UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

POSGRADO EN CIENCIAS BIOLÓGICAS

FACULTAD DE CIENCIAS
BIOLOGÍA EVOLUTIVA

ORIGEN Y EVOLUCIÓN MOLECULAR DE LA DUPLICACIÓN DE GENES EN SISTEMAS DE RNA

TESIS

QUE PARA OPTAR POR EL GRADO DE:

MAESTRO EN CIENCIAS BIOLÓGICAS

PRESENTA:

ALEJANDRO MIGUEL CISNEROS MARTÍNEZ

TUTOR PRINCIPAL DE TESIS:
DR. ANTONIO EUSEBIO LAZCANO-ARAUJO REYES
FACULTAD DE CIENCIAS, UNAM

COMITÉ TUTOR:
DR. ARTURO CARLOS II BECERRA BRACHO
FACULTAD DE CIENCIAS, UNAM
DRA. MARIANA PEIMBERT TORRES
UAM CUAJIMALPA

CD. MX. **MAYO, 2019**

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
POSGRADO EN CIENCIAS BIOLÓGICAS

FACULTAD DE CIENCIAS
BIOLOGÍA EVOLUTIVA

ORIGEN Y EVOLUCIÓN MOLECULAR DE LA DUPLICACIÓN DE GENES EN SISTEMAS DE RNA

TESIS

QUE PARA OPTAR POR EL GRADO DE:

MAESTRO EN CIENCIAS BIOLÓGICAS

PRESENTA:

ALEJANDRO MIGUEL CISNEROS MARTÍNEZ

TUTOR PRINCIPAL DE TESIS:
DR. ANTONIO EUSEBIO LAZCANO-ARAUJO REYES
FACULTAD DE CIENCIAS, UNAM

COMITÉ TUTOR:
DR. ARTURO CARLOS II BECERRA BRACHO
FACULTAD DE CIENCIAS, UNAM
DRA. MARIANA PEIMBERT TORRES
UAM CUAJIMALPA

MÉXICO, CD. MX. **MAYO, 2019**

UNAM
POSGRADO
Ciencias Biológicas

COORDINACIÓN

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
POSGRADO EN CIENCIAS BIOLÓGICAS
FACULTAD DE CIENCIAS
DIVISIÓN ACADÉMICA DE INVESTIGACIÓN Y POSGRADO

OFICIO FCIE/DAIP/0343/2019

ASUNTO: **Oficio de Jurado**

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

Me permito informar a usted que en la reunión ordinaria del Comité Académico del Posgrado en Ciencias Biológicas, celebrada el día **18 de febrero de 2019** se aprobó el siguiente jurado para el examen de grado de **MAESTRO EN CIENCIAS BIOLÓGICAS** en el campo de conocimiento de **Biología Evolutiva** del alumno **CISNEROS MARTÍNEZ ALEJANDRO MIGUEL** con número de cuenta **310597564** con la tesis titulada **"Origen y evolución molecular de la duplicación de genes en sistemas de RNA"**, realizada bajo la dirección del **DR. ANTONIO EUSEBIO LAZCANO-ARAUJO REYES:**

Presidente:	**M. EN C. SUSANA LÓPEZ CHARRETÓN**
Vocal:	**DR. LUIS JOSÉ DELAYE ARREDONDO**
Secretario:	**DRA. MARIANA PEIMBERT TORRES**
Suplente:	**DR. SANTIAGO ÁVILA RÍOS**
Suplente:	**DR. LORENZO PATRICK SEGOVIA FORCELLA**

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

A T E N T A M E N T E
"POR MI RAZA HABLARA EL ESPÍRITU"
Ciudad Universitaria, Cd. Mx., a 22 de abril de 2019

DR. ADOLFO GERARDO NAVARRO SIGÜENZA
COORDINADOR DEL PROGRAMA

POSGRADO EN CIENCIAS BIOLOGICAS
COORDINACIÓN

AGNS/VMVA/ASR/grf*

Agradecimientos institucionales

Agradezco al Posgrado en Ciencias Biológicas y a la Universidad Nacional Autónoma de México por brindarme la oportunidad de formarme como un profesional de las ciencias biológicas en un programa de la más alta calidad.

Agradezco también al apoyo PAPIIT-UNAM (IN223916) y a CONACYT (CVU:814975/ Becario:620137) sin cuyos apoyos no habría sido posible realizar esta tesis.

Finalmente, quiero agradecer a mi tutor principal, el Dr. Antonio Eusebio Lazcano-Araujo Reyes y a los miembros de mi comité tutor, el Dr. Arturo Carlos II Becerra Bracho y a la Dra. Mariana Peimbert Torres que me guiaron en el desarrollo de esta tesis.

Agradecimientos a título personal

Agradezco a mis tutores el Dr. Antonio Lazcano, el Dr. Arturo Becerra y la Dra. Mariana Peimbert, así como a los miembros del jurado la Dra. Susana López, el Dr. Luis Delaye, el Dr. Lorenzo Segovia y el Dr. Santiago Ávila por su apoyo en observaciones, críticas, correcciones, consejos y trámites que permitieron el desarrollo de esta tesis.

Agradezco a los miembros del Laboratorio de Origen de la Vida, Ricardo Hernández, José Campillo, Rodrigo Jácome, Alberto Vázquez, Israel Muñoz, Coral Cruz, Alejandro Álvarez, Germán Alonso, Carolina Rocha, Wolfgang Cottom y Karen Mendoza, por su compañerismo y amabilidad, así como su disposición para escuchar los seminarios sobre los avances de mi proyecto, sobre el cual fueron muy valiosas sus preguntas y observaciones.

Agradezco a mis amigos de la carrera Carlos González, Armando Rodríguez, Iván Linares, Alejandro Gaona, Alondra Vega, Cristina Ramírez, Valeria Falcón, Yaravi Ramírez y Maricela Dircio.

Agradezco también a mis amigos de antaño como Elías Mochán y David Carnero que son casi como hermanos y sé que siempre estarán allí para reunirnos y divertirnos.

Finalmente, agradezco a mi familia, en especial a mis padres Miguel Cisneros y Luz María Martínez, así como a mi hermano Héctor Cisneros que son quienes más esperan de mi, que me aconsejan cuando lo necesito y que siempre me apoyan con amor.

A mis padres y a mi hermano

Tabla de contenido

Resumen

Una parte de la complejidad genómica del último ancestro común de todos los seres vivos puede atribuirse a la expansión de familias de proteínas por duplicación de genes. El inicio de este proceso de expansión podría remontarse hasta el mundo de RNA/proteínas, un etapa de la evolución temprana de la vida posterior al mundo de RNA pero previa al surgimiento de genomas de DNA. Sin embargo, la naturaleza hipotética de esta etapa temprana de la evolución imposibilita el estudio directo de sus propiedades genómicas. Una manera indirecta de estudiar la duplicación de genes en el mundo de RNA/proteínas es usar a los virus de RNA como modelo. La duplicación de genes se ha reportado como un evento poco frecuente en estos virus. Una posibilidad es que esto se deba a la poca conservación de sus secuencias proteínicas a causa de una elevada tasa de mutación. En esta tesis se utiliza la información de la estructura terciaria para estudiar la evolución de tres proteínas que podrían haberse originado por dos eventos de duplicación: las proteínas de la cápside VP1, VP2 y VP3 de virus del orden *Picornavirales*. El análisis sugiere que del primer evento de duplicación surgieron VP2 y una proteína ancestral que, tras un segundo evento de duplicación, dio origen a VP1 y VP3. Esta historia viene acompañada por la transferencia horizontal de dominios P y convergencias conformacionales de los extremos N-terminal de VP2. Este caso de estudio revela que la duplicación de genes en genomas de RNA puede tener implicaciones sobre el rápido ensamblaje de complejos moleculares, la especialización funcional de proteínas y el surgimiento de fuentes de constante adaptación en ambientes cambiantes.

Abstract

A part of the genomic complexity of the last common ancestor of all living beings can be attributed to the expansion of protein families by gene duplication. The beginning of this expansion process could go back to the RNA/protein world, a stage in the early evolution of life after the RNA world but before the emergence of DNA genomes. However, the hypothetical nature of this early stage of evolution precludes the study of its genomic properties. An indirect way to study the duplication of genes in the RNA/protein world is to use RNA viruses as a model. Gene duplication has been reported as an unfrequent event in these viruses. One possibility is that this is due to the low conservation of their protein sequences because of a high mutation rate. In this thesis the tertiary structure information is used to study the evolution of three proteins that could have originated by two duplication events: the capsid proteins VP1, VP2 and VP3 of viruses of the order *Picornavirales*. The analysis suggests that the first duplication event gave rise to VP2 and an ancestral protein which after a second duplication event led to the emergence of VP1 and VP3. This history is accompanied by the horizontal transfer of P domains and conformational convergences of the N-terminal ends of VP2. This case study reveals that gene duplications in RNA genomes may have implications for the rapid assembly of molecular complexes, the functional specialization of proteins and the emergence of sources of constant adaptation to changing environments.

1 Introducción

1.1 El mundo del RNA

Desde 1936 Alexander I. Oparin estableció un esquema robusto sobre el origen heterotrófico de la vida a partir de la síntesis y acumulación de compuestos orgánicos en una atmósfera reductora. Sin embargo, el proceso de transición de la materia orgánica acumulada en la tierra primitiva a la formación de los primeros seres vivos sigue siendo ampliamente debatido (Lazcano, 2014). No obstante, desde 1954 John B. S. Haldane propuso que "el evento crítico al cual mejor nos podríamos referir como el origen de la vida fue el empaquetamiento de varios polímeros auto reproducibles diferentes dentro de una membrana semipermeable" (Haldane, 1954). Como lo describe Lazcano (2012), esta visión sería parte de un conjunto de observaciones independientes que acabarían por sentar las bases teóricas de la hipótesis del mundo del RNA. Una de las cosas que tenían en común estas observaciones, algunas de las cuales fueron expuestas en el primer simposio internacional sobre el origen de la vida, convocado por Oparin en Moscú en 1957, es que tendían a implicar que el RNA es una molécula más antigua que el DNA y las proteínas. Por un lado, la cristalización del virus del mosaico del tabaco (TMV), que se interpretó como el hallazgo de un eslabón entre el mundo mineral y el mundo de lo vivo (Stanley, 1935), y la demostración de que el TMV almacena su información hereditaria en RNA (Fraenkel-Conrat et al. 1957), provocaron que algunos autores asumieran que los virus son antiguos y por lo tanto el RNA también (Stanley, 1959; Fraenkel-Conrat & Singer, 1959). Por otro lado, Brachet (1959), que destacó la simplicidad del RNA comparado con las proteínas,

