of Viruses: Ninth Report of the International Committee on Taxonomy of Viruses*. San Diego, CA: Elsevier Academic Press.
- Koonin, E. V., and Dolja, V. V. (2013). A virocentric perspective on the evolution of life. *Curr. Opin. Virol.* 3, 546–557. doi: 10.1016/j.coviro.2013.06.008
- Koonin, E. V., Senkevich, T. G., and Dolja, V. V. (2006). The ancient virus world and evolution of cells. *Biol. Direct* 1, 1–27. doi: 10.1186/1745-6150-1-1
- Krupovic, M., Prangishvili, D., Hendrix, R. W., and Bamford, D. H. (2011). Genomics of bacterial and archaeal viruses: dynamics within the prokaryotic virosphere. *Microbiol. Mol. Biol. Rev.* 75, 610–635. doi: 10.1128/MMBR.00011-11
- Legendre, M., Bartoli, J., Shmakova, L., Jeudy, S., Labadie, K., Adrait, A., et al. (2014). Thirty-thousand-year-old distant relative of giant icosahedral DNA viruses with a pandoravirus morphology. *Proc. Natl. Acad. Sci. U.S.A.* 111, 4274–4279. doi: 10.1073/pnas.1320670111
- Letunic, I., and Bork, P. (2007). Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics* 23, 127–128. doi: 10.1093/bioinformatics/btl529
- Lobo, F. P., Mota, B. E. F., Pena, S. D. J., Azevedo, V., Macedo, A. M., Tauch, A., et al. (2009). Virus-host coevolution: common patterns of nucleotide motif usage in Flaviviridae and their hosts. *PLoS ONE* 4:e6282. doi: 10.1371/journal.pone.0006282
- Lodish, H. F., Zhou, B., Liu, G., and Chen, C. Z. (2008). Micromanagement of the immune system by microRNAs. *Nat. Rev. Immunol.* 8, 120–130. doi: 10.1038/nri2252
- Moreira, D., and Brochier-Armanet, C. (2008). Giant viruses, giant chimeras: the multiple evolutionary histories of Mimivirus genes. *BMC Evol. Biol.* 8:12. doi: 10.1186/1471-2148-8-12
- Nasir, A., Kim, K. M., and Caetano-Anolles, G. (2012). Giant viruses coexisted with the cellular ancestors and represent a distinct supergroup along with superkingdoms Archaea, Bacteria and Eukarya. *BMC Evol. Biol.* 12:156. doi: 10.1186/1471-2148-12-156
- Nasir, A., Sun, F.-J., Kim, K. M., and Caetano-Anollés, G. (2015). Untangling the origin of viruses and their impact on cellular evolution. *Ann. N.Y. Acad. Sci.* 1341, 61–74. doi: 10.1111/nyas.12735
- Obbard, D. J., and Dudas, G. (2014). The genetics of host-virus coevolution in invertebrates. *Curr. Opin. Virol.* 8, 73–78. doi: 10.1016/j.coviro.2014.07.002
- Obbard, D. J., Gordon, K. H. J., Buck, A. H., and Jiggins, F. M. (2009). The evolution of RNAi as a defence against viruses and transposable elements. *Philos. Trans. R. Soc. B Biol. Sci.* 364, 99–115. doi: 10.1098/rstb.2008.0168
- Parameswaran, P., Sklan, E., Wilkins, C., Burgon, T., Samuel, M. A., Lu, R., et al. (2010). Six RNA viruses and forty-one hosts: viral small RNAs and modulation of small RNA repertoires in vertebrate and invertebrate systems. *PLoS Pathog.* 6:e1000764. doi: 10.1371/journal.ppat.1000764
- Pennisi, E. (2013). Ever-bigger viruses shake tree of life. *Science* 341, 226–227. doi: 10.1126/science.341.6143.226
- Philippe, N., Legendre, M., Doutre, G., Couté, Y., Poirot, O., Lescot, M., et al. (2013). Pandoraviruses: amoeba viruses with genomes up to 2.5 Mb reaching that of parasitic eukaryotes. *Science* 341, 281–286. doi: 10.1126/science.1239181
- Podolsky, S. (1996). The role of the virus in origin-of-life theorizing. *J. Hist. Biol.* 29, 79–126.
- Raoult, D., Audic, S., Robert, C., Abergel, C., Renesto, P., Ogata, H., et al. (2004). The 1.2-megabase genome sequence of mimivirus. *Science* 306, 1344–1350. doi: 10.1126/science.1101485
- Reaney, D. C. (1982). The evolution of RNA viruses. *Annu. Rev. Microbiol.* 36, 47–73. doi: 10.1146/annurev.mi.36.100182.000403
- Yutin, N., and Koonin, E. V. (2012). Hidden evolutionary complexity of nucleocytoplasmic large DNA viruses of eukaryotes. *Virol. J.* 9:161. doi: 10.1186/1743-422X-9-161
- Yutin, N., Wolf, Y. I., and Koonin, E. V. (2014). Origin of giant viruses from smaller DNA viruses not from a fourth domain of cellular life. *Virology* 466, 38–52. doi: 10.1016/j.virol.2014.06.032

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2015 Campillo-Balderas, Lazcano and Becerra. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

## **IX. ANEXO II**

(Artículo enviado)

**A critical assessment of tertiary structure-based phylogenies: the case of hepatitis C virus and human immunodeficiency virus 1 polymerases**

# PLOS Computational Biology

## A critical assessment of tertiary structure-based phylogenies: the case of hepatitis C virus and human immunodeficiency virus 1 polymerases --Manuscript Draft--

<b>Manuscript Number:</b>	PCOMPBIOL-D-18-00518
<b>Full Title:</b>	A critical assessment of tertiary structure-based phylogenies: the case of hepatitis C virus and human immunodeficiency virus 1 polymerases
<b>Short Title:</b>	A critical assessment of tertiary structure-based phylogenies
<b>Article Type:</b>	Research Article
<b>Keywords:</b>	Tertiary structure-based phylogenies; RNA-dependent polymerases; Structural evolution; Hepatitis C virus; Human immunodeficiency virus; Protein crystallography; Deep phylogenies
<b>Corresponding Author:</b>	Rodrigo Jácome, M.D. Universidad Nacional Autonoma de Mexico Mexico, D.F. MEXICO
<b>Corresponding Author Secondary Information:</b>	
<b>Corresponding Author's Institution:</b>	Universidad Nacional Autonoma de Mexico
<b>Corresponding Author's Secondary Institution:</b>	
<b>First Author:</b>	Rodrigo Jácome, M.D.
<b>First Author Secondary Information:</b>	
<b>Order of Authors:</b>	Rodrigo Jácome, M.D. Arturo Becerra Jose Alberto Campillo-Balderas Yolanda López-Vidal Antonio Lazcano Samuel Ponce de León
<b>Order of Authors Secondary Information:</b>	
<b>Abstract:</b>	<p>The number of tertiary structures deposited in the Protein Data Bank has been growing continually during the last decades. The use of tertiary structure-based phylogenies has not followed this trend, despite its potential to uncover remote evolutionary events that primary sequence approaches fail to detect. Owing to their importance as human epidemic pathogens and their significance as antiviral targets, two of the best characterized structures are the homologous hepatitis C virus and the human immunodeficiency virus 1 RNA-dependent polymerases, with more than 100 tertiary structures from each.</p> <p>An evolutionary analysis of these two groups of viral polymerases was performed based on pairwise structural comparisons of selected ligand-free and ligand-bound structures from each of the proteins, followed by the calculations of structural distances and the construction of the corresponding dendrograms.</p> <p>We analyzed the effects that different crystallographic factors had on the topology of the resulting trees. While the ligand-free hepatitis C virus polymerases clustered into the corresponding genotypes and subtypes; the ligand-free reverse transcriptase dendrogram failed to correctly cluster some of the structures. The hepatitis C virus polymerase and the human immunodeficiency virus-1 reverse transcriptase dendrograms built with bound ligands clustered the structures according to the conformational changes triggered by the molecules, failing to reflect the accepted taxonomy of the viruses the structures belong to.</p> <p>As with other databases, there is an important biomedical bias regarding the diversity</p>

	of the hepatitis C and human immunodeficiency virus polymerases' structures, and few genotypes and subtypes are currently represented. As of today, very few works have covered methodological aspects of structure-based phylogenies. To the best of our knowledge, this is the first analysis to address the way in which crystallographic factors like resolution and the presence of ligands alter the expected topology of tertiary structure-based dendrograms, which could lead to incorrect evolutionary inferences.
<b>Suggested Reviewers:</b>	<p>Daniel Ruzek Head of laboratory, Biology Centre CAS ruzekd@paru.cas.cz</p> <p>Marc Delarue Institut Pasteur marc.delarue@pasteur.fr</p> <p>Minna Poranen University of Helsinki Minna.Poranen@helsinki.fi</p> <p>Zaida Luthey-Schulten University of Illinois zan@illinois.edu</p>
<b>Opposed Reviewers:</b>	
<b>Additional Information:</b>	
<b>Question</b>	<b>Response</b>
<p><b>Financial Disclosure</b></p> <p>Please describe all sources of funding that have supported your work. <b>This information is required for submission and will be published with your article, should it be accepted.</b> A complete funding statement should do the following:</p> <p>Include <b>grant numbers and the URLs</b> of any funder's website. Use the full name, not acronyms, of funding institutions, and use initials to identify authors who received the funding.</p> <p><b>Describe the role</b> of any sponsors or funders in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. If the funders had <b>no role</b> in any of the above, include this sentence at the end of your statement: "<i>The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.</i>"</p> <p>However, if the study was <b>unfunded</b>, please provide a statement that clearly indicates this, for example: "<i>The author(s) received no specific funding for this work.</i>"</p> <p>* typeset</p>	<p>R. J gratefully acknowledges the financial support from the Dirección General de Asuntos del Personal Académico (DGAPA), Universidad Nacional Autónoma de México (UNAM). Y.L.V. would like to thank the Programa de Apoyos a Proyectos de Investigación e Innovación Tecnológica (Grant number: IV200315), Dirección General de Asuntos del Personal Académico (DGAPA), Universidad Nacional Autónoma de México (UNAM). A.B., J.A.C.B., and A.L. are thankful to Programa de Apoyos a Proyectos de Investigación e Innovación Tecnológica (Grant Number: IN223916), Dirección General de Asuntos del Personal Académico (DGAPA), Universidad Nacional Autónoma de México (UNAM). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.</p>
<b>Competing Interests</b>	The authors have declared that no competing interests exist.



You are responsible for recognizing and disclosing on behalf of all authors any competing interest that could be perceived to bias their work, acknowledging all financial support and any other relevant financial or non-financial competing interests.

Do any authors of this manuscript have competing interests (as described in the [PLOS Policy on Declaration and Evaluation of Competing Interests](#))?

**If yes**, please provide details about any and all competing interests in the box below. Your response should begin with this statement: *I have read the journal's policy and the authors of this manuscript have the following competing interests:*

**If no** authors have any competing interests to declare, please enter this statement in the box: *"The authors have declared that no competing interests exist."*

\* typeset

#### Data Availability

PLOS journals require authors to make all data underlying the findings described in their manuscript fully available, without restriction and from the time of publication, with only rare exceptions to address legal and ethical concerns (see the [PLOS Data Policy](#) and [FAQ](#) for further details). When submitting a manuscript, authors must provide a Data Availability Statement that describes where the data underlying their manuscript can be found.

Your answers to the following constitute your statement about data availability and will be included with the article in the event of publication. **Please note that simply stating 'data available on request from the author' is not acceptable. If, however, your data are only available upon request from the author(s), you must answer "No" to the first question below, and explain your exceptional situation in the text box provided.**

Do the authors confirm that all data underlying the findings described in their

Yes - all data are fully available without restriction

<p>manuscript are fully available without restriction?</p>	
<p>Please describe where your data may be found, writing in full sentences. <b>Your answers should be entered into the box below and will be published in the form you provide them, if your manuscript is accepted.</b> If you are copying our sample text below, please ensure you replace any instances of <b>XXX</b> with the appropriate details.</p> <p>If your data are all contained within the paper and/or Supporting Information files, please state this in your answer below. For example, "All relevant data are within the paper and its Supporting Information files."</p> <p>If your data are held or will be held in a public repository, include URLs, accession numbers or DOIs. For example, "All <b>XXX</b> files are available from the <b>XXX</b> database (accession number(s) <b>XXX</b>, <b>XXX</b>)." If this information will only be available after acceptance, please indicate this by ticking the box below. If neither of these applies but you are able to provide details of access elsewhere, with or without limitations, please do so in the box below. For example:</p> <p>"Data are available from the <b>XXX</b> Institutional Data Access / Ethics Committee for researchers who meet the criteria for access to confidential data."</p> <p>"Data are from the <b>XXX</b> study whose authors may be contacted at <b>XXX</b>."</p> <p>* typeset</p>	<p>All relevant data are within the paper and its Supporting Information files.</p>
<p>Additional data availability information:</p>	

April 2nd, 2018

**Dear sirs**

We are writing to you to submit our manuscript “A critical assessment of tertiary structure-based phylogenies: the case of hepatitis C virus and human immunodeficiency virus-1 polymerases” by R. Jácome, A. Becerra, J. A. Campillo-Balderas, Y. López-Vidal, A. Lazcano, and S. Ponce de León, for its possible publication in *PLoS Computational Biology*. As discussed in our paper, in the recent years the field of structural biology has seen major advances, and the number of tertiary structures available in public databases keeps growing at an accelerated pace. Although the advantages of tertiary structure-based evolutionary analyses are well recognized, as of today the number of phylogenies based on crystal comparisons is quite limited and several methodological aspects have not been thoroughly addressed. The aim of our work is to analyze the effects that different factors such as crystal resolution and the conformational changes of proteins may have on the construction of evolutionary trees, which can lead to incorrect evolutionary inferences. Since the polymerases of the hepatitis C virus and the human immunodeficiency virus-1 are very attractive therapeutic targets, we have based our work on the comparison of the numerous crystal structures available in the public databases.

Since it is foreseeable that the tertiary structure-based phylogenies will be more frequently used in the future, we believe that *PLoS Computational Biology* is the proper journal to promote the use of tertiary structure-based phylogenies, and hence for publication of our work.

Sincerely,

**Antonio Lazcano****Professor**

1 **A critical assessment of tertiary structure-based phylogenies: the case of**  
2 **hepatitis C virus and human immunodeficiency virus 1 polymerases**

3 Rodrigo Jácome<sup>1¶</sup>, Arturo Becerra<sup>2&</sup>, José Alberto Campillo-Balderas<sup>2¶</sup>, Yolanda López-  
4 Vidal<sup>3&</sup>, Antonio Lazcano<sup>2,4&\*</sup>, Samuel Ponce de León<sup>1,5&</sup>

5 Affiliations

6 <sup>1</sup> División de Investigación, Facultad de Medicina, Universidad Nacional Autónoma de  
7 México, Ciudad de México, C.P. 04510, México

8 <sup>2</sup> Laboratorio de Origen de la Vida, Facultad de Ciencias, Universidad Nacional Autónoma  
9 de México, Ciudad de México, C.P. 04510, México

10 <sup>3</sup> Laboratorio del Microbioma, Facultad de Medicina, Universidad Nacional Autónoma de  
11 México, Ciudad de México, C.P. 04510, México

12 <sup>4</sup> Miembro de El Colegio Nacional, Ciudad de México, México

13 <sup>5</sup> Programa Universitario de Investigación en Salud, Universidad Nacional Autónoma de  
14 México, Ciudad de México, C.P. 04510, México

15

16 \* Corresponding author

17 E-mail: alar@ciencias.unam.mx

18

19 ¶These authors contributed equally to this work

20 &These authors also contributed equally to this work

21

## 22 **Abstract**

23 The number of tertiary structures deposited in the Protein Data Bank has been growing  
24 continually during the last decades. The use of tertiary structure-based phylogenies has not  
25 followed this trend, despite its potential to uncover remote evolutionary events that primary  
26 sequence approaches fail to detect. Owing to their importance as human epidemic pathogens  
27 and their significance as antiviral targets, two of the best characterized structures are the  
28 homologous hepatitis C virus and the human immunodeficiency virus 1 RNA-dependent  
29 polymerases, with more than 100 tertiary structures from each.

30 An evolutionary analysis of these two groups of viral polymerases was performed based on  
31 pairwise structural comparisons of selected ligand-free and ligand-bound structures from  
32 each of the proteins, followed by the calculations of structural distances and the construction  
33 of the corresponding dendograms.

34 We analyzed the effects that different crystallographic factors had on the topology of the  
35 resulting trees. While the ligand-free hepatitis C virus polymerases clustered into the  
36 corresponding genotypes and subtypes; the ligand-free reverse transcriptase dendogram  
37 failed to correctly cluster some of the structures. The hepatitis C virus polymerase and the  
38 human immunodeficiency virus-1 reverse transcriptase dendograms built with bound ligands  
39 clustered the structures according to the conformational changes triggered by the molecules,  
40 failing to reflect the accepted taxonomy of the viruses the structures belong to.

41 As with other databases, there is an important biomedical bias regarding the diversity of the  
42 hepatitis C and human immunodeficiency virus polymerases' structures, and few genotypes  
43 and subtypes are currently represented. As of today, very few works have covered  
44 methodological aspects of structure-based phylogenies. To the best of our knowledge, this is

45 the first analysis to address the way in which crystallographic factors like resolution and the  
46 presence of ligands alter the expected topology of tertiary structure-based dendograms, which  
47 could lead to incorrect evolutionary inferences.

48

#### 49 **Author summary:**

50 The field of protein tertiary structures has seen a significant growth in the last decades. Since  
51 the tertiary structure of proteins is more evolutionarily conserved than primary structure, the  
52 construction of phylogenies using the former has become a valuable tool in the study of deep  
53 evolutionary relationships. However, the use of this methodology remains limited and its  
54 advantages and drawbacks have not been extensively addressed. We have selected two  
55 widely studied tertiary structures for which more than 100 structures are available for each  
56 in public structural databanks: the hepatitis C virus polymerase and the human  
57 immunodeficiency virus-1 reverse transcriptase. We have constructed dendograms for each  
58 protein, using the comparison of ligand-free and ligand-bound tertiary structures, and  
59 analyzed the effect that different crystallographic factors such as resolution, the presence of  
60 mutations or the nature of the ligand had on the topology of the resulting trees. Our work  
61 shows that the construction of tertiary structure-based phylogenies has its nuances, and  
62 several factors have to be considered when selecting the structures and giving a biological  
63 perspective to the results.

#### 64 **Introduction**

65 The year 2014 was proclaimed as the International Year of Crystallography by the United  
66 Nations to honor the centennial of X-ray crystallography. That same year, the Protein Data  
67 Bank [1] surpassed the 100 000 tertiary structures, reaching over 135 000 structures by the

68 end of 2017, most of which have been obtained by X-ray diffraction. During the last 60 years,  
69 X-ray crystallography of proteins and nucleic acids has had a major impact on many diverse  
70 scientific disciplines including molecular biology, structural biology, pharmacology,  
71 medicine, and evolutionary biology [2]. Technical advances like high-throughput expression  
72 systems, crystallization robots, third-generation synchrotrons, reliable remote data collection,  
73 and significant improvements in computational tools for structure solution and refinement,  
74 are cornerstones in obtaining crystal structures of “tricky” molecules such as membrane  
75 proteins or gigantic macromolecular complexes like the ribosome [3]. It is foreseeable that  
76 the number of three-dimensional structures of proteins will continue to grow and its many  
77 valuable scientific applications will go along.

78

79 The amino acid sequence of a protein is the main determinant of its structure, which in turn  
80 translates into its specific function [4]. The study of evolutionary relationships using the  
81 primary structure of proteins has been a key factor in our understanding of the history of life  
82 on Earth. Nevertheless, as of today, establishing that two or more proteins are homologous  
83 comparing their sequences has shown its limits. The evolutionary links in a set of proteins  
84 turn blurry when the similarity of its sequences diminishes and enters what is known as “the  
85 twilight zone”, i.e. when the identity between two protein sequences goes below 20%-30%  
86 [5]. This is a common scenario when looking for distant homologous proteins. Moreover, the  
87 tertiary structure of a protein is more conserved than its primary structure [6, 7], making it a  
88 very valuable tool in the study of the evolution of proteins and organisms. It has been shown  
89 that proteins with very low levels of identity (below 20%) [8, 9] or different functions [10]  
90 may have similar folds with a considerable percentage of their C $\alpha$  atoms spatially  
91 superimposed. Accordingly, the comparison of tertiary structures of proteins has become an

92 alternative in the study of distant homologous proteins, including the construction of  
93 phylogenetic trees. Tertiary-structure based evolutionary trees have had a major impact on  
94 the study of deep phylogenies [11, 12], distantly related proteins [13-15] and evolutionary  
95 relationships between fast-evolving proteins such as those of RNA viruses [9].

96

97 Hepatitis C virus (HCV) and human immunodeficiency virus-1 (HIV-1) are the etiological  
98 agents of highly prevalent infections around the world [16, 17]. These two RNA viruses cause  
99 chronic infections and are a major economic burden to public health institutions, both in  
100 terms of duration of treatment as well as the long-term medical repercussions, e.g., HCV is  
101 the leading cause of hepatic transplant in adults in the U.S.A., and 1-4% of the chronically-  
102 infected patients will develop hepatocellular carcinoma [18]. Moreover, the lack of approved  
103 efficient vaccines for these viral infections has prompted the pharmacological industry, as  
104 well as the biomedical community, in the pursuit for effective and safe treatments capable of  
105 long-lasting viral control and/or eradication. Following the approval of Zidovudine in 1987  
106 for the treatment of HIV-1 infection, antiviral therapy has come a very long way. Today,  
107 there are six types of anti-HIV-1 [19] and three types of anti-HCV drugs [20] aimed at various  
108 molecular targets. These significant advances notwithstanding, viral factors such as the  
109 persistence of HIV-1 in latent memory CD4<sup>+</sup> cells [21], the high mutation rate of RNA  
110 viruses [22], and the high probability of preexisting resistant variants or their appearance  
111 soon after the initiation of treatment [23], have complicated the development of preventive  
112 and therapeutic measures.

113 Due to its critical role in the viral cycles, the RNA-dependent polymerase is one of the key  
114 molecular targets in the development of HCV and HIV-1 antivirals; this is an RNA-  
115 dependent RNA polymerase (protein NS5B) in the case of HCV, and a reverse transcriptase



116 (RT) in HIV-1. These homologous enzymes have the characteristic right-hand shape of the  
117 DNA and RNA polymerases Superfamily, with three conserved functional subdomains, i.e.  
118 palm, fingers and thumb, plus a connection and an RNase H domain in the case of the HIV-  
119 1 RT, and a fingertips subdomain in HCV NS5B which gives this latter a closed-hand shape  
120 (Fig 1). Like the other members of this enzyme superfamily, these polymerases have a two-  
121 metal ion mechanism of action in which two universally conserved aspartates located in the  
122 palm subdomain coordinate  $Mg^{2+}$  ions to achieve the nucleophilic attack [24]. The use of X-  
123 ray crystallography has played a key part in the discovery of new antiviral drugs and the  
124 understanding of paramount aspects such as their binding sites and the mechanisms  
125 underlying antiviral resistance [25, 26].

126 **Fig 1. Three-dimensional representation of HCV NS5B and HIV-1 RT.** (A) Hepatitis C  
127 virus NS5B apo structure (modified from PDB 1C2P). The subdomains are colored as  
128 follows: yellow – fingers; red – thumb; green – palm; orange – fingertips; grey – C-linker.  
129 (B) Human immunodeficiency virus 1 reverse transcriptase (modified from PDB 1DLO).  
130 The subdomains have the same colors as in a), the connection domain is in light blue and the  
131 RNase H domain is in dark blue.

132

133 Due to their biomedical significance, more than a hundred three-dimensional structures from  
134 each of these two enzymes have been determined and are available at public databases. As  
135 discussed here, this significant amount of data, allowed us to perform a thorough  
136 phylogenetic analysis of the available NS5B and RT crystal structures. We selected 31 HCV  
137 polymerases and 19 HIV-1 RT structures based on the presence or absence in these crystal  
138 structures of nucleic acids and the chemical diversity of bound ligands. Then we constructed  
139 dendograms by making structural comparisons between them to see the possibility of  
140 constructing structural phylogenetic trees of closely related RNA-dependent polymerases  
141 that can also reflect the empirical taxonomy for both viruses. Finally, we have critically

142 assessed the effects of the different inherent crystallographic factors (e.g. resolution, bound  
143 substrates and ligands) in the construction of tertiary structure-based evolutionary trees.

144

## 145 **Results**

### 146 **Hepatitis C virus**

147 **Biological diversity.** Parallel and complementary searches with the terms “hepatitis c virus  
148 AND polymerase” were performed on the RCSB PDB [1] and the European PDB [27]. This  
149 strategy was implemented since the taxonomy identifier assigned to all the HCV structures  
150 in the RCSB PDB is limited to “Taxonomy ID: 11103 – Hepatitis C virus”, which did not  
151 allow us to assign the structures to a specific taxon. The search in the PDBe allowed us to  
152 identify 192 (as of August 2017) NS5B structures belonging to different HCV genotypes and  
153 subtypes (Table 1).

154

155

156

157

158

159

160 **Table 1. Number of the available Hepatitis C virus NS5B crystal structures.**

Genotype	Subtype	Isolate	Number of structures
Genotype 1	1b	HC-J4	50
		BK	36
		Con1	2
	1a	Not specified	27
		H	1
Genotype 2	2a	H77	2
		1	1
		Not specified	5
	2b	JFH-1	21
		HC-J6	2
Unclassified	Not specified	4	
	HC-J8	2	
Total			192

168

169 The NS5B structures belong to genotypes 1 (124 structures, 65%) and 2 (29 structures, 15%).  
170 Thirty-nine (20%) of the structures are only tagged with the Hepatitis C virus taxonomy  
171 identifier 11103. As shown in Table 1, the structures belong to only four subtypes: subtype  
172 1b is the most abundant (115/192), followed by subtype 2a (27/192), subtype 1a (9/192) and  
173 subtype 2b (2/192). From the 153 structures that could be assigned, 70% belong to three  
174 isolates: subtype 1b isolate HC-J4, subtype 1b isolate BK, and subtype 2a isolate JFH-1.