especuló que el RNA se pudo haber formado espontáneamente por polimerización no enzimática, y Belozerskii (1959), tras discutir sobre la versatilidad del RNA en diferentes procesos celulares, concluyó que la gran especialización y diferenciación del DNA sugería que este se originó después que el RNA. Al inicio de la década de los 60s, Handler (1963) y Eakin (1963), de manera independiente, notaron que muchas coenzimas son nucleótidos o derivados de nucleótidos y propusieron que estas son vestigios de una etapa, previa a la aparición de las proteínas, en la cual el metabolismo era mediado por coenzimas. Sin embargo, los primeros en sugerir de manera explícita que los primeros seres vivos carecían de DNA y de proteínas, y que sus funciones eran desempeñadas por el RNA fueron Alexander Rich (1962), Carl E. Woese (1967), Francis Crick (1968) y Leslie E. Orgel (1968). En primer lugar, Rich reconoció que a) la síntesis prebiótica de aminoácidos y su posible polimerización para formar proteínas primitivas no explicaba el surgimiento de la síntesis de proteínas basada en ácidos nucleicos, b) que hay múltiples razones estereoquímicas para creer que los polinucleótidos pueden catalizar su propia replicación pero no así en el caso de los poliaminoácidos y c) que la mayor estabilidad del DNA refleja un proceso de especialización, a través de evolución Darwiniana, a partir de un ácido nucleico más versátil (Rich, 1962). En segundo lugar, la ideas generales detrás de las hipótesis desarrolladas de manera independiente por Woese, Crick y Orgel implican que los ribosomas originales podían haber estado compuestos solo de RNA y que el RNA podría actuar como molde y como enzima capaz de catalizar su propia replicación (Woese, 1962; Crick, 1968; Orgel, 1968). En el caso de Crick, su perspicacia y audacia incluso lo llevó a afirmar que “la primer enzima era una molécula de RNA con propiedades de replicasa” (Crick, 1968). En la década de los 70s, Hartman (1975) y White III (1976), de manera independiente, hicieron observaciones

similares a las de Handler y Eakin, sin embargo, ninguna de las propuestas relacionadas con las propiedades catalíticas de los ácidos nucleicos o su relación con las coenzimas y sus implicaciones sobre la evolución temprana de la vida fueron tomadas con suficiente seriedad y fueron consideradas como plenamente especulativas hasta que en la década de los 80s Thomas Cech (Kruger et al. 1982) y Sidney Altman (Guerrier-Takada et al. 1983), de manera independiente, descubrieron moléculas de RNA con actividad catalítica, a las cuales se les conoce desde entonces como ribozimas. Poco tiempo después, Gilbert (1986) acuñó el término de "el mundo del RNA", expresión que permitió la difusión y popularidad de esta hipótesis que plantea que en etapas tempranas de la evolución de la vida el RNA funcionaba como molécula catalítica y como material portador de la información hereditaria, en un escenario ausente de proteínas y DNA. La variedad de propiedades catalíticas de las ribozimas y los derivados de ribonucleótidos, así como las propiedades genéticas del RNA, son el principal sustento de esta hipótesis y le otorgan al RNA un papel protagónico durante el origen y evolución temprana de la vida (Lazcano, 2014).

El escenario propuesto por Gilbert (1986) se puede dividir en tres fases: I) En el mundo del RNA, las moléculas de RNA se habrían ensamblado a partir de una sopa de nucleótidos. Entonces, las moléculas de RNA se habrían auto replicado y evolucionado a través de mutaciones y recombinaciones, adquiriendo nuevas funciones y adaptándose a nuevos nichos. Con ayuda de cofactores como la nicotinamida adenina dinucleótido, las ribozimas habrían adquirido una gran variedad de funciones catalíticas. II) En el mundo de RNA/proteínas, las ribozimas habrían comenzado a sintetizar proteínas, a través de moléculas de RNA capaces de

unirse a aminoácidos activados, ordenando a los aminoácidos a través de una molécula de RNA molde y catalizando la polimerización de los aminoácidos con moléculas de RNA como las del sitio catalítico del ribosoma. Eventualmente, las enzimas habrían adquirido el rol principal en los procesos catalíticos debido a una mayor efectividad y rapidez para realizar las reacciones. III) Finalmente, en el mundo de DNA/RNA/proteínas habría aparecido el DNA, copiado a partir del RNA por transcripción reversa, como un portador confiable de la información hereditaria. A partir de la evolución del DNA de cadena doble habría surgido la posibilidad de almacenar la información de manera estable y de corregir los errores de la replicación. Así, el DNA habría acabado por relegar al RNA a su rol de intermediario entre el DNA y las proteínas.

1.2 Duplicación de genes durante la evolución temprana de la vida

Los catálogos del contenido de genes del último ancestro común (LCA o cenancestro) sugieren que este ya era un organismo complejo, similar a las bacterias contemporáneas, capaz de realizar varios procesos biológicos esenciales. Esta complejidad puede ser atribuida a duplicaciones, transferencias horizontales (HGT) y a la pérdida de genes durante la evolución temprana de la vida (Becerra et al., 2007). No obstante, el repertorio funcional y el tamaño de los genomas tempranos no solo se pudieron haber incrementado a través del aprovechamiento del material genético preexistente (e.g. HGT, duplicación y fusión de genes (Kaessmann, 2010)), sino que también a través de mecanismos capaces de originar genes *de novo* (e.g. mutaciones en secuencias no codificantes (Kaessmann, 2010), mutaciones en el cuerpo de un gen que originan marcos de lectura sobrelapados (Delaye et al., 2008) o errores en la replicación, debidos al

deslizamiento de la polimerasa, que generan regiones de baja complejidad (Velasco et al., 2013)). Sin embargo, desde que Susumu Ohno, en su libro *Evolution by Gene Duplication* (1970), destacó el papel de la duplicación de genes en la evolución temprana de los vertebrados, este mecanismo ha sido considerado como una de las principales fuerzas de la evolución. La lógica de Ohno parte de la observación de que la selección natural es muy poco tolerante con mutaciones que alteran la función de genes codificantes. Si no pueden ocurrir grandes cambios evolutivos por acumulación de mutaciones, entonces debe haber otro mecanismo que permita la adquisición de nuevos genes con funciones previamente inexistentes. Teniendo esto en consideración, argumenta que la duplicación de genes, a través de la generación de copias redundantes, permite que una de las copias sea ignorada por la implacable presión de la selección natural y por lo tanto acumule mutaciones que deriven en una nueva función. Las hipótesis de Ohno no solo se han confirmado sino que también han impulsado el desarrollado diversos modelos que explican a la evolución por duplicación de genes.

La importancia evolutiva de este fenómeno se puede visualizar fácilmente por el hecho de que ha sido registrado en altas proporciones en los tres dominios de la vida (17-44% en Bacteria, ~30% en Archaea y 30-65% en Eukarya) (Zhang, 2003), por lo cual es considerado como el mecanismo más importante para el incremento en tamaño y complejidad de los genomas (Lazcano, 1995). Como lo demuestra la fuerte correlación que existe entre el número de genes homólogos surgidos por duplicación (parálogos) y el tamaño del genoma (Gevers et al., 2004), la expansión de las familias de proteínas por duplicación de genes es un fenómeno que se encuentra íntimamente relacionado con el crecimiento de los genomas, y podría explicar en gran medida la

evolución de los genomas desde que eran compuestos por unos cuantos genes primordiales hasta su composición por varios miles de genes (Lazcano, 1995; Magadum et al., 2013). Por otro lado, se ha observado una relación inversamente proporcional entre el tamaño del genoma y la tasa de mutación, de manera que las entidades biológicas con mayor tasa de mutación presentan los genomas más pequeños. Estos son los viroides, seguidos por los virus de RNA, los virus de DNA de cadena sencilla, virus de DNA de cadena doble, bacterias y, finalmente, eucariontes (Gago et al., 2009). En este sentido, se ha sugerido que los primeros genomas eran pequeños y susceptibles a altas tasas de mutación y que la tendencia evolutiva ha sido, gracias al origen de las cadenas dobles de DNA, la reducción de la tasa de error debido al incremento en la fidelidad de las polimerasas (Holmes, 2011). Esto habría permitido la evolución de genomas más grandes que, por duplicación de genes, habrían adquirido un mayor potencial codificante (Lazcano et al., 1992).

El origen de los primeros genes podría remontarse hasta el mundo del RNA (Kaessmann, 2010), y la duplicación de genes también. Por ejemplo, el sitio de formación del enlace peptídico en el ribosoma, llamado centro peptidil transferasa (PTC), es un elemento estructural universal de los ribosomas que podría ser un remanente de una ribozima presente en el mundo del RNA. Esta ribozima habría dado lugar a una maquinaria antigua, conocida como el “proto-ribosoma”, capaz de formar enlaces peptídicos y de polimerizar péptidos no codificados. El PTC, dispuesto de una manera casi simétrica, está formado por dos regiones, A y P, estructuralmente muy similares que consisten de un motivo tallo-codo-tallo similar al de los tRNA. Estas dos regiones, y por lo tanto el proto-ribosoma, podrían ser resultado de un evento de duplicación (Krupkin et

al., 2011). Sin embargo, establecer el surgimiento de la duplicación de genes es un problema que tiene que ver con el concepto de gen. Por ejemplo, durante el mundo del RNA, el genotipo y el fenotipo estaban representados por la misma molécula, una relación entre información y función que se puede ver ejemplificada en el caso de los viroides, que son sistemas de RNA con genomas que no codifican para proteínas, pero sí, en algunos casos, para estructuras terciarias con actividad catalítica (Flores et al., 2012). En otras palabras, la distinción entre genotipo y fenotipo no era tan clara como hoy en día en donde la información es almacenada de manera lineal en el DNA y la función es determinada por las estructuras tridimensionales de las proteínas. Por lo tanto, si convenimos en definir a los genes como marcos de lectura abiertos, entonces tendríamos que reconocer que la duplicación de genes habría comenzado a lo sumo en el mundo de RNA/ proteínas, una etapa representada por el surgimiento de proteínas codificadas, por el desplazamiento funcional de las ribozimas por enzimas, dejando a las coenzimas como vestigios de una etapa anterior, y por la aparición y expansión de las primeras familias de proteínas (White III, 1976). El número de familias de proteínas conocidas, que pueden componerse desde uno hasta 2,000 miembros, incrementa con cada genoma secuenciado, y el número real podría ser mucho mayor a 60,000 (Kunin et al., 2003). Muchas de estas familias de proteínas se podrían haber expandido en diferentes etapas evolutivas y dentro de diversos grupos taxonómicos lidiando con una amplia variedad de ambientes (Kunin et al., 2003; Gevers et al., 2004). Sin embargo, es posible que varias de estas familias hayan surgido a partir de la amplificación de genes ancestrales que codificaban para proteínas originadas previo al surgimiento de genomas de DNA (Lazcano et al., 1992).