175 **Dendograms.** Once the structures had been grouped into genotypes and subtypes, we looked  
176 for structures that did not have nucleic acids bound to the active site nor small molecules  
177 bound to the distinct allosteric sites previously described for NS5B [28, 29]. We also sought  
178 structures with a similar number of amino acid residues present in the crystallographic  
179 structures. Six structures belonging to different HCV subtypes were identified that fulfilled  
180 the aforementioned criteria: (1) genotype 1 - subtype 1a isolate H77 (PDB 2XI2); (2) subtype  
181 1b BK (PDB 1C2P); (3) subtype 1b isolate HC-J4 (PDB 1NB4); (4) genotype 2 - subtype 2a

182 isolate HC-J6 (PDB 2XWH); (5) subtype 2a JFH-1 (PDB 3I5K); and (6) subtype 2b isolate  
 183 HC-J8 (PDB 3GSZ). A distance-based dendrogram was then built using the Structural  
 184 Alignment Score (SAS) [30] calculated from the pairwise comparisons of the selected  
 185 structures. Table 2 shows a list of the structures used for the construction of the dendrograms.  
 186

187 **Table 2. List of the HCV NS5B crystal structures used in the construction of ligand-free**  
 188 **and ligand-bound dendrograms.**

PDB ID	Subtype	Resolution (Å)	Ligands	Ligand chemical group	Allosteric binding site
1C2P	<b>1b BK</b>	1.9	-		
1NB4	<b>1b J4</b>	2	-		
2XI2	<b>1a H77</b>	1.8	-		
2XWH	<b>2a J6</b>	1.8	-		
3I5K	<b>2a JFH1</b>	2.2	-		
3GSZ	<b>2b J8</b>	1.9	-		
1NB6	<b>1b J4</b>	2.6	UTP		
2XI3	<b>1a H77</b>	1.7	GTP		
4WTJ	<b>2a JFH1</b>	2.2	dsRNA		
3H98	<b>1bBK</b>	1.9	B5P	Benzothiadiazine	Palm
3CO9	<b>1bBK</b>	2.1	3MS	Pyridazinone	Palm
2QE5	<b>1bBK</b>	2.6	617	Anthranilic acid derivative	Palm
2D3Z	<b>1bBK</b>	1.8	FIH	Thiophene-based inhibitor	Thumb NNI2
1NHV	<b>1bBK</b>	2.9	154	Phenylalanine derivative	Thumb NNI2
4KB7	<b>1bBK</b>	1.85	690	Benzofuran	Palm
4EO8	<b>1bBK</b>	1.8	o53	Tri-substituted acylhydrazine	Thumb NNI2
4JJS	<b>1bJ4</b>	2.2	1M9	Anthranilic acid derivative	Thumb NNI2
4JU4	<b>1bJ4</b>	2.4	1O3	Anthranilic acid derivative	Thumb NNI2
4JU1	<b>1bJ4</b>	2.9	1NZ	Quinazolinone	Thumb NNI2
4JTY	<b>1bJ4</b>	2.6	1NV	Quinazolinone	Thumb NNI2
4J08	<b>1bJ4</b>	2.1	1JH	Anthranilic acid derivative	Thumb NNI2
3HKY	<b>1bJ4</b>	1.9	IX6	Benzodiazepine pol inhibitors	Palm
3GOL	<b>1bJ4</b>	2.85	XND	Benzodiazepine pol inhibitors	Palm
3UPH	<b>1bJ4</b>	2	0C1	Dihydrofuranoindole	Palm
3U4O	<b>1bJ4</b>	1.77	08E	Indole C2acyl sulfonamide	Palm
3FQL	<b>1bCon1</b>	1.8	79Z (Nesbuvir)	Benzofuran	Palm
4KHM	<b>1<sup>a</sup></b>	1.7	1PV	Benzofuran	Palm
3HKW	<b>1a 18</b>	1.55	IX6	Benzodiazepine pol inhibitors	Palm
3QGH	<b>1a Bartenschlager</b>	2.14	63F	N-benzyl-4-heteroaryl-1(phenylsulphonyl)piperazin-2-carboxamide	Palm

5TWM	2aJFH-1	1.97	7NG	Benzofuran	Palm
3HVO	2b J8	2	VGI	1H-benzo[de]isoquinoline-1,3 (2H)-diones	Thumb NNI2

189 dsRNA = double-stranded RNA; NNI = non-nucleotide inhibitor

190 The unrooted dendogram (Fig. 2A) shows two distinct branches, each one clustering one of  
191 the two HCV genotypes. Within the two genotypes, each subtype groups well in a different  
192 clade. The average root mean square deviation (RMSD) when comparing HCV NS5B  
193 structures of the same genotype was, for genotype 1 - 0.63 Å for 558 residues and 89%  
194 identity, and for genotype 2 - 0.51 Å for 557 residues and 90% identity. When comparing the  
195 structures of different genotypes the average RMSD was: 0.78 Å for 557 residues and 74%  
196 identity.

197 **Figure 2 Dendograms built by the pairwise comparison of hepatitis C virus NS5B**  
198 **tertiary structures.** (A) HCV NS5B ligand-free unrooted dendogram. The branches are  
199 colored as follows: green – genotype 1; blue- genotype 2. The information at each node  
200 includes: the genotype and the subtype of the structure and its PDB code. (B) Unrooted  
201 dendogram of HCV NS5B structures with bound ligands. The branches are colored as  
202 follows: dark blue – subtype 1b BK; cyan – subtype 1b J4; purple – subtype 1b Con1; light  
203 green – subtype 1a H77; dark green – subtype 1a; orange – subtype 2a; red – subtype 2b. The  
204 apo structures are those with a colored frame. The information at each node includes: the  
205 PDB code and the allosteric binding site of its ligand (TH – thumb; P – Palm). The colored  
206 shapes correspond to the different chemical groups ligands belong to: blue circle – double-  
207 stranded RNA; blue arrow – GTP; blue triangle – UTP; yellow circle – benzofuran; yellow  
208 triangle – benzodiazepine; yellow arrow – carboxamide; star – quinazolinone; rectangle –  
209 anthranilic acid derivative; trapezium – phenylalanine derivative; pentagon – thiophene-  
210 based inhibitors; hexagon – dihydrofuranoindole; arc – indole sulfonamide; cross –  
211 benzothiadiazine; X – pyridazinone; rhombus – acylhydrazine.

212

213 The effect that the ligands might have for the construction of dendrograms based on the  
214 comparison of tertiary structures was then analyzed. To do so, we added structures with  
215 different types of ligands including nucleotides (UTP and GTP) and a double-stranded RNA  
216 comprising a template and a primer, as well as small molecules belonging to different

217 chemical groups bound to the HCV NS5B allosteric sites, thumb non-nucleoside inhibitor  
218 binding site (NNI) 2 and palm NNI (Table 2) [28, 29]. Although most works describe two  
219 NNI palm binding sites [28, 29], we considered them as one [31, 32], since many of the  
220 residues the ligands interact with overlap. The subtypes' structures that had not been included  
221 in the previous step due to the presence of ligands, i.e. subtypes 1b isolate Con1 and subtype  
222 1a isolate 1, were considered for this set of pairwise structural comparisons. We did not  
223 include structures with ligands bound to the thumb NNI1 allosteric site because the available  
224 structures have been genetically modified and large fragments of the NS5B structure have  
225 been deleted from the crystallographic structure.

226 Unlike the dendrogram built with the NS5B ligand-free structures, in which the branches  
227 cluster the distinct HCV genotypes and subtypes accordingly, in this tree only genotypes are  
228 grouped (Fig 2B). One of the main branches clusters genotype 2 structures with two  
229 exceptions, which are discussed below. This major branch is further divided in two smaller  
230 ones. One of them includes the structures from subtype 2a, while the other includes structures  
231 from subtype 2b: the structure without ligands and one structure with VGI, an NNI bound to  
232 the thumb 2 allosteric site (PDB 3HVO). The structure of subtype 1a bound to a carboxamide  
233 (PDB 3QGH) stems from this branch prior to the genotype 2 structures (*vide infra* for an  
234 extended discussion).

235 The genotype 1 branches do not separate the corresponding subtypes, and the ligand-free  
236 structures are interspersed with those crystallized in the presence of ligands. The structures  
237 from genotype 1 cluster in three big branches (Fig 2B).

238 One of the branches groups all the subtype 1b BK structures and three subtype 1b J4 crystals.  
239 The polymerases with ligands bound to the palm-binding site are closer to the ligand-free 1b

240 BK crystal (1C2P), while those with ligands bound to the thumb NNI2 site are further away.  
241 Stemming close to the origin of this branch there are two branches with subtype 1b J4  
242 crystals. One of these branches groups the ligand-free structure and the polymerase with a  
243 nucleotide bound to the active site (1NB4 and 1NB6, respectively), and another branch that  
244 corresponds to a structure bound to a benzodiazepine inhibitor in the palm allosteric site  
245 (3GOL).

246 The next big branch clusters structures from various subtypes with small molecules bound to  
247 the palm allosteric site plus the subtype 1a H77 ligand-free crystal. Close to the origin of this  
248 clade, we observe a branch clustering three subtype 1b J4 structures with diverse types of  
249 inhibitors bound close to the active site, namely benzodiazepine pol inhibitor (3HKY), indole  
250 sulfonamide-based inhibitor (3U4O) and a dihydrofuranoindole (3UPH). Next, there is a long  
251 branch of a subtype 2a polymerase with a benzofuran inhibitor (5TWM), followed by a  
252 branch that includes the 1a H77 ligand-free crystal (2XI2) plus the structure with bound GTP  
253 (2XI3), and another branch comprising two structures with benzofuran inhibitors (subtype  
254 1a - 4KHM and a subtype 1b Con1 - 3FQL) and one structure with a benzodiazepine inhibitor  
255 (3HKW).

256 Finally, there is one branch clustering all the subtype 1b J4 structures with anthranilic acid  
257 derivatives and quinazolinones bound to the thumb NNI2 site, plus a very long branch  
258 corresponding to the subtype 2a polymerase with a double-stranded RNA consisting of a 4-  
259 mer template and 2-mer primer bound to the structure (4WTJ).

260 To have a better understanding of the tree topology, a more detailed structural analysis of  
261 some pairwise comparisons was performed using the programs MatchMaker and  
262 Match/Align included in the UCSF Chimera software [33]. These programs allowed us to

263 make the corresponding three-dimensional superpositions and visualize the local and global  
264 conformational changes occurring in the NS5B protein when the ligands are bound, and  
265 which alter the topology of the ligand-free tree.

266 As shown in Fig 3, most of the conformational changes occurring in NS5B in the presence  
267 of ligands are found in the thumb subdomain and its surrounding structures including the tip  
268 of the fingertips and the C-terminal linker structure [34].

269 **Figure 3. Structural superposition of representative HCV NS5B structures with the**  
270 **apo structure (PDB 1C2P) using the MatchMaker and Match/Align programs.** (A)  
271 Three-dimensional rendering of NS5B; colors are the same as in Fig 1. Bottom: NS5B “top”  
272 view. (B) Structural superposition of NS5B bound to thumb NNI2 inhibitor (PDB 2D3Z);  
273 right – “top” view. (C) Structural superposition of NS5B bound to palm NNI inhibitor (PDB  
274 3UPH); right – “top” view. (D) Structural superposition of NS5B bound to palm NNI  
275 inhibitor (PDB 3FQL); right – “top” view. The color palette in (B), (C) and (D) is set  
276 according to the distance between the superimposed  $\alpha$ -carbon atoms (in Å). The ligand of the  
277 three structures is colored in green.

278 Different ligands are bound to the thumb NNI2 allosteric site, including phenylalanine  
279 derivatives (1NHV), thiophene-based sulphonamides (2D3Z), quinazolinones (4JU1, 4JTY),  
280 and anthranilic acid derivatives (4J08, 4JU4, 4JJS). Most of the conformational changes  
281 associated with these molecules are located in the thumb subdomain, not only in the allosteric  
282 binding site, but also in the residues surrounding the C-terminal linker structure [34] and the  
283 fingertips (Fig 3B).

284 Ligands bound to the palm allosteric site include dihydrofuranoindoles (PDB 3UPH,  
285 molecule 0C1), indole C2 acyl sulfonamides (PDB 3U4O, molecule 08E), 1,5  
286 benzodiazepine derivatives (PDB 3HKW, molecule IX6), pyrimidine derivatives like  
287 molecule B5P (PDB 3H98), and 2-acylbenzofuran flavonoids such as molecule 79Z (PDBs  
288 3FQL). As Figure 3C shows, the binding of ligands to this allosteric site causes  
289 conformational changes in the fingertips, the residues between structural motifs C and D of



290 the palm subdomain, and the C-terminal linker. The superposition of 3FQL with the apo  
291 structure 1C2P (Fig 3D) shows larger conformational changes (compared to 3UPH, Fig 3C)  
292 in several structures of the thumb subdomain, the C-terminal linker, the thumb residues  
293 interacting with it, the fingertips, the residues 85 to 109 located at the “top” of the fingers  
294 subdomain, and the residues between structural motifs C and D of the palm subdomain.

295 The only three structures that did not cluster with their corresponding genotypes are 3QGH,  
296 5TWM and 4WTJ. As shown in S1 Table, these structures have the highest RMSDs and the  
297 smallest number of superimposed residues in the all-against-all pairwise comparisons.

298 To understand these differences, all the polymerases with ligands bound to the palm allosteric  
299 binding sites were superimposed. The structural conservation among the superimposed  
300 polymerases is very high, and most of its atoms have an RMSD below 1 Å; some elements  
301 from the thumb subdomain and from the “top” of the fingers subdomain show higher mobility  
302 with RMSDs near 1 Å. The most mobile structures are located in the fingertips subdomain  
303 and at the proteins’ C-terminus linker structure, adjacent to the priming loop. These regions  
304 have an RMSD equal or above 2 Å.

305 The structural superposition shows a different conformation of 3QGH near the highly mobile  
306 C-terminus linker structure. Instead of the long loop present in the polymerases, the residues  
307 Gly549 to Ile560 form a short helical element followed by a  $\beta$ -strand (Fig 4A). In the case of  
308 crystal 5TWM, the superposition shows that in all the structures, residues 25 to 35 form two  
309 short helices. However, the first short helix is not present in 5TMW, shortening this portion  
310 of the protein (Fig 4B), which reduces the number of superimposed C $\alpha$  atoms in the structural  
311 comparisons.