Se han identificado varias familias de proteínas bien conservadas y ampliamente distribuidas cuya expansión parece haber ocurrido antes de la divergencia de los tres dominios de la vida (Becerra et al., 2007). Algunos ejemplos incluyen a i) grandes familias como la de los transportadores ABC (Clayton et al., 1997), ii) familias compuestas por pocos parálogos como los factores de elongación EF-G y EF-Tu (Iwabe et al., 1989) o las subunidades hidrofílicas alfa y beta de la ATPasa tipo-F (Gogarten et al., 1989), y iii) proteínas formadas por un par de módulos homólogos en tándem que sugieren un evento de duplicación seguido de una fusión, tal como en el caso de HisA, una isomerasa involucrada en la biosíntesis de histidina (Alifano et al., 1996). El análisis de estas familias de proteínas universalmente conservadas nos puede dar indicios sobre la organización y complejidad genómica de la población a partir de la cual divergió el LCA. Por ejemplo, aunque se desconocen organismos que presenten un solo factor de elongación o una sola subunidad hidrofílica de la ATPasa tipo-F, podemos inferir que el LCA desciende de células menos complejas, que tenían genomas más cortos y que codificaban para procesos más simples (Becerra et al., 2007). Si se toma en consideración que el origen monofilético de las ribonucleótido reductasas y la conservación de dominios que interactúan directamente con el RNA o que intervienen en la biosíntesis de RNA y nucleótidos sugieren, respectivamente, que el cenancestro tenía un genoma de DNA y que este es un resultado evolutivo del mundo de RNA/proteínas (Becerra et al., 2007), se puede especular que algunas de las familias universalmente conservadas se expandieron en poblaciones de células que codificaban su proteínas en genomas de RNA (Figura 1).

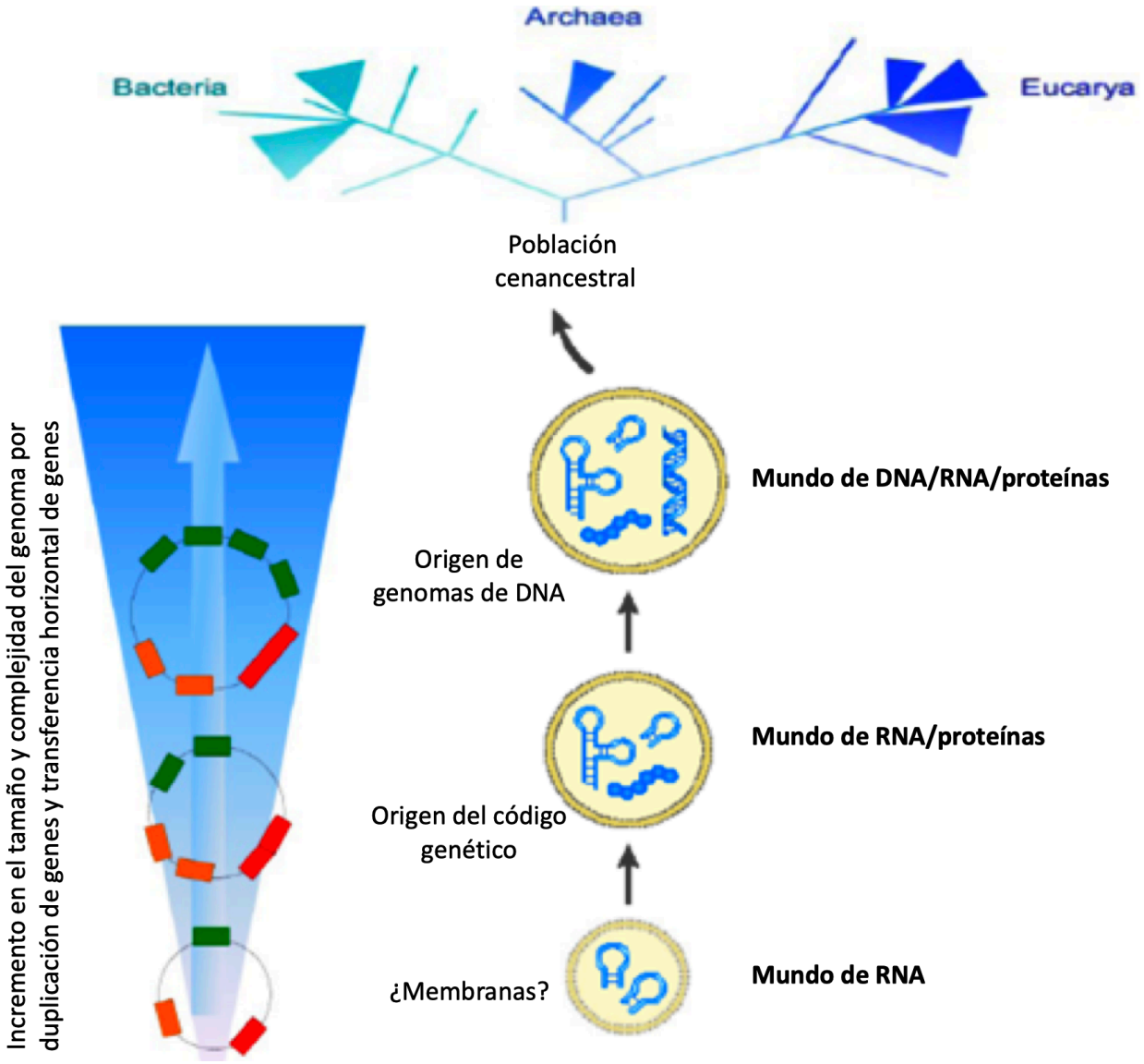


Figura 1. Eventos evolutivos previos al LCA y a la divergencia de los tres dominios de la vida. Los genes que habrían originado a grandes familias de genes, a familias compuestas por pocos miembros, y a proteínas compuestas por dos módulos homólogos, están representados por rectángulos verdes, anaranjados y rojos, respectivamente. Modificado de Becerra et al., 2007.

1.3 Virus de RNA como modelo del mundo de RNA/proteínas

Para obtener más información sobre la estabilidad de las duplicaciones de genes en genomas de RNA se pueden usar a los virus de RNA como modelo del mundo de RNA/ proteínas. Los virus de RNA son las únicas entidades biológicas que siguen codificando sus proteínas en genomas de RNA, y aunque su origen parece ser más bien reciente (Campillo-Balderas et al., 2015), podrían proporcionarnos indicios valiosos sobre las propiedades de los genomas previo a la aparición de genomas de DNA (Reyes-Prieto et al., 2012). Por ejemplo, es posible que la antes mencionada relación inversamente proporcional entre la tasa de mutación y el tamaño de los genomas (que muestra a los virus de RNA con mayores tasas de mutación y genomas más pequeños que los virus de DNA de doble cadena) refleje algunas características de la evolución temprana de los genomas (Gago et al., 2009; Holmes, 2011). Para esto, vale la pena señalar algunas de las posibles similitudes y diferencias entre los virus de RNA y las hipotéticas células del mundo de RNA/proteínas. Para empezar, una de las diferencias más fundamentales tiene que ver con el hábito estrictamente parasitario de los virus de RNA, el cual tiende a estar asociado con la reducción en el tamaño de los genomas. Consecuentemente, los virus de RNA dependen fuertemente de las maquinarias celulares, principalmente la maquinaria de traducción, para completar sus ciclos de replicación viral. En principio, se podría pensar que la mayoría de las células del mundo de RNA/proteínas habrían sido de vida libre, y que estas ya habrían tenido cierta autonomía metabólica y genética. Por otro lado, es posible que haya una relación entre el tamaño del virión y el tamaño del genoma (Cui et al., 2014). Si, de acuerdo con VirialZone (Hulo et al., 2011), consideramos que las cápsides más grandes de virus de RNA alcanzan hasta 300 nm

de diámetro, y lo comparamos con el diámetro de las células más pequeñas conocidas, que corresponde a las bacterias parásitas de la especie *Mycoplasma genitalium*, que puede ir desde 300 hasta 400 nm (Tully et al., 1981), podemos pensar, aunque no haya manera segura de saberlo, que los genomas de las células de vida libre del mundo de RNA/proteínas no habrían estado tan limitados por el tamaño de la célula. Independientemente de estas diferencias, es posible que, si la capacidad de corrección de las polimerasas evolucionó en genomas de DNA (García-Meza et al., 1994), las células del mundo de RNA/proteínas habrían tenido tasas de mutación similares a las de los virus de RNA. Además, se ha sugerido que durante el mundo de RNA/proteínas se pudieron haber superado las limitaciones debidas a las propiedades de los genomas a través de la evolución de genomas segmentados (Reyes-Prieto et al., 2012), de manera similar a como se observa en algunos virus de RNA, cuyos genomas segmentados tienden a ser más grandes que los de otros virus (Holmes, 2009).

1.4 Los virus y la duplicación de genes

La clasificación de Baltimore agrupa a los virus, tanto de DNA como de RNA, en siete grupos diferentes de acuerdo con la composición y mecanismos de procesamiento del genoma para la replicación del virus (Holmes, 2009) (Figura 2). Los primeros dos grupos incluyen a virus de DNA de cadena doble (grupo I: dsDNA) y DNA de cadena sencilla (grupo II: ssDNA). Los genomas de virus de RNA pueden ser de cadena doble (grupo III: dsRNA), cadena sencilla positiva (grupo IV: ssRNA(+)) o negativa (grupo V: ssRNA(-)), o de cadena sencilla retrotranscrita (grupo VI: ssRNA(RT)). El último grupo corresponde a virus de DNA de doble cadena retro transcrita (grupo VII: dsDNA(RT)). La cadena positiva corresponde a aquella que funciona directamente como mRNA, mientras que la cadena negativa tiene que transcribirse a positiva para poder ser traducida. En el caso de los retrovirus, el RNA de cadena sencilla es retrotranscrito por la transcriptasa reversa para dar lugar a un genoma de DNA de cadena doble que se integra en el genoma del hospedero para ser transcrito por la maquinaria celular (Holmes, 2009). En el caso de los virus de RNA de doble cadena, a partir del duplex paterno, se sintetiza una sola cadena positiva que sirve como molde para su propia cadena negativa complementaria, de modo que ninguna cadena paterna pasa a la siguiente generación, siendo un ejemplo de replicación conservativa basada en cadena sencilla a pesar de poseer genomas de cadena doble (Reanney, 1982).

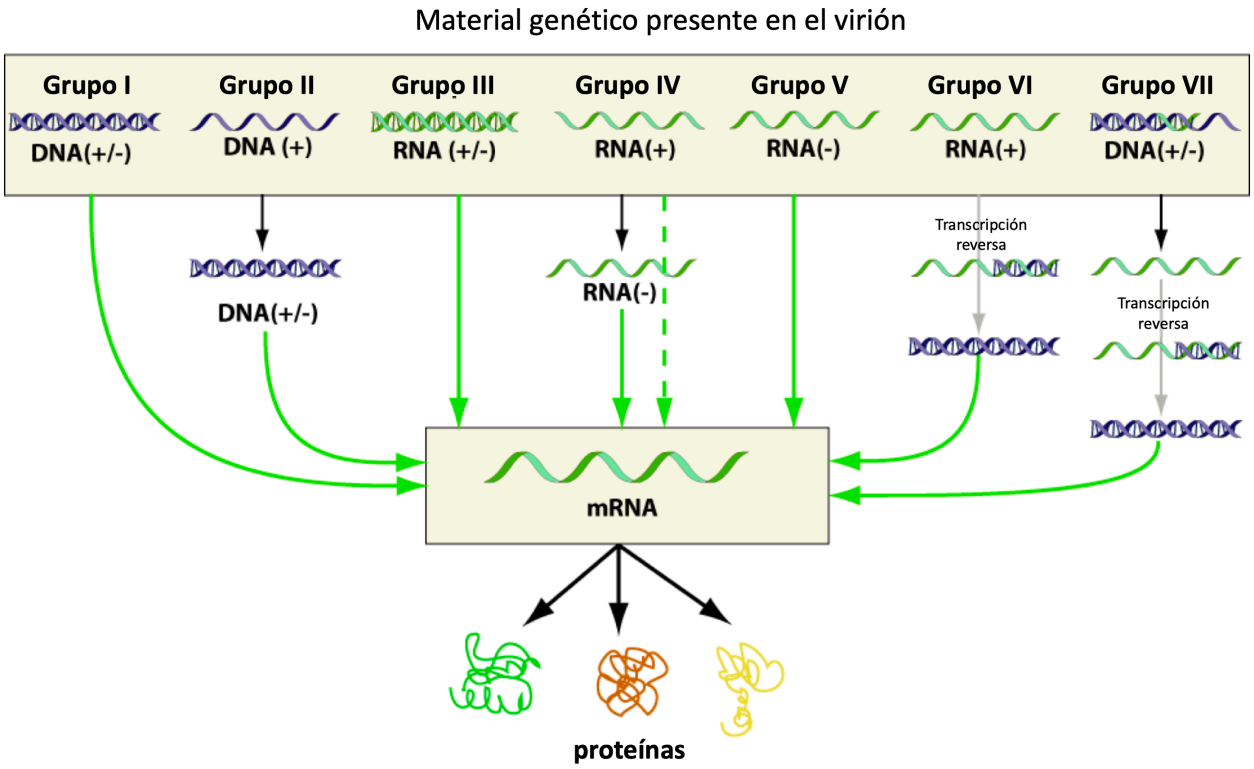


Figura 2. Clasificación de Baltimore. Las cadenas de DNA y RNA se muestran ilustradas en color azul y verde, respectivamente. Las flechas negras delgadas indican la síntesis de una cadena complementaria de DNA o RNA. Las flechas verdes continuas indican transcripción. La flecha verde punteada hace referencia a que al genoma de los virus ssRNA(+) equivale a un mRNA. Las flechas grises indican transcripción reversa. Las flechas negras gruesas indican traducción. Modificada de ViralZone (www.expasy.org/viralzone).

La duplicación de genes se ha reportado como un evento frecuente en la evolución de virus de dsDNA (Shackelton & Holmes, 2004). En un estudio reciente (Gao et al., 2017) se analizaron 201 genomas de virus de dsDNA y se encontraron casos de duplicación en 42.3% de estos virus. Las 1874 proteínas que identificaron como parálogas se distribuyen en 612 familias de proteínas compuestas desde 2 hasta 61 miembros. Además, se encontró una correlación positiva entre el número de parálogos y el tamaño de los genomas, que en virus de dsDNA

pueden llegar hasta 2473 kbp (*Pandoravirus salinus*). Se conocen varios ejemplos de duplicaciones en las familias *Adenoviridae* (Davison et al., 2003), *Herpesviridae* (McGeoch & Davison, 1999) y *Poxviridae* (Hughes & Friedman, 2005), y se ha sugerido que algunas de estas podrían entenderse a partir de las interacciones entre las defensas antivirales del hospedero y los mecanismos de evasión del virus (Gao et al., 2017). En cambio, en virus de RNA se ha reportado a la duplicación de genes como un evento poco frecuente (Simon-Loriere & Holmes, 2013). En el estudio realizado por Simon-Loriere y Holmes (2013), solamente se identificaron nueve casos significativos a nivel de secuencia: cuatro en ssRNA(+), tres en ssRNA(RT) y dos en ssRNA(-). Cabe destacar que solo un caso corresponde a un virus segmentado (*Benyvirus* ssRNA(+)) y ninguno a virus de dsRNA.

Al parecer, a los virus de RNA no les faltan los mecanismos que pueden generar duplicaciones, como la replicación por elección de copia o el reclutamiento de segmentos (Simon-Loriere & Holmes, 2012), de manera que la poca frecuencia con la que se observan duplicaciones en estos virus puede deberse a las presiones que limitan el crecimiento de sus genomas. Por ejemplo, la alta tasa de mutación, la incapacidad para abrir dobles hélices o la competencia con otros virus y la interacción con el sistema inmune del hospedero que podrían favorecer una rápida replicación (Reanney, 1982; Holmes, 2009). La importancia de la capacidad de corrección y de la capacidad de abrir dobles hélices para el crecimiento de los genomas se hace evidente en los virus de dsDNA. Por ejemplo, las polimerasas replicativas y de reparación de las familias A y B solamente se encuentran en el 10% de los virus de dsDNA con genomas menores a 40 kbp, pero se encuentran en el 100% de los virus de dsDNA con genomas mayores a

140 kbp. En el caso de las helicasas replicativas, no solamente se sigue un patrón similar relacionado con el tamaño del genoma sino que, además, son las enzimas involucradas en la replicación que se encuentran codificadas con mayor frecuencia en los genomas de virus de dsDNA (Kazalauskas et al., 2016). En el caso de los virus de RNA (que en promedio tienen genomas de 10 kbp), se ha sugerido que el genoma tan excepcionalmente grande de los virus de ssRNA(-) de la familia *Coronaviridae* (hasta 33.452 kbp del *Ball python nidovirus*) se debe a que estos codifican un dominio helicasa Hel y un domino exoribonucleasa 3'-5' ExoN involucrados en la apertura de dobles hélices de RNA y en la reparación durante la replicación, respectivamente (Holmes, 2009). Otra estrategia que se ha observado en virus de RNA que llegan a tener genomas cercanos a los 30 kbp es la segmentación de genomas como en el caso de los virus de dsRNA de la familia *Reoviridae*. Sin embargo, el estudio de Simon-Loriere & Holmes (2013) no encontró casos de duplicación en ninguna de estas familias de virus de RNA. Por otro lado, en un trabajo de evolución dirigida en el cual se introdujeron duplicaciones de manera artificial en un virus ssRNA(+) de la familia *Potyviridae*, se observó una pérdida de adecuación. En este estudio se sugirió que algunos factores adicionales que podrían afectar la estabilidad de las duplicaciones en virus de RNA podrían ser el correcto procesamiento de la poliproteína y el uso de más recursos para la replicación y expresión de un genoma más grande (Willemsen et al., 2016). A pesar de todas las evidencias que argumentan en contra de la estabilidad de las duplicaciones en virus de RNA, es posible que la alta tasa de mutación de los virus de RNA, además de limitar el tamaño de sus genomas, dificulte la detección de genes parálogos por comparación de estructuras primarias de proteínas. Esta aproximación solo podría reconocer duplicaciones ocurridas en su historia evolutiva más reciente, pasando por alto casos

que podrían revelar a la duplicación de genes como un evento más frecuente de lo que se tiene pensado. Una estrategia alternativa para detectar casos elusivos de proteínas parálogas es la comparación de estructuras terciarias de proteínas que, es bien sabido, conservan mejor las señales de homología.

2 Antecedentes

2.1 Duplicación de genes en virus de RNA

En el estudio realizado por Simon-Loriere y Holmes (2013) usaron BLASTP (valor e < 10-5) para analizar casos de duplicación en 1198 virus, encontrando solo nueve casos significativos. Cuatro corresponden a ssRNA(+): dos en la familia *Closteroviridae* (proteínas de la cápside CP y CPm, y CPm1 y CPm2), uno en el género *Benyvirus* (factor de patogenicidad p25 y p26) y uno en la familia *Picornaviridae* (proteína con función de cebador Vpg). Dos corresponden a ssRNA(-): ambos en la familia *Rhabdoviridae* (glicoproteína G y Gns, y U1 y U2 de función desconocida). Y tres corresponden a ssRNA(RT): todos en la familia *Retroviridae* (proteína de proliferación celular y reguladora de la transcripción orfA y orfB, orf1 y orf2 de función desconocida, y vpr y vpx asociadas a la supresión de la fase G2 de la mitosis y la translocación nuclear del complejo de preintegración, respectivamente). No se encontró ningún caso asociado a dsRNA.

Como antecedente directo a este proyecto (Cisneros-Martínez, 2016), recientemente se identificaron, por comparación de estructuras terciarias, cuatro casos adicionales de duplicación en proteínas de virus de RNA. Estos incluyen a las proteínas de la cápside de virus ssRNA(+) del orden de los *Picornavirales* (VP1, VP2 y VP3), las cisteína proteasas de virus ssRNA(+) de la familia *Picornaviridae* (2A y 3C) y las subunidades KP6α y KP6β de una proteína citotóxica de un virus de dsRNA de la familia *Totiviridae*. Juntando los casos de duplicación identificados por

comparación de secuencias con aquellos identificados por comparación de estructuras terciarias, se obtiene un total de 13 casos reportados hasta ahora (Tabla 1). Aunque 13 casos de duplicación aún parecen ser muy pocos, es destacable la conservación de estas proteínas parálogas a pesar de todas las presiones que podrían limitar el tamaño de estos genomas. Esto sugiere que la estabilidad de las duplicaciones en genomas de RNA depende de que la ventaja de tener una duplicación sea mayor a la desventaja de tener un genoma más grande (Cisneros-Martínez, 2016). Ya que la estructura terciaria permitió identificar a estas proteínas como homólogas, es posible que la construcción de dendrogramas de similitud estructural pueda corroborar la relación paráloga de estas proteínas o revelar una historia evolutiva más compleja involucrando procesos de HGT. Esta tesis se enfoca en el caso de las cápsides de los virus del orden *Picornavirales*.

Tabla 1. Lista de los 13 casos de duplicación de genes identificados por comparación de secuencias (Simon-Loriere & Holmes, 2013) y por comparación de estructuras terciarias de proteínas (Cisneros-Martínez, 2016). En negritas se destacan las proteínas de la cápside del orden *Picornavirales*.