312 **Figure 4. Structural superposition of the structures that did not cluster with their**  
313 **corresponding genotypes.** (A) Close up of the C-linker region highlighting (in yellow) the  
314 structure of 3QGH, which is different from the rest of the structures. (B) Close up of the  
315 fingertips region highlighting (in green) the structure of 5TWM, which is different from the  
316 rest of the structures. (C) Structural superposition between 1C2P and 4WTJ showing the  
317 conformational changes caused by the presence of a double-stranded RNA. The range of  
318 RMSDs (in Å) between the C $\alpha$  of the two structures goes from blue (0 Å) to red ( $\geq 2$  Å).

319 As shown in S1 Table, the structure with the smallest number of superimposed residues and  
320 the one with the highest RMSDs is 4WTJ, which corresponds to a subtype 2a polymerase  
321 with a double-stranded RNA consisting of a 4-mer template and 2-mer primer bound. The  
322 structural superposition of this crystal with a ligand-free structure (Fig 4C) shows that the  
323 protein has a larger number of conformational changes. The N-terminus fragment and the  
324 thumb subdomains are the moieties with the highest mobility, and a considerable proportion  
325 of their residues have RMSDs above 2 Å, or cannot be superimposed since they are out of  
326 the algorithm's range.

327

### 328 **Human immunodeficiency virus 1**

329 **Biological diversity.** In the case of HIV-1, we used a similar strategy and made  
330 complementary searches in the RCSB PDB and the PDB in Europe, since the Taxonomy  
331 identifier assigned to all the HIV-1 structures in the former is “Taxonomy ID – 11676 –  
332 Human immunodeficiency virus 1”. A search in the PDBe with the terms “Human  
333 immunodeficiency virus” and “reverse transcriptase” yielded 222 results (as of August 2017),  
334 with the following distribution: HIV type 1 BH10: 69 structures (31%); HIV-1  
335 M:B\_HXB2R: 64 structures (29%); HIV type (CLONE 12): 1 structure; Human  
336 immunodeficiency virus 2: 1 structure; unclassified HIV 1: 87 structures (39%). The viral  
337 source of structure 4ZHR was the recombinant form NY5/BRU, according to the original  
338 report [35].

339 **Dendograms.** For the HIV-1 RT structural comparisons, only the p66 subunit was  
 340 considered, since it is the catalytic subunit in the p66/p51 heterodimer that forms the HIV-1  
 341 reverse transcriptase.

342 Firstly, we sought the ligand-free RT structures, i.e., the enzyme without nucleic acid  
 343 complexes or bound ligands. We identified ligand-free structures from strains BH10, HXB2  
 344 and the recombinant NY5/BRU. The list of crystal structures used for the dendograms'  
 345 construction is shown in Table 3.

346 **Table 3. List of the HIV-1 RT crystal structures used in the construction of ligand-free**  
 347 **and ligand-bound dendograms.**

PDB ID	Strain	Resolution (Å)	Mutations	Substrate/Ligands	Type	Template length (nt)	Primer length (nt)
1DLO	BH10	2.7	C280S	NA			
1HMV	BH10	3.2	0	NA			
1HQE	BH10	2.7	C280S - K103N	NA			
1HVU	BH10	4.75	0	NA			
1QE1	BH10	2.85	M184I - C280S	NA			
3KLI	BH10	2.65	M41L, D67N, K70R, T215Y, K219Q, Q258C, C280S, I559V	NA			
4ZHR	NY5/BRU recombinant clone pNL4-3	2.6	0	NA			
1JLE	HXB2	2.8	Y188X, C280X	NA			
1RTJ	HXB2	2.35	0	NA			
2YKN	BH10	2.12	N57S, F227L, E478Q	YKN	NNRTI		
4ICL	BH10	1.8	K172A, K173A, C280S	Rilpivirine	NNRTI		
1IKV	BH10	3	K103N, E478Q	Efavirenz	NNRTI		
1JKH	HXB2	2.5	Y188X, C280X	Efavirenz	NNRTI		
3MEG	HXB2	2.8	K103N	Rilpivirine	NNRTI		
1TVR	Clone12	3	C280S	TB9	NNRTI		
3KLG	BH10	3.65	M41L, D67N, K70R, T215Y, K219Q, Q258C, C280S, I559V	DNA/DNA		27	22
3V6D	BH10	2.75	Q258C, C280S, D498N	DNA/DNA		27	21
5J2M	HXB2	2.43	Q258C, C280S	DNA/DNA		27	21
3KJV	HXB2	3.1	Q258C, C280S	DNA/DNA		27	21

348 NNRTI = Non-nucleotide reverse transcriptase inhibitors; nt = nucleotides

349 The RMSD for the BH10 ligand-free structures was 1.123 Å for an average of 536  
350 superimposed residues and a SAS of 0.21. The mean SAS for the comparisons between the  
351 BH10 and the recombinant NY5/BRU structures was 0.266 (mean RMSD 1.425 Å for 536  
352 superimposed residues). The mean SAS for the comparisons between the BH10 and the  
353 HXB2 structures was 0.467 (mean RMSD 2.075 Å for 449 superimposed residues) (S2  
354 Table). The dendrogram (Fig 5A) shows two distant, well-resolved branches, each one  
355 clustering the structures from strains BH10 and HXB2, respectively. Interspersed between  
356 these two clades, there are two branches of BH10 structures (PDBs 1HMV and 1HVU) with  
357 the branch of the recombinant form NY5/BRU between them. It is important to note that  
358 these two BH10 structures that did not cluster with the other BH10 structures are the ones  
359 with the lowest resolution (3.2 Å and 4.75 Å, respectively).

360 **Figure 5. HIV-1 RT ligand-free dendrogram and structural superposition of BH10 and**  
361 **HXB2 RT structures.** (A) HIV-1 RT ligand-free unrooted dendrogram. The branches are  
362 colored as follows: blue – strain BH10; red – group M subtype B HXB2; yellow –  
363 recombinant form NY5/BRU. The information at each node includes: the PDB code and the  
364 resolution of the structure. (B) Structural superposition of BH10 and HXB2 VIH-1 RT  
365 structures using the MatchMaker and Match/Align programs. The BH10 structure is in gold;  
366 the HXB2 structure is in light blue. The red line represents the distance between residues  
367 Asp256 of both structures (in Å).

368  
369 It is interesting to note that there is a difference of approximately 90 superimposed residues  
370 when the HXB2 structures are compared to the structures from strains BH10 and NY5/BRU,  
371 which results in large structural distances. A visual inspection of a structural alignment  
372 between 1DLO (strain BH10) and 1RTJ (strain HXB2) (Fig 5B) shows that the BH10  
373 structure is in a closed conformation in which the thumb and several fragments of the fingers  
374 subdomain are rotated enclosing the active site, e.g. the distance between thumb residue Asp  
375 256 of both structures is 16.09 Å (Fig 5B). A conformational change is also evident in the  
376 loop connecting palm's motifs D and E, and in the RNaseH domain. However, most of the

377 palm subdomain and the connection domain do not seem to undergo major conformational  
378 changes.

379 We then looked for structures crystallized in the presence of different substrates and ligands,  
380 i.e. protein-nucleic acid complexes and protein-NNRTI complexes (Table 3). In the case of  
381 protein-nucleic acid complexes, we identified crystal structures from strains BH10 and  
382 HXB2 with double-stranded DNA in which the template and the primer strands had a similar  
383 length. In the case of protein-NNRTI complexes, we sought for structures from different  
384 strains with the same bound ligands. We identified BH10 and HXB2 structures with efavirenz  
385 and rilpivirine bound to the active site. We also added the RT from HIV-1 Clone 12 (PDB  
386 1TVR), that is bound to molecule TB9, a NNRTI, and another BH10 structure with molecule  
387 YKN (PDB 2YKN), which is also a NNRTI.

388 The lowest SAS and RMSD, and the largest number of superimposed residues were obtained  
389 when similar complexes were compared i.e. ligand-free vs ligand-free, protein-DNA vs  
390 protein-DNA, and protein-ligand vs protein-ligand (S2 Table). The SAS rises when different  
391 complexes are compared, which is mainly due to the reduced number of superimposed  
392 residues between the complexes (S2 Table).

393 The unrooted dendogram (Fig 6A) has two main branches. The first of them is subdivided in  
394 two clades, one that includes the BH10 strain ligand-free structures and another one that  
395 includes the structures with bound DNA and the ligand-free structure from the recombinant  
396 NY5/BRU strain (4ZHR) diverging close to the node.

397 **Figure 6. HIV-1 RT dendogram with bound ligands and structural superposition of**  
398 **representative structures.** (A) Unrooted dendogram of HIV-1 RTs with bound ligands. The  
399 colors are as follows: blue – strain BH10; red – group M subtype B HXB2; yellow –  
400 recombinant form NY5/BRU; green – CLONE 12. The information at each node includes:  
401 the PDB code, the resolution of the structure and the bound substrate/ligand. (B) Structural

402 superposition of selected VIH-1 RT structures using the MatchMaker and Match/Align  
403 programs. The BH10 apo structure is in gold (PDB 1DLO); the HXB2 apo structure is in  
404 light blue (1RTJ); the BH10 structure with double-stranded DNA is in purple (PDB 3V6D);  
405 the HXB2 structure with Rilpivirine is in yellow (PDB 3MEG).

406 The second main branch can also be subdivided. One of these subdivisions includes  
407 structures from HXB2 strain, namely the two ligand-free structures (1JLE & 1RTJ) and a RT  
408 bound to efavirenz (1JKH). The second subdivision includes all the other structures with  
409 NNRTIs. One of its branches includes the RT bound to YKN (2YKN), the HXB2 RT with  
410 rilpivirine (3MEG) and the BH10 RT with efavirenz (1IKV), while the other branch groups  
411 the BH10 RT bound to rilpivirine (4ICL) and the Clone 12 RT structure bound to TB9  
412 (1TVR). It is interesting to note that the RT structures bound to efavirenz are not the closest  
413 from one another (*vide infra*).

414 We took one structure from each of the four branches described above and made a structural  
415 superposition in Chimera's MatchMaker and Match-Align algorithms (Fig 6B). As shown in  
416 Fig 6B, the RNase H and the Connection domains superimpose with only minor changes in  
417 some of the loops. However, the polymerase's subdomains undergo major structural  
418 rearrangements. The palm subdomain seems to be the one with the fewest motions, although  
419 some of its elements show an evident flexibility, e.g. the loop connecting palm's motif D  
420 with motif E and motif E, *per se*, have different rearrangements. As mentioned above, the  
421 thumb subdomain shows an open and a closed configuration. The thumb of the protein-DNA  
422 complex (3V6D), the protein-ligand complex (3MEG) and the HXB2 ligand-free structure  
423 (1RTJ) exhibit an open configuration, while the BH10 ligand-free structure is in the closed  
424 configuration. The fingers subdomain also shows different conformations. The presence of  
425 double-stranded DNA pulls this subdomain towards the active site, while the binding of  
426 rilpivirine in the palm's non-nucleoside inhibitor binding site pushes the fingers away.

## 427 **Discussion**

428 In this article, we have analyzed the different HCV NS5B and HIV-1 RT tertiary structures  
429 available in public structural data banks. Pairwise structural comparisons were performed  
430 between a set of selected structures from each of the two replicases, and two dendograms  
431 were built for each polymerase: one with ligand-free structures, and one with bound ligands  
432 (nucleic acids or non-nucleoside polymerase inhibitors). The aim of the structure-based  
433 dendograms was to analyze the possibility of using this methodology to study evolutionary  
434 links between closely related proteins, and to study the effects of different protein  
435 crystallization properties like the resolution level and the presence of ligands in the  
436 construction of these dendograms.

437 The field of protein tertiary structures determination has seen major breakthroughs in the past  
438 two decades and the number of structures as well as their diversity has grown exponentially.  
439 However, even with the introduction of high-throughput techniques and structural genomics  
440 projects [36] that reduce significantly the limitations of tertiary-structures' obtention, the  
441 field is still far from the primary-structure sequencing techniques that can obtain millions of  
442 sequences from one single experiment. For instance, a search in the NCBI protein databank  
443 with the terms "Human immunodeficiency virus 1" and "reverse transcriptase" yields more  
444 than 260 000 results, compared with the 222 HIV-1 RT structures available in the PDBe.  
445 This difference is even more remarkable when the biological diversity of the available  
446 structures is considered.

447 Hepatitis C virus is classified in seven genotypes, 67 subtypes and more than 20 provisional  
448 subtypes [37]. It is therefore somewhat disappointing that only the polymerases from two  
449 viral genotypes and four subtypes have been crystallized, and that 60% of these structures

450 belong to one subtype. Moreover, from a biomedical perspective, as of today, there are no  
451 HCV genotype 3 NS5B crystal structures. Genotype 3 has been associated with higher rates  
452 of hepatic steatosis, increased fibrosis progression, and hepatocellular carcinoma [38] as well  
453 as lower SVR rates compared to the other genotypes in Sofosbuvir-based treatments,  
454 particularly in patients with cirrhosis [39, 40]. Different groups have performed  
455 bioinformatics analyses of HCV genotype 3 NS5B in an attempt to understand its clinical  
456 and biochemical particularities [41, 42]. Molecular docking calculations with Sofosbuvir  
457 predicted better binding-affinities for HCV genotypes 2 and 1 compared to genotype 3 [41].  
458 Moreover, Mugosa et al. [42] found eight highly variable residues in genotype 3A NS5B  
459 sequences. Most of these residues were located in the palm subdomain, but in their three-  
460 dimensional model of the protein, only position 219 had a functional role, and the rest of the  
461 variable residues appeared to have a structural role [42]. Considering that approximately 30%  
462 of the total number of HCV infections are caused by genotype 3 [43], and that the genetic  
463 distance between genotypes is considerable, the lack of its NS5B structure represents a major  
464 limitation for the improvement of the current treatment regimes.