Organización del genoma	Familia o género	Proteínas parálogas	Nivel de detección
ssRNA(+)	*Closteroviridae*	CP-CPm	Secuencia
		CPm1-CPm2	Secuencia
	Picornaviridae	VPg-VPg	Secuencia
		2A-3C	Estructura terciaria
		VP1-VP2-VP3	Estructura terciaria
	Dicistroviridae	**VP1-VP2-VP3**	Estructura terciaria
	Secovivirdae	**S-L(NTD)-L(CTD)**	Estructura terciaria
	Benyvirus	p25-p26	Secuencia
ssRNA(-)	*Rhabdoviridae*	G-Gns	Secuencia
		U1-U2	Secuencia
ssRNA(RT)	*Retroviridae*	orfA-orfB	Secuencia
		orf1-orf2	Secuencia
		vpr-vpx	Secuencia
dsRNA	*Totiviridae*	KP6α-KP6β	Estructura terciaria

2.2 Cápsides de virus del orden *Picornavirales*

En algunos virus del orden *Picornavirales*, como los de la familia *Picornaviridae*, los genes de la cápside se traducen como una sola poliproteína que es procesada proteolíticamente en VP0, VP3 y VP1. VP0 luego es procesada sin intervención de una proteasa en VP4 y VP2, salvo en los géneros *Kobuvirus* y *Parechovirus*, en los cuales permanece como VP0 (Sabin et al., 2016; Kalynych et al., 2016). En virus de otras familias, como *Iflaviridae* o *Dicistroviridae*, en lugar de que VP4 sea una extensión N-terminal de VP2, esta se encuentra como extensión N-

terminal de VP3 (Liljas et al., 2002; Kalynych et al., 2016). En la familia *Secoviridae* no hay VP4 y, en algunos virus de esta familia, los dominios que ocupan las posiciones equivalentes a VP1, VP2 y VP3 (domino A, C y B, respectivamente) permanecen unidos formando una sola proteína de tres dominios (e. g. *Nepovirus*) o una proteína grande de dos dominios (L) y otra chica de un dominio (S) (e. g. *Comovirus*). VP1, VP2 y VP3 son barriles β de tipo *jelly roll*, de aproximadamente 250 residuos de largo, formados por ocho hebras antiparalelas (hebras B-I). 60 copias de la poliproteína estructural, compuesta por una copia de VP1, VP2 y VP3 se ensamblan para formar una cápside icosaédrica de aproximadamente 30 nm de diámetro. Las 60 copias dan un total 180 barriles β, usualmente representados como unidades trapezoides. Debido a que en estos virus la unidad asimétrica se compone de tres proteínas diferentes, su número de triangulación T, que es un indicador del tamaño de la cápside equivalente al número de subunidades en la unidad asimétrica, se denomina T=pseudo3 (T=p3). Esto los distingue de otros virus con cápsides T=3, sin representantes en el orden *Picornavirales*, como los de la familia *Tombusviridae* o *Solemoviridae*, en los cuales la cápside se ensambla a partir de la misma proteína (CP) expresada 180 veces. En este caso, cada subunidad ocupa ambientes estructurales diferentes (A, C y B) (Rossmann et al., 1985; Rossmann & Johnson, 1989) (Figura 3). En ocasiones, el dominio *jelly roll* de estas proteínas de la cápside es llamado dominio *shell* (S) para distinguirlo de un dominio protruyente (P) presente en el extremo C-terminal de algunas proteínas de la cápside. Tanto las cápsides T=3 como las T=p3 tienen cuatro ejes de simetría: en los *Picornavirales*, I) el eje quíntuple que se encuentra en el centro de los pentámeros, en donde colindan cinco copias de VP1 estabilizadas con el extremo N-terminal de VP3 y con VP4; II) el eje doble que se encuentra en la frontera entre pentámeros, involucrando principalmente a VP2;

III) el eje triple formado por tres pentámeros que en su intersección presentan tres copias intercaladas de VP2 y VP3; y IV) el eje quasi triple (q3) que se encuentra en el centro del triángulo formado por VP1, VP2 y VP3 (Rossmann & Johnson, 1989).

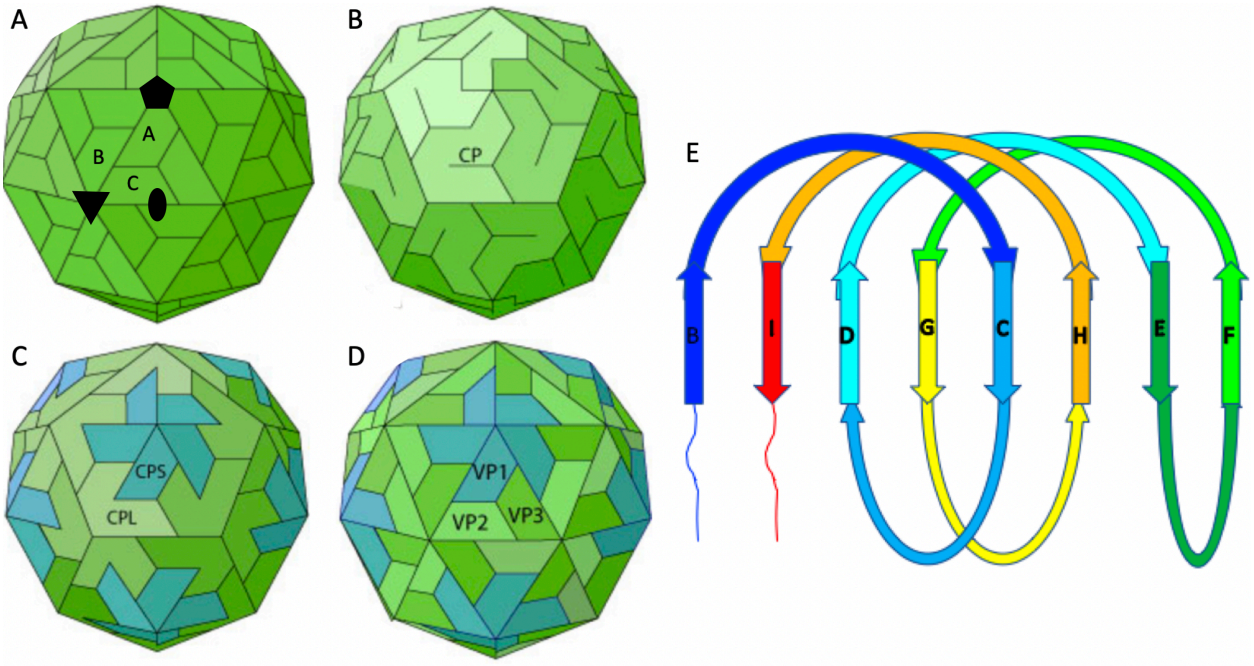


Figura 3. Estructura general de las cápsides de virus del orden *Picornavirales*. (A) Para comparación, una cápside T=3 típica de virus de plantas como los *Solemoviridae*. Se ensambla a partir de 180 repeticiones de la misma proteína de la cápside. Solo en esta se muestran los ejes de simetría doble (óvalo), triple (triángulo) y quíntuple (pentágono). (B) Cápside T=p3 típica de los *Nepovirus* de la familia *Secoviridae* del orden *Picornavirales*, en los cuales los tres dominios *jelly roll* permanecen unidos. (C) Cápside T=p3 típica de los *Comovirus* de la familia *Secoviridae* del orden *Picornavirales*, en donde solo uno de los dominios *jelly roll* es procesado por la proteasa. La cápside se ensambla a partir de una proteína corta (CPS) de un solo dominio *jelly roll* y una proteína larga (CPL) de dos dominios *jelly roll*. (D) Cápside T=p3 típica de *Enterovirus* de la familia *Picornaviridae* del orden *Picornavirales*. VP1, VP2 y VP3 son procesados proteolíticamente. (E) Representación esquemática del dominio *jelly roll* coloreado de amino terminal a carboxilo terminal en un gradiente del azul al rojo. Caricaturas A-D modificadas de ViralZone

(www.expasy.org/viralzone). La nomenclatura de los elementos de estructura secundaria están basados en el *Southern bean mosaic virus* (SBMV) del género *Sobemovirus* (Rossmann et al., 1985).

2.3 Evolución de las cápsides de virus del orden *Picornavirales*

El origen por duplicación de las subunidades VP1, VP2 y VP3 de las cápsides de los virus del orden *Picornavirales* se ha sugerido con anterioridad (Chandrasekar & Johnson, 1997; Liljas et al., 2002). Esta relación no se ha podido confirmar a nivel de secuencia (Simon-Loriere & Holmes, 2013). Sin embargo, ya hay evidencia cuantitativa sobre la similitud estructural entre estas proteínas (Cisneros-Martínez, 2016). Respecto a la evolución de estas proteínas de la cápside, se han hecho varios dendrogramas de similitud estructural usando a los protómeros completos (Tuthill et al., 2009; Wang et al 2015; Kalynych et al., 2016; Sabin et al., 2016; Spurny et al., 2017). Estos análisis ayudan a establecer las relaciones filogenéticas entre los virus del orden *Picornavirales*, pero no permiten poner a prueba la posible relación paráloga entre las subunidades del protómero (VP1, VP2 y VP3). El árbol de protómeros que incluye a más familias y géneros del orden de los *Picornavirales* es el de Spurny et al. (2017). En este, determinaron la estructura de la cápside del *Black queen cell virus* (BQCV), un virus de la familia *Dicistroviridae*, perteneciente al género *Triatovirus*. El árbol incluye a virus de la familia *Dicistroviridae* e *Iflaviridae*, que infectan invertebrados, y a virus de la familia *Picornaviridae* que infectan vertebrados (Spurny et al., 2017) (Figura 4). En ningún estudio se han incluido los modelos estructurales de las proteínas de la cápside de los virus de la familia *Secoviridae*. Estos virus infectan plantas y sus proteínas de la cápside se han propuesto como eslabones en la

evolución de cápsides T=p3 a partir de T=3 por eventos de duplicación (Chandrasekar & Johnson, 1997).