465 Four HIV-1 lineages can be recognized, all of which can be tracked to independent cross-  
466 species transmission [44]. Group M is the most prevalent and has a worldwide distribution,  
467 group O is responsible for several thousand cases in Western Africa, while groups N and P  
468 have only been reported in a handful of cases in Cameroon. Globally, group M subtype C is  
469 the most prevalent, while the subtype B is the most prevalent in America and Western Europe  
470 [44, 45]. The biological diversity of the available crystallized reverse transcriptases is also  
471 quite limited. These RT structures come from only four strains, all of which are laboratory-  
472 adapted strains and, from these, only the Group M subtype B HXB2 can be properly



473 classified; the rest are only tagged as “HIV-1 unknown group”. Most of the biochemical and  
474 structural studies in which the effects of anti-HIV medications are tested have been  
475 performed using group M subtype B strains [46]. However, like every other RNA virus, HIV-  
476 1 has an extremely high mutation rate and fast replicative cycles, which is reflected in genetic  
477 distances between subtypes of approximately 30%. It is therefore not surprising to find  
478 differences between subtypes in terms of replicative fitness, viral loads, transmission routes,  
479 and drug resistance [47]. We can expect to find a wider diversity of HIV-1s in future  
480 biochemical and structural studies if more efficient and universal treatments are pursued.

481 The current biological diversity of the NS5B and the RT structures exhibits a clear  
482 biomedical bias, and it is easy to understand that most of the available structures belong to  
483 strains, which are more prevalent in the United States and Europe. However, it would be  
484 important to have a wider array of structures to obtain a better insight on the structural  
485 diversity of these proteins and a better understanding of certain biomedical factors like  
486 treatment response rates and antiviral resistance.

487 The construction of phylogenetic trees using the tertiary structures of proteins is not new  
488 [30], but its use has not been as extended as the primary structure-based approach. The main  
489 goal of tertiary structure-based phylogenetic trees has been the unraveling of evolutionary  
490 links between proteins with a high degree of structural conservation, but low levels of  
491 sequence conservation, usually below the “twilight zone” [5], where “traditional”  
492 phylogenetic approaches fail to provide robust evolutionary hypothesis [9, 48-50].  
493 Nevertheless, the structure-based methodology has its downsides and there are several  
494 aspects related to the structure *per se* that must be taken into account.

495 The HCV NS5B ligand-free dendrogram (Fig 2A) shows that the methodology allowed the  
496 separation of the structures according to genotype and subtype, even when the genetic  
497 distance between subtypes is small, i.e. the genetic distance between genotypes is  
498 approximately 30%, while the distance between subtypes is close to 15%. Since the structural  
499 distance between genotypes is sufficiently large, the separation can be recognized even when  
500 ligand-bound structures were included (Fig 2B).

501 However, as shown in Figure 5A, in the case of the HIV-1 ligand-free RT dendrogram, such  
502 clear subtype clustering is not observed. This might be due to the BH10 RT structures wide  
503 array of resolutions. Perhaps not surprisingly, structures with a resolution better than 3 Å are  
504 closer than those with lower resolutions. Structures with poorer resolutions are less accurate,  
505 and there might be significant differences regarding the precise location of the  $\alpha$ -carbon  
506 backbone atoms compared to the same structures with better resolutions [51]. This is  
507 reflected in that the main difference in the comparisons between the BH10 structures is the  
508 RMSD and not the number of aligned residues. This phenomenon is equivalent to long-  
509 branch attraction in primary-sequence based phylogenies, which is known to severely alter  
510 tree topologies [52].

511 When ligands are considered for the pairwise structural comparisons, the topology of the  
512 dendograms changes in a significant way (Figs 2B and 6A). Instead of clustering according  
513 to genotypes and subtypes, the structures cluster depending on the type of bound substrate  
514 and the conformational changes caused by those molecules.

515 In the case of HCV, the fact that there is basically no difference between the number of  
516 superimposed residues when structures with and without ligands are considered implies that  
517 the topology is given by the RMSD of each comparison, i.e., the tree topology is the outcome

518 of the conformational changes occurring in the presence of ligands in the crystal structure. A  
519 similar work by Caillet-Saguy et al [34] analyzed the conformational variability of HCV  
520 subtype 1b NS5B in the presence of NNIs, and found that the protein comprises several  
521 mobile fragments, some of them far from the NNI allosteric binding sites. The displacement  
522 of these regions in the presence of NNI may disturb the conformational changes required by  
523 the thumb subdomain to attain a proper initiation configuration providing a common  
524 mechanism of action for these dissimilar molecules [34]. Our analysis identified the same  
525 mobile regions on NS5B and similar conformational changes in a wider diversity of subtypes.

526 The main determinant of the ligand-bound HIV-1 RT dendrogram topology is the nature of  
527 the complex, i.e. ligand-free structures, protein-nucleic acid complexes, and protein-inhibitor  
528 complexes. The reverse transcriptase shows an enormous range of conformational changes,  
529 with the biggest motions taking place in the fingers and thumb subdomains [53, 54]. Without  
530 ligands, the molecule has a closed conformation in which the fingers and thumb subdomains  
531 block the active site; in the presence of a primer/template, these subdomains rotate and  
532 “grasp” the double-stranded nucleic acids. Finally, in the presence of an incoming dNTP, the  
533 fingers subdomain rotates and allows catalysis to take place [55]. In the presence of a non-  
534 nucleoside RT inhibitor, the conformational changes are distinct: the palm subdomain’s  
535 primer grip is in an open conformation that limits the motions of the thumb subdomain, which  
536 will remain in an open conformation, thus interfering with the movements required to  
537 accommodate the incoming nucleotides and preventing further polymerization [55]. Previous  
538 structural comparisons of RTs structures bound to rilpivirine gave large RMSDs values and  
539 showed that the protein can adopt different conformations in the presence of the same drug  
540 [56]. The position of one of the drugs’ chemical moieties within the active site is different in

541 the wild type structure compared to the mutated ones, which, in turn, changes the position of  
542 two  $\beta$  sheets [56]. This might explain the fact that structures 3MEG and 4ICL, both with  
543 rilpivirine, but with different mutations (3MEG K103N; 4ICL K172A, K173A, C280S), are  
544 found on different branches in our dendrogram (RMSD 1.740 Å for 537 superimposed  
545 residues, SAS 0.340). The number of superimposed residues diminishes considerably when  
546 different RT complexes are compared, even when the identity between the proteins is always  
547 above 96%. In this case, the artifact stems from the fact that the displacement of the structural  
548 elements is so large that the different algorithms used to perform structural comparisons fail  
549 to superimpose these fragments, or yield extremely high RMSDs values, which alters the  
550 corresponding structural distances and dendograms.

551 It would thus be erroneous to use the term “phylogenetic tree” in the above cases, since the  
552 constructed dendograms reflect the polymerases’ phenotypic differences caused by the  
553 ligands binding, and not an evolutionary scenario in which the structural deviations are due  
554 to the accumulation of mutations over evolutionary times.

555 This work shows that, in spite of their limitations, tertiary structure-based dendograms can  
556 be a useful tool in the study of different evolutionary scenarios. Even for proteins with high  
557 levels of identity, as is the case of the polymerases selected for this work, the methodology  
558 allowed a correct taxonomical separation of the crystal structures in which no ligands were  
559 present. Even in those circumstances, it is important to make a careful selection of the  
560 structures, since inherent crystallographic factors such as the resolution of the structure may  
561 modify the expected dendogram topology.

562 Nevertheless, one of the current, most relevant limitations is the scant biological diversity of  
563 the biological entities from which the available structures originated. As shown in this work,

564 the abundance of a given crystallographic structure is frequently biased and driven by specific  
565 biomedical and economic interests, which can be very different from evolutionary or  
566 biological ones.

567 An aspect that partially explains the aforementioned limitation is that protein crystallization  
568 remains in many ways an artisanship, and it is impossible to predict the appropriate  
569 conditions in which a protein crystal structure will be obtained. Factors not pertaining to the  
570 protein, such as solvent concentration and chemical environment, room temperature, X-ray  
571 sources and wavelengths, can all have a direct impact on the electrochemical interactions  
572 within the macromolecule and the determination of its final tridimensional structure.  
573 Moreover, aspects such as the size of the protein of interest, the quaternary structure of a  
574 given complex, the presence of highly mobile regions within the molecule, the dependence  
575 of a ligand or a substrate for a protein to be crystallized, can determine other factors like the  
576 final resolution and the quality of the tertiary structure. Unlike the primary structure of a  
577 protein, which is fixed, a protein in solution is always in movement, and unpredictable results  
578 may occur such as unexpected arrangements or local conformational changes (*vide supra*).  
579 Finally, there is always room for certain levels of subjectivity when interpreting the electron  
580 density maps obtained in the diffraction experiment. In an effort to improve the quality of the  
581 structure during the refinement process, the risk of over-interpretation is always latent and  
582 atoms may be assigned to questionable regions within the electron density maps [51]. In most  
583 cases, these artifacts do not have a significant repercussion. However, if those atoms  
584 participate in important biological processes such as catalysis or ligand binding, the incorrect  
585 spatial assignment of a single atom, may have major biological implications [51]. As can be

586 seen, attention must be paid to every single detail when selecting the tertiary structures of  
587 interest in order to yield the most accurate evolutionary inferences possible.

588 The methodological and theoretical framework of tridimensional structure-based  
589 phylogenies needs to be further developed. For instance, as of today there are no statistical  
590 tools equivalent to the bootstrap or the jack-knife tests, that could add robustness to the  
591 proposed evolutionary hypothesis.

592 The role that other techniques like cryo-EM will play in the future remains to be seen. The  
593 advances of this imaging technology have allowed the obtention of a growing number of  
594 structures in the near-atomic range ( $<4 \text{ \AA}$ ) [57, 58], with resolutions near or even below  $2 \text{ \AA}$   
595 [59-61]. Moreover, even though most of the cryo-EM structures available comprise  
596 molecules with high symmetry such as viral capsids, in recent years, the structures of  
597 different enzymes have been resolved. Eventually, more and more structures obtained by  
598 different techniques will converge and new questions regarding tertiary structure-based  
599 phylogenies will arise.

600 We believe that, as of today, tertiary structure-based phylogenies remain underrated and sub-  
601 utilized, and it is clear that more bench work needs to be done to improve the current  
602 limitations and constraints of this methodology. In spite of these pitfalls, the construction of  
603 phylogenies based on tertiary structure comparisons has shown its potential and it is easy to  
604 see that in the near future it will be an invaluable asset in the reconstruction of evolutionary  
605 conundrums.

## 606 **Material and methods**

### 607 **Tertiary structures' selection**

608 A primary search for HCV NS5B and HIV-1 RT structures was performed on the PDB web  
609 site using the terms “Hepatitis C virus AND polymerase” and “Human immunodeficiency  
610 virus AND reverse transcriptase”. For each protein, a customized table was created and  
611 downloaded for further analysis. We performed a similar search in the European Protein Data  
612 Bank [27] as a complementary approach.

613 Once the HCV NS5B and HIV-1 RT structures were identified, we separated the structures  
614 belonging to the different taxonomical levels, i.e. genotypes and subtypes, when possible.

615 Equivalent structures among the HCV and HIV-1 subtypes were then located, to define a set  
616 of structures without bound ligands, structures with similar ligands, i.e. double-stranded  
617 nucleic acids, antivirals, and small molecules bound to similar allosteric sites.

618 In the case of HCV, due to the diversity in the chemical groups the ligands belong to, we first  
619 selected a series of different chemical groups: benzothiadiazines, anthranilic acid derivatives,  
620 quinazolinones, benzodiazepine polymerase inhibitors, benzofurans. Next, we launched a  
621 search using the “Similar Ligands” tool within the PDB [1] to broaden our sample.

622 Finally, the selection criteria included the integrity of the structure; therefore, we discarded  
623 those structures in which the intended mutations caused the loss of big fragments of the  
624 protein (for a broader explanation, see the results section).

625 The selected structures were used for the subsequent structural comparisons.

### 626 **Structural alignments and dendogram construction**

627 The pairwise structural alignments of the selected HCV NS5B and HIV-1 RT structures were  
628 performed in the web-based Secondary Structure Matching (SSM) program [62] using  
629 default parameters. For each of the pairwise comparisons we calculated the Structural  
630 Alignment Score ( $SAS = RMSD \times 100 / \text{number of aligned residues}$ ) [30], and we built a

631 structural distance-matrix. Then, for each of the two proteins, two different matrices were  
632 built, one for ligand-free structures and another one for structures with substrates and ligands.  
633 Distance-based dendograms were constructed for each of the structural distance matrices  
634 using the FITCH algorithm included within the PHYLIP package version 3.695. The  
635 dendograms were visualized and edited with FigTree.  
636 The visualization and rendering of the three-dimensional structures were performed using  
637 Chimera version 1.11 [33]. The MatchMaker and Match/Align programs included within  
638 Chimera were used to make the detailed analysis of the representative structural  
639 superpositions.

640

## 641 **Acknowledgments**

642 R. J gratefully acknowledges the financial support from the Dirección General de Asuntos  
643 del Personal Académico (DGAPA), UNAM. YLV, AB, JACB, and AL are thankful to  
644 Dirección General de Asuntos del Personal Académico, Universidad Nacional Autónoma  
645 de México (DGAPA- PAPIIT).

646

## 647 **References**

- 648 1. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The  
649 Protein Data Bank. *Nucleic Acids Res.* 2000; 28: 235-242.
- 650 2. Shi Y. A glimpse of structural biology through X-ray crystallography. *Cell.* 2014;  
651 159:995-1014.
- 652 3. Garman EF. Developments in X-ray crystallographic structure determination of  
653 biological macromolecules. *Science.* 2014; 343:1102-1108.



- 654 4. Anfinsen CB. Principles that govern the folding of protein chains. *Science*. 1973; 181:  
655 223–230.
- 656 5. Rost B. Twilight zone of protein sequence alignments. *Prot Eng*. 1999; 12: 85-94.
- 657 6. Clothia C, Lesk AM. The relation between the divergence of sequence and structure  
658 in proteins. *EMBO J*. 1986; 5:823-826.
- 659 7. Illergard K, Ardell DH, Elofsson A. Structure is three to ten times more conserved  
660 than sequence: a study of structural response in protein cores. *Proteins*. 2009; 77:499-  
661 508.
- 662 8. Gan HH, Perlow RA, Roy S, Ko J, Wu M, Huang J, et al. Analysis of protein  
663 sequence/structure similarity relationships. *Biophys J*. 2002; 83:2781-2791.
- 664 9. Jácome R, Becerra A, Ponce de León S, Lazcano A. Structural analysis of monomeric  
665 RNA-dependent polymerases: evolutionary and therapeutic implications *PLoS ONE*.  
666 2015; 10(9):e0139001. doi:10.1371/journal.pone.0139001
- 667 10. Agarwai G, Rajavel M, Gopal B, Srinivasan N. Structure-based phylogeny as a  
668 diagnostic for functional characterization of proteins with a cupin fold. *PLoS ONE*.  
669 2009; 4(5):e5736.
- 670 11. O'Donoghue P, Luthey-Schulten Z. On the evolution of structure in aminoacyl-tRNA  
671 synthetases. *Microbiol Mol Biol Rev*. 2003; 67:550-573.
- 672 12. O'Donoghue P, Sethi A, Woese CR, Luthey-Schulten Z. The evolutionary history of  
673 Cys-tRNACys formation. *Proc Nat Acad Sci U.S.A*. 2005; 102:19003-19008.
- 674 13. Garau G, Di Guilmi AM, Hall BG. Structure-based phylogeny of the metallo- $\beta$ -  
675 lactamases. *Antimicrob Agents Chemother*. 2005; 49:2777-2784.

- 676 14. Bahar MW, Graham SC, Stuart DI, Grimes JM. Insights into the evolution of a  
677 complex virus from the crystal structure of Vaccinia virus D13. *Structure*.  
678 2011;19:1011-1020
- 679 15. Lakshmi B, Mishra M, Srinivasan N, Archunan G. Structure-based phylogenetic  
680 analysis of the lipocalin family. *PLoS ONE*. 2015; 10: e0135507.  
681 Doi:10.1371/journal.pone.0135507.
- 682 16. GBD 2015 HIV collaborators. Estimates of global, regional, and national incidence,  
683 prevalence, and mortality of HIV, 1980–2015: the Global Burden of Disease Study  
684 2015. *Lancet HIV*. 2016; 3:e361-387.
- 685 17. Petruzzello A, Marigliano S, Loquercio G, Cozzolino A, Cacciapuoti C. Global  
686 epidemiology of hepatitis C virus infection: an up-date of the distribution and  
687 circulation of hepatitis C virus genotypes. *World J Gastroenterol*. 2016; 22:7824-  
688 7840.
- 689 18. Lingala S, Ghany MG. Natural history of hepatitis C. *Gastroenterol Clin N Am*. 2015;  
690 44: 717-734.
- 691 19. Pau AK, George JM. Antiretroviral therapy: current drugs. *Infect Dis Clin North Am*.  
692 2014; 28:371-402.
- 693 20. Carter W, Connelly S, Struble K. Reinventing HCV treatment: past and future  
694 perspectives. *J Clin Pharmacol*. 2016; Sep 22. doi: 10.1002/jcph.830.
- 695 21. Blankson JN, Siliciano JD, Siliciano RF. Finding a cure for human immunodeficiency  
696 virus-1 infection. *Infect Dis Clin North Am*. 2014; 28:633-650.
- 697 22. Duffy S, Shackelton LA, Holmes EC. Rates of evolutionary change in viruses:  
698 patterns and determinants. *Nat Rev*. 2008; 9:267-276

- 699 23. Sarrazin C. The importance of resistance to direct antiviral drugs in HCV infection in  
700 clinical practice. *J Hepatol.* 2016; 64:486-504.
- 701 24. Steitz TA. DNA- and RNA-dependent polymerases. *Curr Opin Struct Biol.* 1993;  
702 3:31-38.
- 703 25. Mayhoub AS. Hepatitis C RNA-dependent RNA polymerase inhibitors: a review of  
704 structure-activity and resistance relationships; different scaffolds and mutations.  
705 *Bioorg Med Chem.* 2012; 20:3150-3161.
- 706 26. Frey KM. Structure-enhanced methods in the development of non-nucleoside  
707 inhibitors targeting HIV reverse transcriptase variants. *Future Microbiol.* 2015;  
708 10:1767-1772.
- 709 27. Velankar S, van Ginkel G, Alhroub Y, Battle GM, Berrisford JM, Conroy MJ, et al.  
710 PDBe: improved accessibility of macromolecular structure data from PDB and  
711 EMBL. *Nucleic Acids Res.* 2015; 44(D1): D385-395.
- 712 28. Hang JQ, Yang Y, Harris SF, Leveque V, Whittington HJ, Rajyaguru S, et al. Slow  
713 binding inhibition and mechanism of resistance of non-nucleoside polymerase  
714 inhibitors of Hepatitis C virus. *J Biol Chem.* 2009; 284:15517-15529.
- 715 29. Sofia MJ, Chang W, Furman PA, Mosley RT, Ross BS. Nucleoside, nucleotide, and  
716 non-nucleoside inhibitors of hepatitis C virus NS5B RNA-dependent RNA-  
717 polymerase. *J Med Chem.* 2012; 55:2481-2531.
- 718 30. Subbiah S, Laurents DV, Levitt M. Structural similarity of DNA-binding domains of  
719 bacteriophage repressors and the globin core. *Current Biology.* 1993; 3:141-148
- 720 31. Bartenschlager R, Lohman V, Penin F. The molecular and structural basis of  
721 advanced antiviral therapy for hepatitis C virus infection. *Nat Rev Microbiol.* 2013;  
722 11: 482-496.

- 723 32. Yeung KS, Beno BR, Parcella K, Bender JA, Grant-Young KA, Nickel A, et al.  
724 Discovery of a Hepatitis C NS5B replicase palm site allosteric inhibitor (BMS-  
725 929075) advanced to phase 1 clinical studies. *J Med Chem.* 2017; 60: 4369-4385.
- 726 33. Petersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, et al.  
727 UCSF Chimera-a visualization system for exploratory research and analysis. *J*  
728 *Comput Chem.* 2004; 25:1605-1612.
- 729 34. Caillet-Saguy C, Simister PC, Bressanelli S. An objective assessment of  
730 conformational variability in complexes of Hepatitis C virus polymerase with non-  
731 nucleoside inhibitors. *J Mol Biol.* 2011; 414:370-384.
- 732 35. Nakamura A, Tamura N, Yasutake Y. Structure of the HIV-1 reverse transcriptase  
733 Q151M mutant: insights into the inhibitor resistance of HIV-1 reverse transcriptase  
734 and the structure of the nucleotide-binding pocket of Hepatitis B virus polymerase.  
735 *Acta Cryst.* 2015; F71: 1384-1390.
- 736 36. Nair R, Lui J, Soong TT, Acton TB, Everett JK, Kouranov A, et al. Structural  
737 genomics is the largest contributor of novel structural leverage. *J Struct Funct*  
738 *Genomics.* 2009; 10:181-191.
- 739 37. Smith DB, Bukh J, Kuiken C, Muerhoff AS, Rice CM, Stapleton JT, et al. Expanded  
740 classification of hepatitis C virus into 7 genotypes and 67 subtypes: updated criteria  
741 and genotype assignment web resource. *Hepatology.* 2014; 59:318-327.
- 742 38. Goossens N, Negro F. Is genotype 3 of the hepatitis C virus the new villain?  
743 *Hepatology.* 2014; 59: 2403-2412.
- 744 39. Jacobson IM, Gordon SC, Kowdley KV, Yoshida EM, Rodríguez-Torres M,  
745 Sulkowski MS, et al. Sofosbuvir for hepatitis C genotype 2 or 3 in patients without  
746 treatment options. *N Engl J Med.* 2013; 368:1867-1877

- 747 40. Lawitz E, Mangia A, Wyles D, Rodríguez-Torres M, Hassanein T, Gordon SC, et al.  
748 Sofosbuvir for previously untreated chronic Hepatitis C infection. *N Eng J Med.*  
749 2013; 368:1878-1887.
- 750 41. Di Maio VC, Cento V, Mirabelli C, Artese A, Costa G, Alcaro S, et al. Hepatitis C  
751 virus genetic variability and the presence of NS5B resistance-associated mutations as  
752 natural polymorphisms in selected genotypes could affect the response to NS5B  
753 inhibitors. *Antimicrob Agents Chemother.* 2014; 58:2781-2797.
- 754 42. Mugosa B, Cella E, Lai A, Lo Presti A, Blasi A, Vratnica Z, et al. Hepatitis C virus  
755 genotype 3A in a population of injecting drug users in Montenegro: Bayesian and  
756 evolutionary analysis. *Arch Virol.* 2017; doi:10.1007/s00705-017-3224-5
- 757 43. Messina JP, Humphreys I, Flaxman A, Brown A, Cooke GS, Pybus OG, et al. Global  
758 distribution and prevalence of Hepatitis C virus genotypes. *Hepatology.* 2015; 61:77-  
759 87.
- 760 44. Hemelaar J. The origin and diversity of the HIV-1 pandemic. *Trends Mol Med.* 2012;  
761 18 182-192.
- 762 45. Beloukas A, Psarris A, Giannelou P, Kostaki E, Hatzakis A, Paraskevis D. Molecular  
763 epidemiology of HIV-1 infection in Europe: an overview. *Infect Gen Evol.* 2016;  
764 46:180-189.
- 765 46. Kantor R. Impact of HIV-1 pol diversity on drug resistance and its clinical  
766 implications. *Curr Opin Infect Dis.* 2006; 19:594-606.
- 767 47. Hemelaar J. Implications of HIV-1 diversity for the HIV-1 pandemic. *J Infect.* 2013;  
768 66:391-400.

- 769 48. Balaji S, Srinivasan N. Comparison of sequence-based and structure-based  
770 phylogenetic trees of homologous proteins: inferences on protein evolution. *J Biosci.*  
771 2007; 32:83-96.
- 772 49. Cerny J, Cerna Bolfikova B, Valdes JJ, Grubhoffer L, Ruzek D. Evolution of tertiary  
773 structure of viral RNA dependent polymerases. *PLoS ONE.* 2014; 5:e96070.  
774 doi:10.1371/journal.pone.0096070 PMID: 24816789
- 775 50. Mönttinen HA, Ravantti JJ, Stuart DI, Poranen MM. Automated structural  
776 comparisons clarify the phylogeny of the right-hand-shaped polymerases. *Mol Biol*  
777 *Evol.* 2014; 31: 2741–2752 doi: 10.1093/molbev/msu219 PMID: 25063440
- 778 51. Wlodawer A, Minor W, Dauter Z, Jaskolski M. Protein crystallography for non-  
779 crystallographers, or how to get the best (but not more) from published  
780 macromolecular structures. *FEBS J.* 2008; 275:1-21.
- 781 52. Bergsten J. A review of long-branch attraction. *Cladistics.* 2005; 21:163-193.
- 782 53. Hu WS, Hughes SH. HIV-1 reverse transcription. *Cold Spring Harb Perspect Med.*  
783 2012; 2:a006882.
- 784 54. Das K, Arnold E. HIV-1 reverse transcriptase and antiviral drug resistance. Part 1.  
785 *Curr Opin Virol.* 2013; 3:111-118.
- 786 55. Singh K, Marchand B, Kirby KA, Michailidis E, Sarafianos SG. Structural aspects of  
787 drug resistance and inhibition of HIV-1 reverse transcriptase. *Viruses;* 2010; 2:606-  
788 638.
- 789 56. Lansdon EB, Brendza KM, Hung M, Wang R, Mukund S, Jin D, et al. Crystal  
790 structures of HIV-1 reverse transcriptase with Etravirine (TMC125) and Rilpivirine  
791 (TMC128): implications for drug design. *J Med Chem.* 2010; 53:4295-4299.