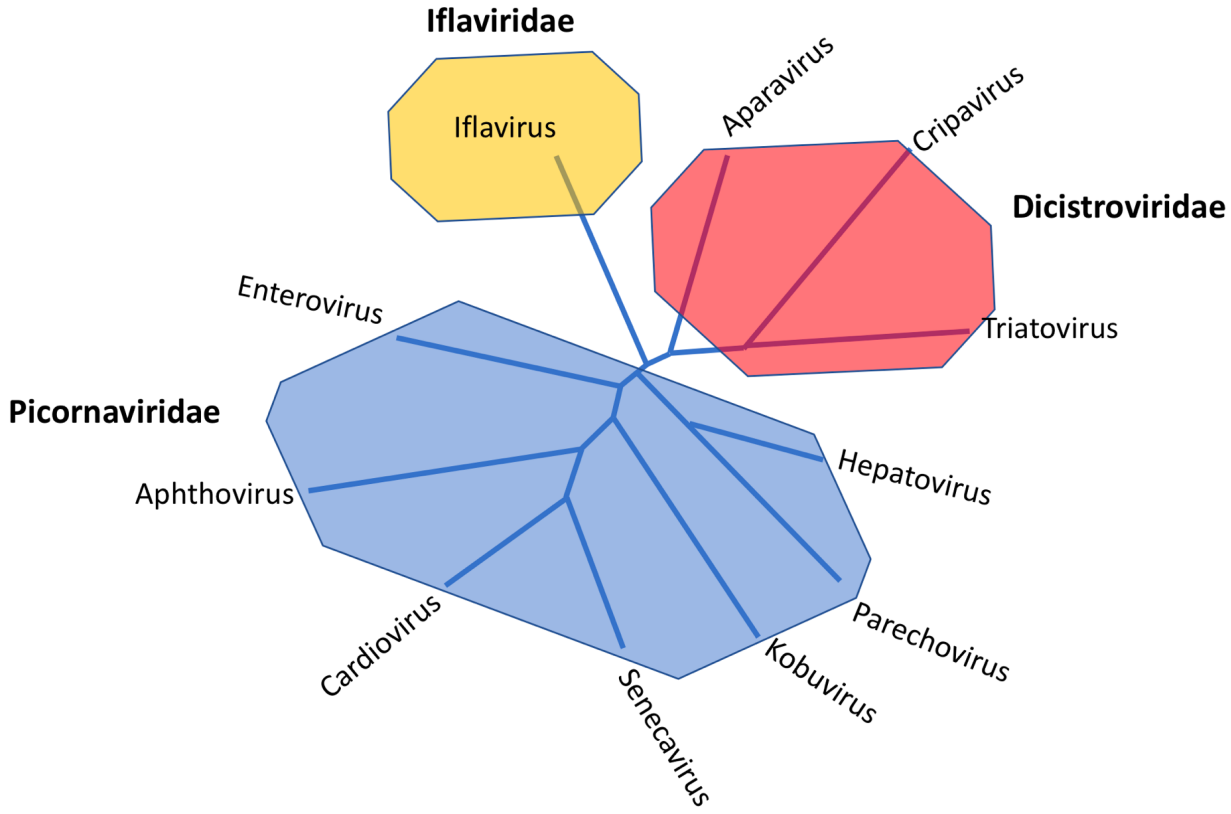


Figura 4. Esquema que resume los árboles de similitud estructural realizados con los protómeros completos. Los fondos azul, rojo y anaranjado engloban a los géneros de las familias *Picornaviridae*, *Dicistroviridae* e *Iflaviridae*, respectivamente. Basada en Spurny et al., (2017).

En un estudio reciente (Krupovic & Koonin, 2017), se muestra que las estructuras de *Dicistroviridae*, *Iflaviridae* y *Picornaviridae* (solo VP1), se encuentran estructuralmente relacionadas con las cápsides *single jelly roll* (SJR) de las familias *Tombusviridae*, *Solemoviridae*, *Hepeviridae*, *Astroviridae*, *Caliciviridae*, *Tymoviridae* y *Circoviridae* (ssDNA). Además, proponen que las proteínas de cápsides virales de tipo *jelly roll* se pudieron haber originado por reclutamiento de proteínas celulares que presentan el mismo plegamiento. Es

posible que estas proteínas celulares hayan sido algunas con unión a carbohidratos, lo cual le habría brindado a los virus un mecanismo de unión a la superficie celular del hospedero (Krupovic & Koonin, 2017).

3 Objetivos

Los objetivos de esta tesis parten de la hipótesis de que los genomas del mundo de RNA/ proteínas crecieron, principalmente, por acumulación de genes duplicados. Este estudio pretende abordar el problema a partir de las propiedades de los genomas de RNA, ya sean de virus de RNA o de las hipotéticas células del mundo de RNA/proteínas. Además, se pretende explotar el potencial de la información estructural de las proteínas para detectar eventos evolutivos en genomas con altas tasas de divergencia. En particular, se pretende analizar la evolución de las proteínas de la cápside de los virus del orden *Picornavirales* a través de la evaluación de las señales de homología a nivel de estructura primaria, secundaria y terciaria de las proteínas VP1, VP2 y VP3. Así mismo, se desea identificar los homólogos lejanos de estas proteínas de la cápside para reconstruir su historia evolutiva a partir de la construcción de dendrogramas de similitud estructural.

4 Metodología

4.1 Similitud pareada a nivel de estructura primaria, secundaria y terciaria

Para conocer el porcentaje de identidad de secuencia entre VP1, VP2 y VP3, se seleccionó un modelo estructural representante del género *Enterovirus* que tuviera la mayor resolución (PDB: 4q4w) (Zocher et al. 2014). Las cadenas se compararon con el servidor en línea de DALI All against all (Holm & Laakso, 2016) (enero de 2018) y se extrajo el porcentaje de identidad y el valor Z pareado, así como los alineamientos pareados de secuencia y de estructura secundaria.

4.2 Señales de homología a nivel de secuencia

Para saber si las secuencias de VP1, VP2 y VP3 conservan cierta información que permita detectarlas, de manera recíproca, como homólogos lejanos, se realizaron tres búsquedas independientes con JACKHMMER (Finn, et al. 2011) (versión en línea en febrero de 2018) contra la base de datos PDB y BLOSUM45 como matriz de sustitución inicial. Los demás parámetros se usaron con la configuración establecida por default. Cada búsqueda se llevó hasta la convergencia (14, 7 y 9 iteraciones, respectivamente), generando 302 resultados significativos para VP1, 199 para VP2 y 285 para VP3. Los tres archivos de resultados se concatenaron en un solo archivo (786 secuencias) al cual se le redujo la redundancia al 90% de identidad con CD-HIT (Li & Godzik, 2006; Fu et al. 2012) (versión 4.6.1-2012-08-27). Con las secuencias

resultantes (138 secuencias) se realizó un alineamiento múltiple de secuencia usando información estructural con PROMALS3D (Pei et al., 2008) (versión en línea en mayo de 2018). El alineamiento se analizó con Jalview (versión 2.10.2), coloreando a los residuos por las propiedades físico-químicas de sus cadenas laterales (alifático/hidrofóbico (ILVAM) = rosa; aromático (FWY) = anaranjado; positivo (KRH) = azul; negativo (DE) = rojo; hidrofílico (STNQ) = verde; conformacionalmente especial (PG) = magenta; formador de puentes disulfuro (C) = amarillo) solo si se encontraban en por lo menos el 90% del total de secuencias. La conservación se analizó de la misma manera por grupos (VP1s, VP2s y VP3s). Para analizar la conservación de estos residuos desde un enfoque estructural, se seleccionaron 29 secuencias representativas del alineamiento múltiple y se alinearon sus modelos estructurales con el programa STAMP (Russell & Barton, 1992) (scanslide=5, scanscore=5) contenido en la herramienta Multiseq (Roberts et al. 2006) del programa VMD (Humphrey et al. 1996) (versión 1.9.3). Las superposiciones se visualizaron con Chimera (Pettersen et al. 2004) (versión 1.11.2).

4.3 Homólogos lejanos

Para extender la lista de homólogos, un conjunto de los resultados de JACKHMMER agrupados al 30% de identidad de secuencia con PSI-CD-HIT (27 secuencias) se alineó estructuralmente con STAMP. El alineamiento múltiple de secuencias resultante se usó como archivo de entrada para HHpred (Zimmermann et al., 2018) (versión en línea en febrero de 2018), con el cual se realizó una búsqueda contra el PDB, seleccionando las opciones HHblits -> uniclust30 para el método de generación del alineamiento múltiple y limitando la probabilidad

mínima de hits a 30%. Por otro lado, se realizaron tres búsquedas independientes (VP1, VP2 y VP3) con DALI. Del resultado contra el PDB25, se consideraron como homólogos previamente no detectados a aquellos con Z mayor a 2 en por lo menos dos de las tres búsquedas. De cada género viral representado al final de las búsquedas con HHpred y DALI (HHpred: 23 géneros y uno no asignado; DALI: tres géneros) se seleccionó un archivo de coordenadas PDB con la mayor resolución (determinado por cristalografía de rayos-x). Esto generó una lista de 53 proteínas o dominios homólogos derivados de 27 archivos de coordenadas PDB. Los 53 modelos estructurales se volvieron a alinear con STAMP para realizar otra búsqueda en HHpred con el alineamiento múltiple de secuencia resultante. En esta búsqueda solamente se encontró un homólogo previamente no detectado. En total se obtuvieron 54 proteínas o dominos homólogos.

4.4 Relación evolutiva

Para saber si la señal de homología a nivel de secuencia es suficiente para resolver la relación evolutiva de estas proteínas se realizó un árbol de máxima verosimilitud con PhyML (Guindon et al. 2010) (versión 3.1). El alineamiento de secuencia fue generado a partir de una nueva superposición estructural, con el programa STAMP, de los 54 modelos estructurales. El modelo evolutivo fue seleccionado con ProtTest (Darriba et al. 2011) (versión 3.2). La filogenia se realizó con BLOSUM62 + G + F, con cuatro categorías de tasa de sustitución, parámetro gamma = 6.623, árbol inicial con BioNJ, con optimización de la longitud de ramas, SPRs en la búsqueda de las topologías del árbol y 100 bootstrap. El resultado de esta filogenia motivó el análisis de la relación evolutiva de estas proteínas a través de la comparación de sus modelos

estructurales. Primero, a partir de la misma superposición estructural de la cual se extrajo el alineamiento múltiple de secuencia, se usó la herramienta Multiseq del programa VMD para generar una matriz de distancias derivadas de la medida de homología Q_H (O'Dongue et al. 2003) la cual se define como:

$$Q_H = N[q_{aln} + q_{gap}]$$

En donde q_{aln} es una medida de la fracción de pares de distancias entre carbonos α que son iguales o similares entre dos estructuras alineadas, q_{gap} es una estimación de la desviación estructural inducida por inserciones en cada proteína del par alineado y N es una normalización que depende del número de residuos alineados, el número de residuos que forman parte de las inserciones, el numero de inserciones en cada proteína del par alineado y el número de inserciones simultáneas (para más detalles ver O'Dongue et al. 2003). Los valores de Q_H varían de 0 a 1, en donde 1 equivale a proteínas idénticas, mientras que en la medida de distancia derivada de Q_H 0 equivale a proteínas idénticas. Esta matriz se introdujo en el programa FITCH del paquete PHYLIP (Felsenstein, 1989) (versión 3.695) para realizar el árbol de similitud estructural. Se realizaron 100 árboles con orden aleatorio diferente y rearreglos globales por poda y reinjerto de ramas, y un árbol consenso de mayoría de todos los árboles muestreados. Adicionalmente, se realizó un dendrograma de similitud estructural evaluada por superposiciones pareadas con el programa SSM (Krissinel & Henrick, 2004) (PDBe Fold v2.59. (src3) 14 Apr 2014, usada en línea en marzo de 2018) y una matriz de distancias estimadas a partir del *Structural Alignment Score* (SAS) (Subbiah, 1993), el cual es calculado a partir de la fórmula:

$$SAS = \frac{100RMSD}{Naln}$$

En donde RMSD es la desviación cuadrática media entre carbonos α y Naln el número de residuos alineados. En el caso de un modelo estructural sin asignación de estructura secundaria (3zxa) SSM no puede realizar los alineamientos. En este caso, todos los valores de RMSD y Naln se obtuvieron con la herramienta Match-Align de Chimera a partir de las superposiciones estructurales de STAMP. La misma estrategia se siguió para 11 comparaciones en las que SSM arrojó valores de RMSD exageradamente altos o Naln exageradamente bajos. Los árboles se analizaron con base en los caracteres estructurales que justifican las agrupaciones como, por ejemplo, las inserciones particulares de cada proteína, y en términos de los hospederos. Adicionalmente, se analizó el origen de los dominios P (no incluidos en las comparaciones estructurales), presentes en algunos de estas proteínas de la cápside, con búsquedas de DALI.

5 Resultados y discusión

5.1 Similitud pareada a nivel de estructura primaria, secundaria y terciaria

Las comparaciones pareadas confirman que hay una mayor conservación de la estructura terciaria que de la estructura primaria en estas proteínas. A nivel de secuencia, VP2 y VP3 tienen 11% de identidad, mientras que VP1 y VP2, como VP1 y VP3, tienen apenas 10% de identidad. En cambio, el valor Z indica una similitud estructural significativa entre las tres proteínas (>2). Estos son 7.6 entre VP1 y VP2, 9.5 entre VP1 y VP3, y 11.1 entre VP2 y VP3 (Tabla 2).

Tabla 2. Comparación pareada entre VP1, VP2 y VP3. Porcentaje de identidad en la diagonal superior (celdas con fondo blanco) y valor Z en la diagonal inferior (celdas con fondo gris).

	VP1	**VP2**	**VP3**
VP1	-	10	10
VP2	7.6	-	11
VP3	9.5	11.1	-

Al observar los alineamientos pareados se hace evidente la gran divergencia de las secuencias. En cada alineamiento hay diferentes posiciones idénticas, sin embargo, las únicas identidades que coinciden en los tres alineamientos corresponden a un par de prolinas al final de las hebras βE y βG, respectivamente. En cambio, los elementos de estructura secundaria coinciden muy bien, con la salvedad de algunas inserciones (Figura 5).

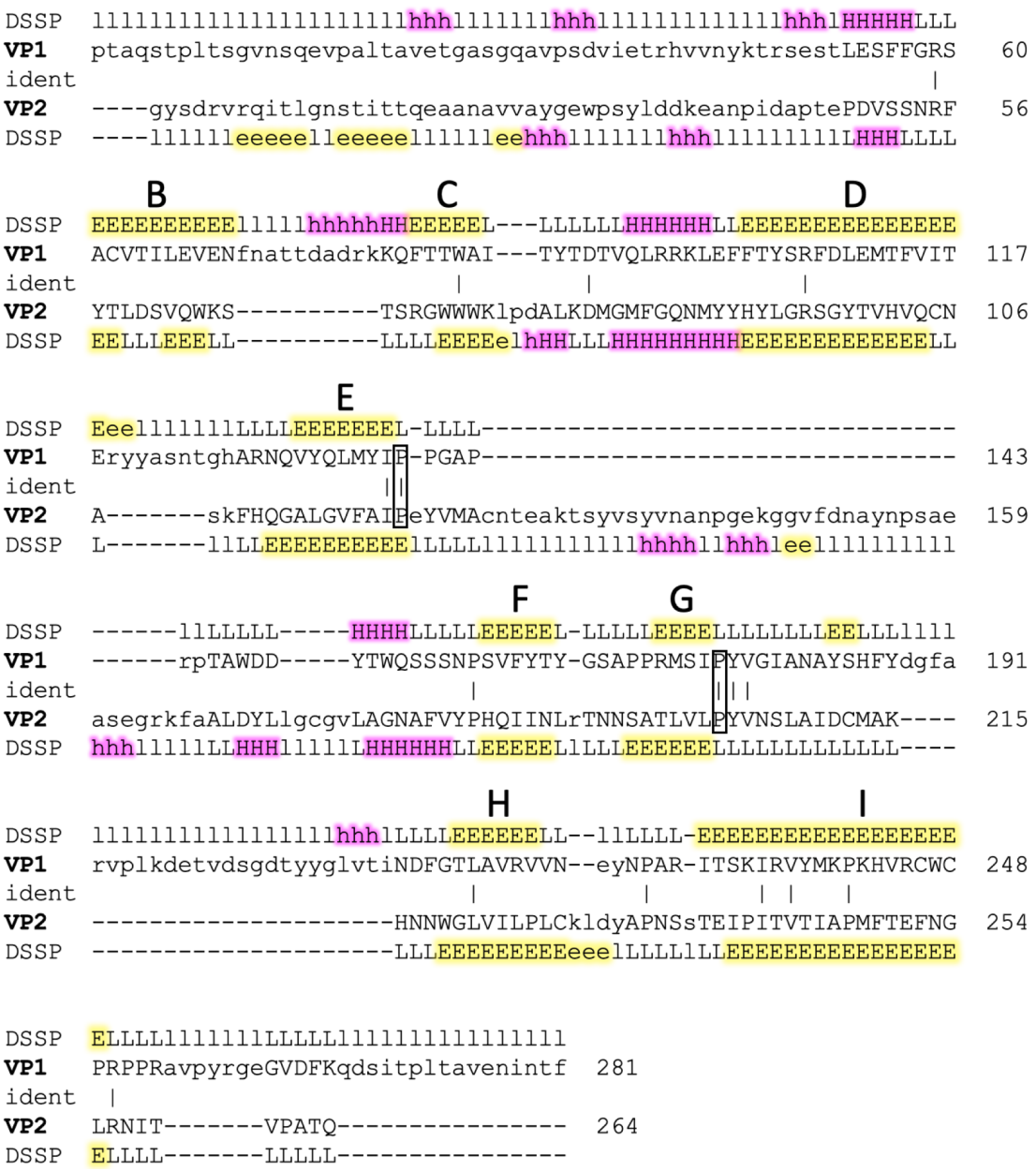


Figura 5. Alineamiento pareado de estructura primaria y secundaria derivado de la superposición estructural con DALI. Se muestra el alineamiento entre VP1 y VP2. Para alineamiento entre VP1 y VP3, y entre VP2 y VP3 ver Anexo 1. Cada hebra β (B-I) está señalada en la parte superior del alineamiento. Las hebras están resaltadas en amarillo y las hélices en magenta. El par de prolinas ubicadas al final de las hebras βE y βG que coinciden en los tres alineamientos están señaladas con rectángulos.

A nivel de estructura terciaria las regiones de mayor conservación corresponden a las hebras β que forman el barril. Las mayores diferencias corresponden al extremo N-terminal y a las inserciones particulares de estas proteínas. Estas son a) el llamado *loop Foot and mouth disease virus* (FMDV) en VP1 (entre las hebras βG y βH), b) el *puff* en VP2 (entre las hebras βE y βF) y c) el *knob* en VP3 (en la hebra βB) (Rossman et al., 1985; Rossmann & Johnson, 1989) (Figura 6). Estas inserciones forman un relieve particular que, en *Enterovirus*, se caracteriza por la presencia de una depresión denominada "cañón", en donde se une el receptor celular CD155 o Pvr (Strauss et al., 2015)

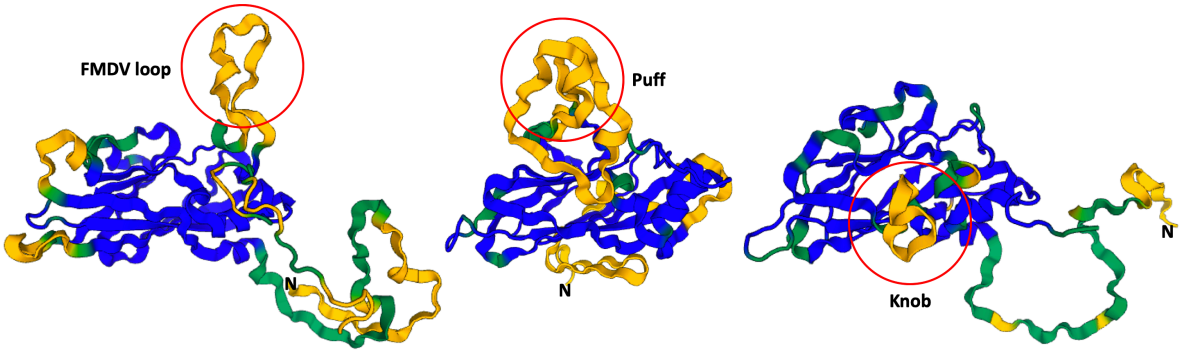


Figura 6. Regiones estructuralmente conservadas en VP1, VP2 y VP3. A la izquierda VP1, en el centro VP2 y a la derecha VP3. Las regiones mejor conservadas se muestran en azul, y las menos conservadas en amarillo. El *loop* FMDV de VP1, el *puff* de VP2 y el *knob* de VP3 están señalados por círculos rojos.

Puesto que en las búsquedas de parálogos por comparación de secuencias no se detectaron a estas proteínas de la cápside (Simon-Loriere & Holmes, 2013), era previsible obtener tan bajos porcentajes de identidad en las comparaciones pareadas. En cambio, su similitud estructural es evidente. El hecho de que las mayores diferencias corresponden a inserciones que tienen implicaciones en la unión al receptor celular y en la formación de

superficies antigénicas (Rossman et al., 1985; Rossmann & Johnson, 1989), sugiere que estas interacciones con el hospedero han supuesto fuertes presiones de selección durante la evolución de estas proteínas.