- 792 57. Binshtein E, Ohi MD. Cryo-Electron microscopy and the amazing race to atomic  
793 resolution. *Biochemistry*. 2015; 54: 3133-3141.
- 794 58. Vonck J, Mills DJ. Advances in high-resolution cryo-EM of oligomeric enzymes.  
795 *Curr Opin Struct Biol*. 2017; 46: 48-54.
- 796 59. Bartesaghi A, Merk A, Banerjee S, Matthies D, Wu X, Milne JS, et al. 2.2 Å  
797 resolution cryo-EM structure of beta-galactosidase in complex with a cell-permeant  
798 inhibitor. *Science*. 2015; 348: 1147-1151.
- 799 60. Banerjee S, Bartesaghi A, Merk A, Rao P, Bulfer SL, Yan Y, et al. 2.3 Å resolution  
800 cryo-EM structure of human p97 and mechanism of allosteric inhibition. *Science*.  
801 2016; 351: 871-875.
- 802 61. Merk A, Bartesaghi A, Banerjee S, Falconieri V, Rao P, Davis MI, et al. Breaking  
803 cryo-EM resolution barriers to facilitate drug discovery. *Cell*. 2016; 165: 1698-1707.
- 804 62. Krissinel E, Henrick K. Secondary-structure matching (SSM), a new tool for fast  
805 protein structure alignment in three dimensions. *Acta Crystallogr D Biol Crystallogr*.  
806 2004; 60:2256-68

807

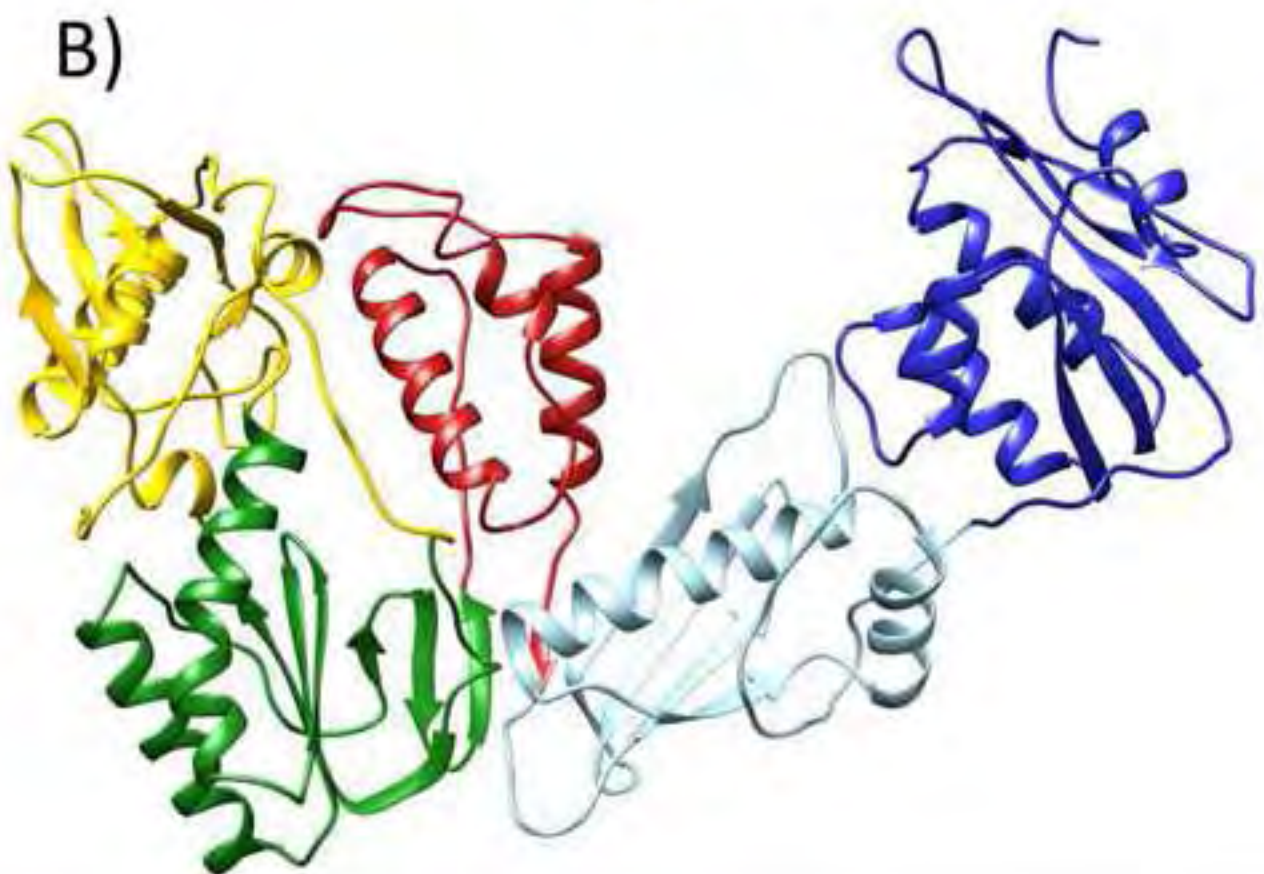
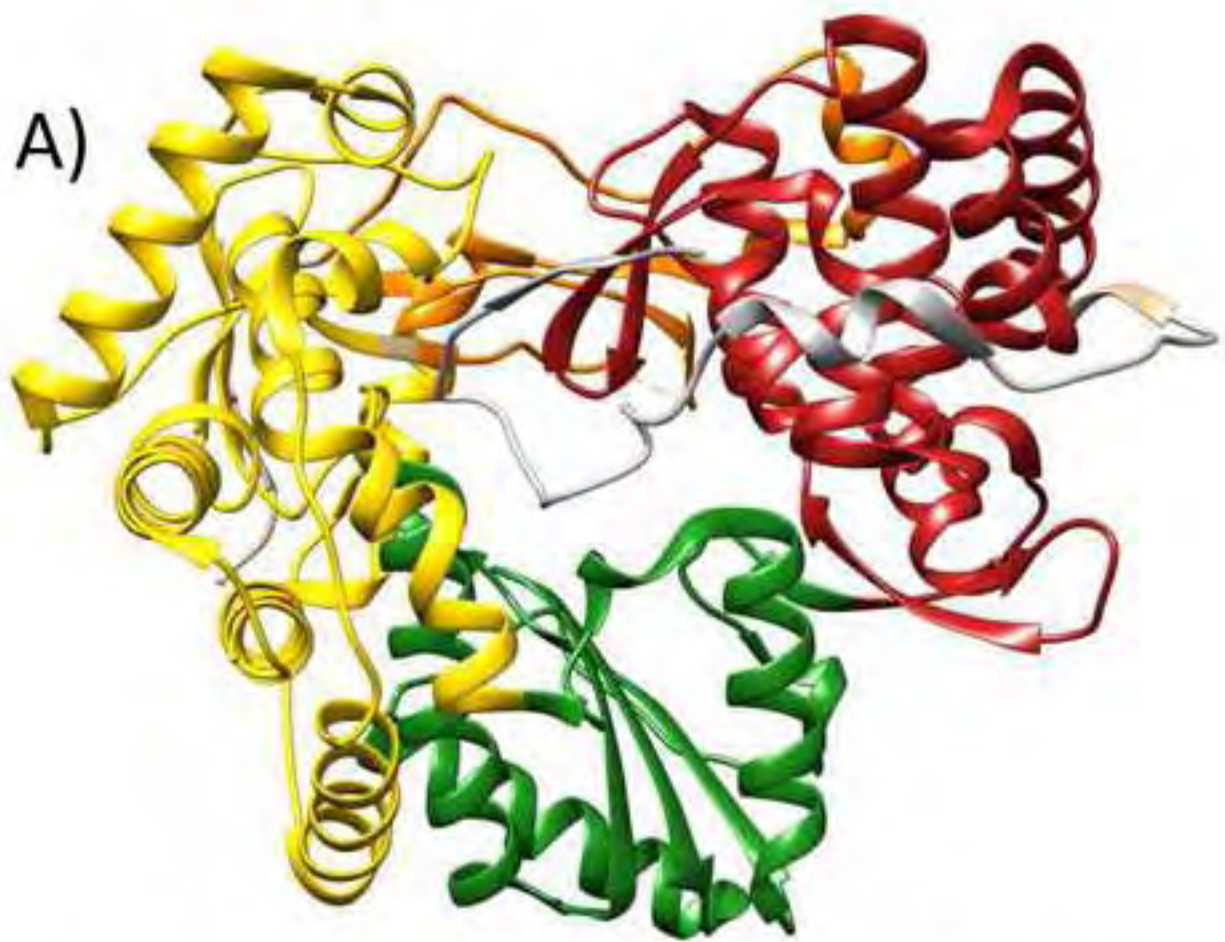
## 808 **Supporting information**

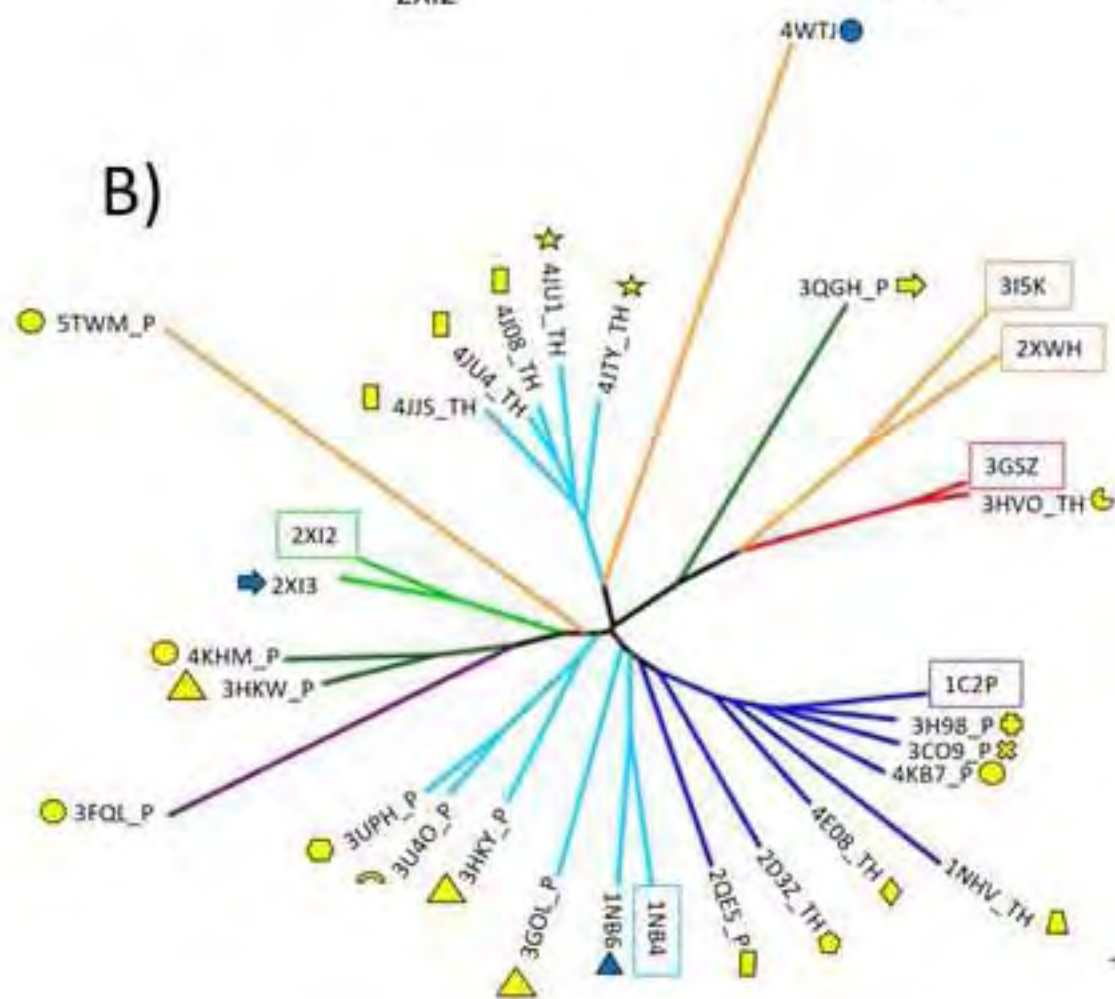
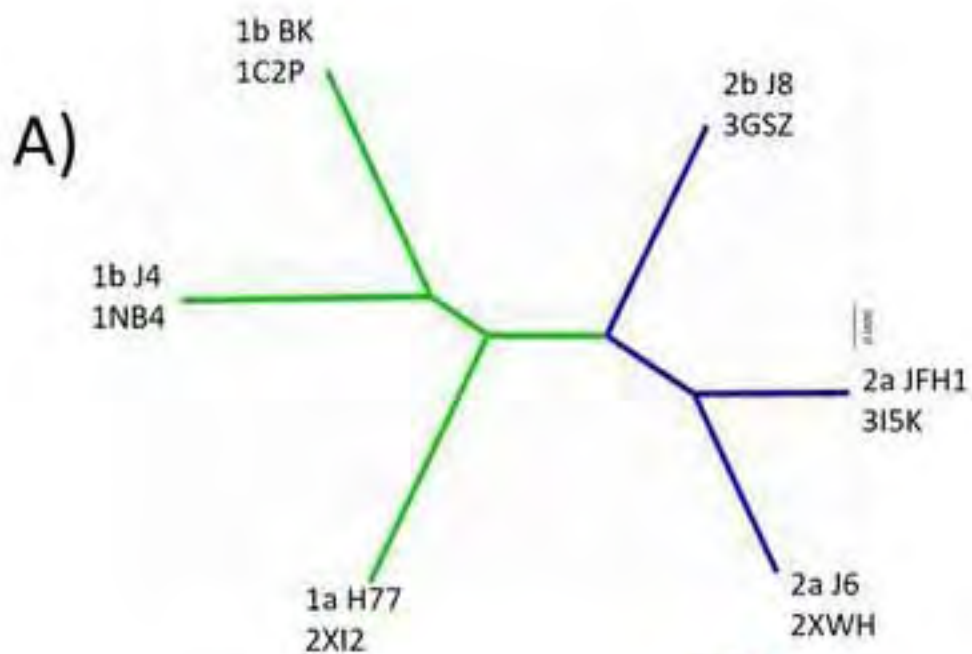
809 **S1 Table. HCV NS5B pairwise structural comparisons matrix.** The lower part of the  
810 matrix shows the calculated Structural Alignment Score for each structural comparison. The  
811 upper part of the matrix shows the RMSD (in Å) and the number of superimposed residues  
812 (in parenthesis) for each structural comparison. The PDB identifiers are shaded according to  
813 HCV subtype: blue – genotype 1 subtype BK; cyan – genotype 1b subtype J4; purple –  
814 genotype 1b subtype Con1; light green – genotype 1a subtype H77; dark green – genotype 1

815 subtype a; orange – genotype 2 subtype a; red – genotype 2 subtype b. For further details of  
816 the structures, please refer to Table 2.

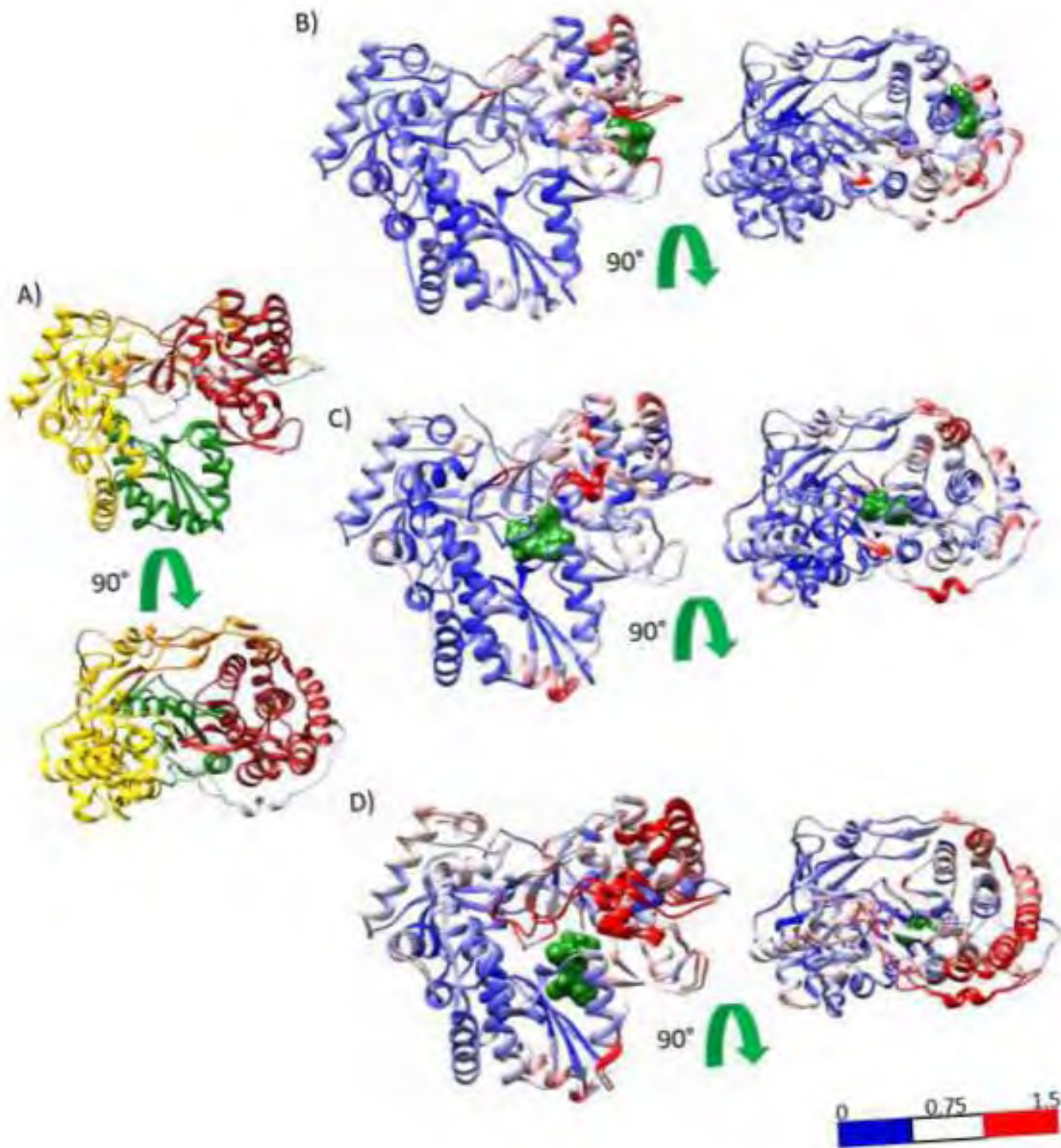
817 **S2 Table. HIV-1 RT pairwise structural comparisons matrix.** The lower part of the matrix  
818 shows the calculated Structural Alignment Score for each structural comparison. The upper  
819 part of the matrix shows the RMSD (in Å) and the number of aligned residues (in parenthesis)  
820 for each structural comparison. The PDB identifiers are shaded according to HIV subtype:  
821 blue – strain BH10; red – group M subtype B strain HXB2; yellow – NY5/BRU recombinant  
822 clone pNL4-3; green – strain Clone 12. The thick lines limit the matrix areas in which similar  
823 complexes are compared, i.e. ligand-free structures, structures with non-nucleoside RT  
824 inhibitors, and structures with double-stranded DNA, respectively. For further details of the  
825 structures, please refer to Table 3.

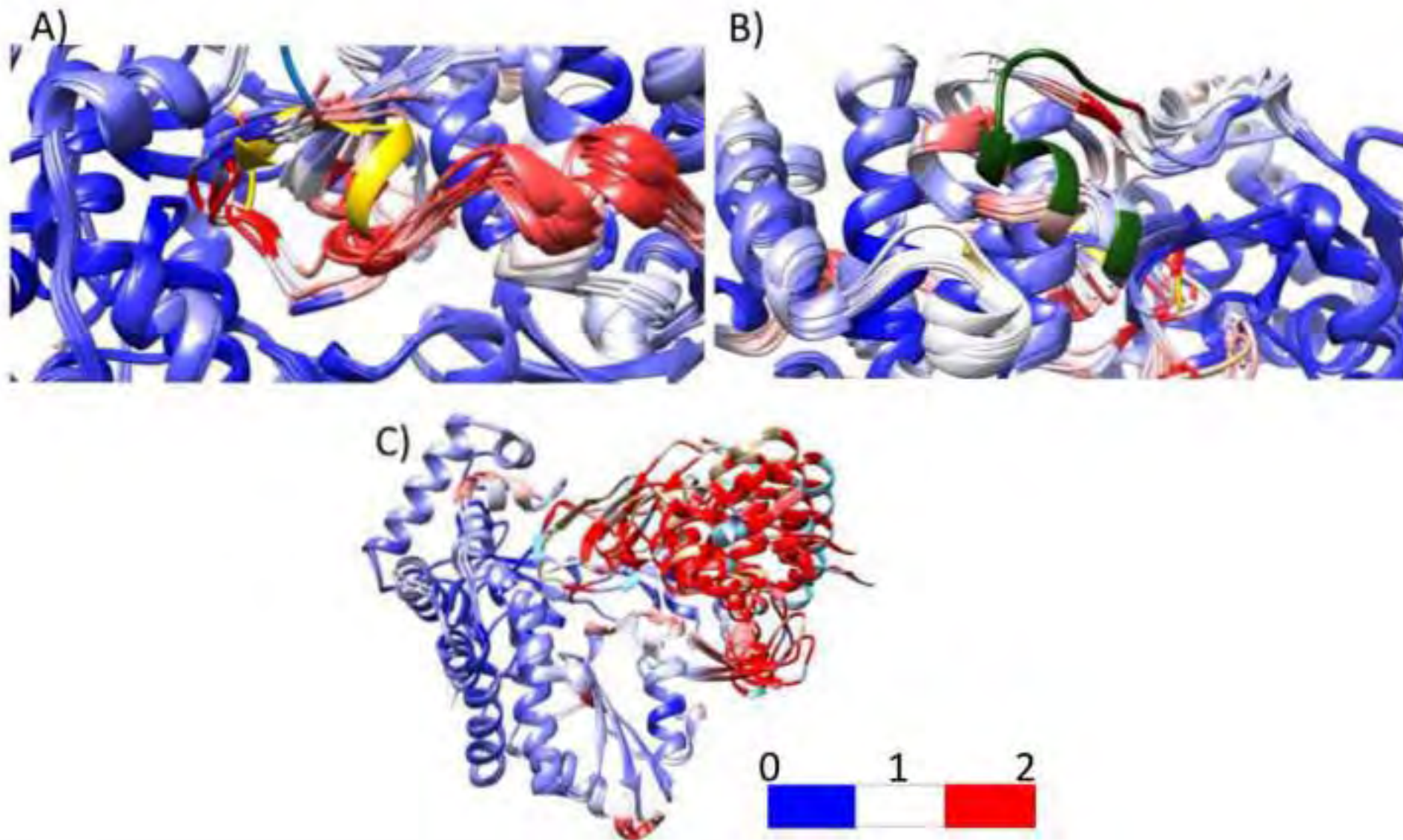




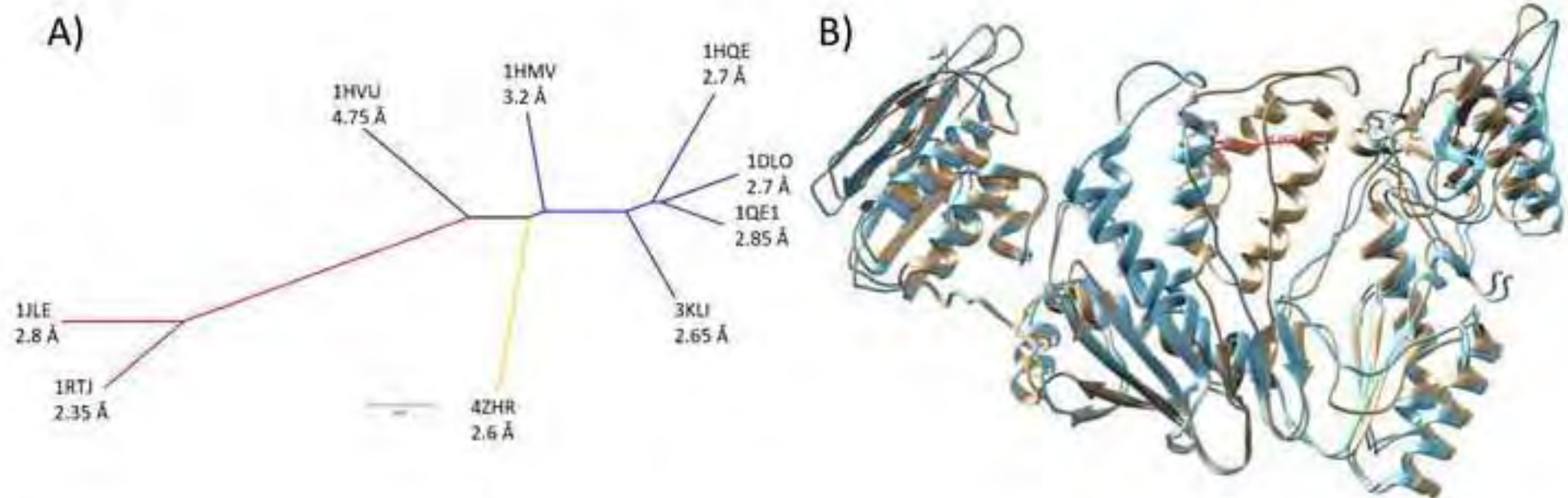


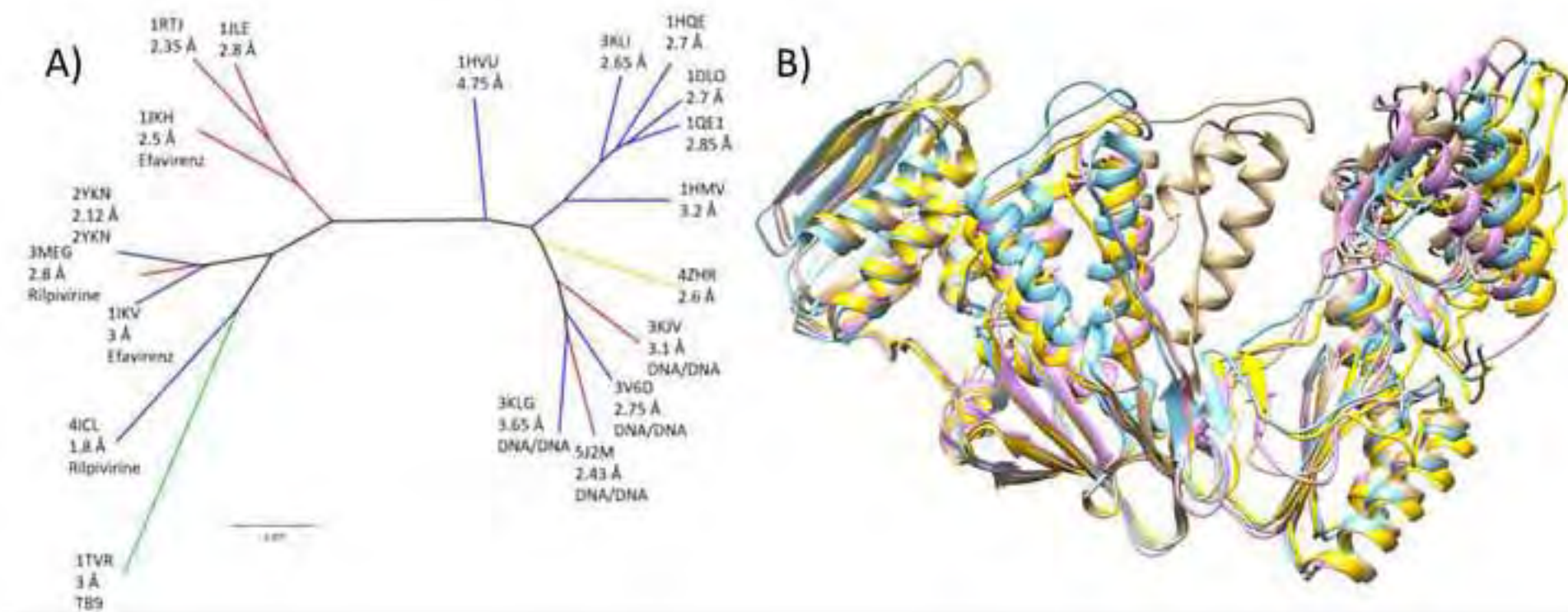














Click here to access/download  
**Supporting Information**  
S1\_Table.pdf





Click here to access/download  
**Supporting Information**  
S2\_Table.pdf