5.2 Señales de homología a nivel de secuencia

De las tres búsquedas con JACKHMMER, la búsqueda con VP3 (285 hits significativos en 9 iteraciones) es la que encuentra más pronto a las otras proteínas de la cápside. Desde la primer iteración encuentra a proteínas de la cápside (CP) de virus de la familia *Caliciviridae* (no *Picornavirales* con cápside T=3) y, a partir de la segunda, encuentra representantes de VP1 y VP2. En la búsqueda con VP2 (199 hits significativos en 7 iteraciones) se encuentran representantes de VP3 a partir de la segunda iteración, cápsides de virus de la familia *Caliciviridae* a partir de la tercera, y ningún representante de VP1 en las siete iteraciones. Finalmente, en la búsqueda con VP1 (302 hits significativos en 14 iteraciones) se encuentran representantes de VP3 a partir de la tercera iteración, representantes de VP2 a partir de la quinta, y cápsides de virus de la familia *Caliciviridae* hasta la octava. Es posible que con VP1 se hayan requerido de más iteraciones para encontrar a VP3 y VP2 debido a que es la más divergente de las tres proteínas (Liljas et al., 2002). La misma razón podría explicar por qué VP1 no se encontró en la búsqueda con VP2. En cuanto a la búsqueda con VP3, es posible que VP1 y VP2 aparezcan como homólogos lejanos en menos iteraciones debido a que VP3 es la más conservada de las tres

(Liljas et al., 2002). Esto podría implicar que su primer perfil refleja mejor los patrones de conservación de estas proteínas.

La distribución taxonómica de las tres búsquedas incluye a VP1, VP2 y VP3 de virus de los géneros *Senecavirus*, *Cardiovirus*, *Aphthovirus*, *Enterovirus*, *Kobuvirus*, *Hepatovirus* y *Parechovirus* de la familia *Picornaviridae*, y CP de los géneros *Vesivirus*, *Norovirus* y *Lagovirus* de la familia *Caliciviridae*. También se encontraron VP2 y VP3 del género *Iflavirus* de la familia *Iflaviridae*, y los géneros *Cripavirus*, *Triatovirus* y *Aparavirus* de la familia *Dicistroviridae*, pero ninguna VP1 de estos virus que infectan invertebrados. Esta ausencia podría deberse a que la búsqueda inició con proteínas de *Enterovirus*, pero, sobre todo, a la gran divergencia de VP1. La estrecha relación entre las familias *Picornaviridae* y *Caliciviridae* es bien conocida y por lo tanto encontrar estas CP no es motivo de sorpresa. La relación entre los *Caliciviridae* y los virus del orden *Picornavirales* es evidente debido a la conservación de una helicasa, una proteasa y la polimerasa de RNA dependiente de RNA (RdRp) (Koonin et al., 2008). Además, sus genomas tienen una cola de poli(A) 3' y llevan una proteína Vpg unida al extremo 5'. No está claro que Vpg sea homóloga, sin embargo, en todos estos virus proporciona un OH 3' que permite el inicio de la replicación (King et al., 2012). Por otro lado, cabe destacar que no se encontraron las proteínas de la cápside de los virus de la familia *Secoviridae*, los cuales son incluidos en el orden *Picornavirales*.

En el alineamiento múltiple realizado con las 138 secuencias que resultaron del agrupamiento con CD-HIT (41 VP1, 46 VP2, 45 VP3 y 6 CP de *Caliciviridae*) se pueden observar pocos residuos conservados dentro de cada grupo (8 en VP1, 10 en VP2, 5 en VP3). Menos aún, en el conjunto unido de VP1, VP2 y VP3, en el cual solo se conservan las mismas dos prolinas al final de βE y βG detectadas por las comparaciones pareadas (Figura 7). Sin embargo, esto no significa que estos pocos residuos conservados son las únicas señales de homología que permiten detectar a estas proteínas como homólogos lejanos a nivel de secuencia, pues, de hecho, hay otros residuos menos conservados y varias similitudes que podrían dar cuenta de los patrones de conservación de esta familia (Anexo 2). Además, es posible que los pocos residuos conservados que se muestran en el alineamiento reflejen un sesgo debido a una sobrerrepresentación de secuencias de *Enterovirus* (58.5% de VP1, 52.2% de VP2 y 53% de VP3). No obstante, se puede corroborar la gran divergencia a nivel de secuencia que hay en esta familia de proteínas.

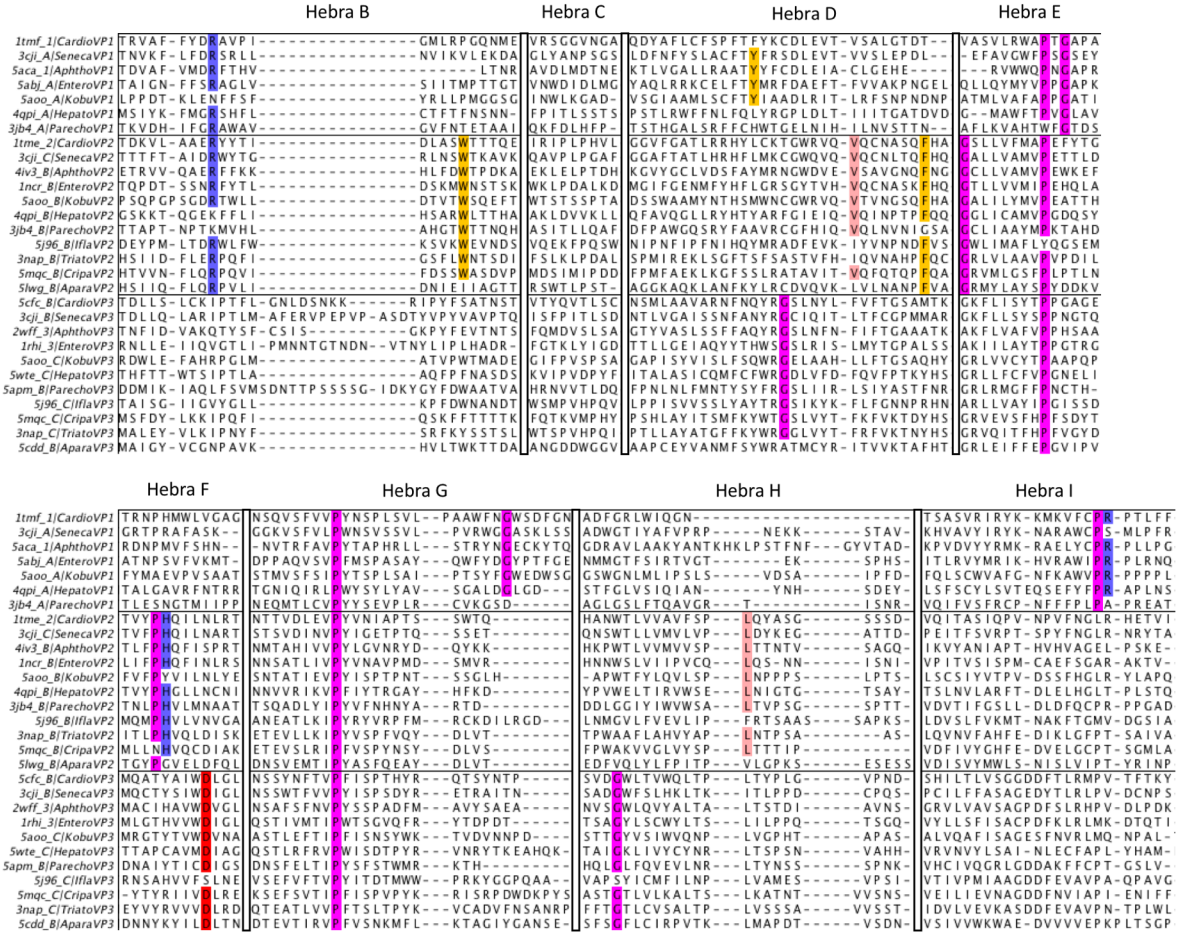


Figura 7. Alineamiento múltiple de secuencias. Por simplicidad, solo se muestra un representante de cada género para VP1, VP2 y VP3 (CP de *Caliciviridae* no mostradas). Además, el alineamiento está recortado para mostrar solo las regiones correspondientes a las ocho hebras β.

El alineamiento múltiple presentado aquí es similar al presentado en Liljas et al. (2002). Ambos usan un enfoque estructural para realizar los alineamientos de secuencia y coinciden en la identificación de muchos de los residuos conservados. La diferencia radica en que Liljas et al. (2002) realizaron alineamientos independientes para VP1, VP2 y VP3, mientras que en este estudio, asumiendo que son proteínas homólogas, se realizó un solo alineamiento con todas las VP1, VP2 y VP3. Algunas coincidencias incluyen a los residuos, identificados en el *Cricket paralysis virus* (CrPV) (PDB:1b35), Arg51 y Pro183 de VP1,

Gly119, Pro143, Asp166, Pro178 y Gly212 de VP3, y Arg78, Trp87, Phe137, Gly140, Pro148, Pro171, Pro188 y Leu216 de VP2. Solo dos de los ocho residuos conservados en VP1 coinciden con los descritos por Liljas et al. (2002). Esto se debe a que en el presente alineamiento solo se ven los residuos de VP1 conservados en la familia *Picornaviridae*, pues no se incluyen secuencias de virus que infectan invertebrados. En VP3 y VP2 las coincidencias son casi exactas. Solamente en el caso de VP2 se identificaron dos residuos conservados que no se mencionan en Liljas et al. (2002), la valina en βD y la histidina en βF (Val130 y Arg172, respectivamente en 1b35), y en VP3 Liljas et al. (2002) identificaron a un Asp243 al final de βI. Otro residuo muy conservado en las tres proteínas es una tirosina, a veces sustituida por fenilalanina o triptófano, inmediatamente después de la prolina conservada al final de βG. Como se describe en Liljas et al. (2002), muchos de los residuos conservados son prolinas o glicinas que se ubican en los extremos de las hebras β o directamente sobre las vueltas (Figura 8). Estos residuos podrían tener una función en el mantenimiento de la conformación correcta de las proteínas. Por ejemplo, Liljas et al. (2002) reportan que Gly119 de VP3 tiene ángulos phi y psi que solo son permitidos en la glicina. Otros residuos como Trp87, Leu142, Leu216 y posiblemente Val130 de VP2 (Anexo 4), forman un grupo de residuos hidrofóbicos con cadenas laterales orientadas hacia el interior del barril que estabilizan su estructura terciaria (Liljas et al., 2002). Hay otros residuos conservados, como las argininas conservadas en VP1 y VP2, Phe137 de VP2 y Asp166 y Asp243 de VP3, que podrían formar contactos intraprotómero e interprotómeros. Por ejemplo, Phe137 interactúa con otras cadenas laterales en el eje triple (Liljas et al., 2002). En el caso de las CP de *Caliciviridae* se puede observar que estas presentan más

residuos característicos de VP2 que de VP1 y VP3. Estos son Trp87, Gly140, Pro171, Lys/ His172 y Leu216. Además, también conservan las prolinas de βE y βG (Anexo 3).

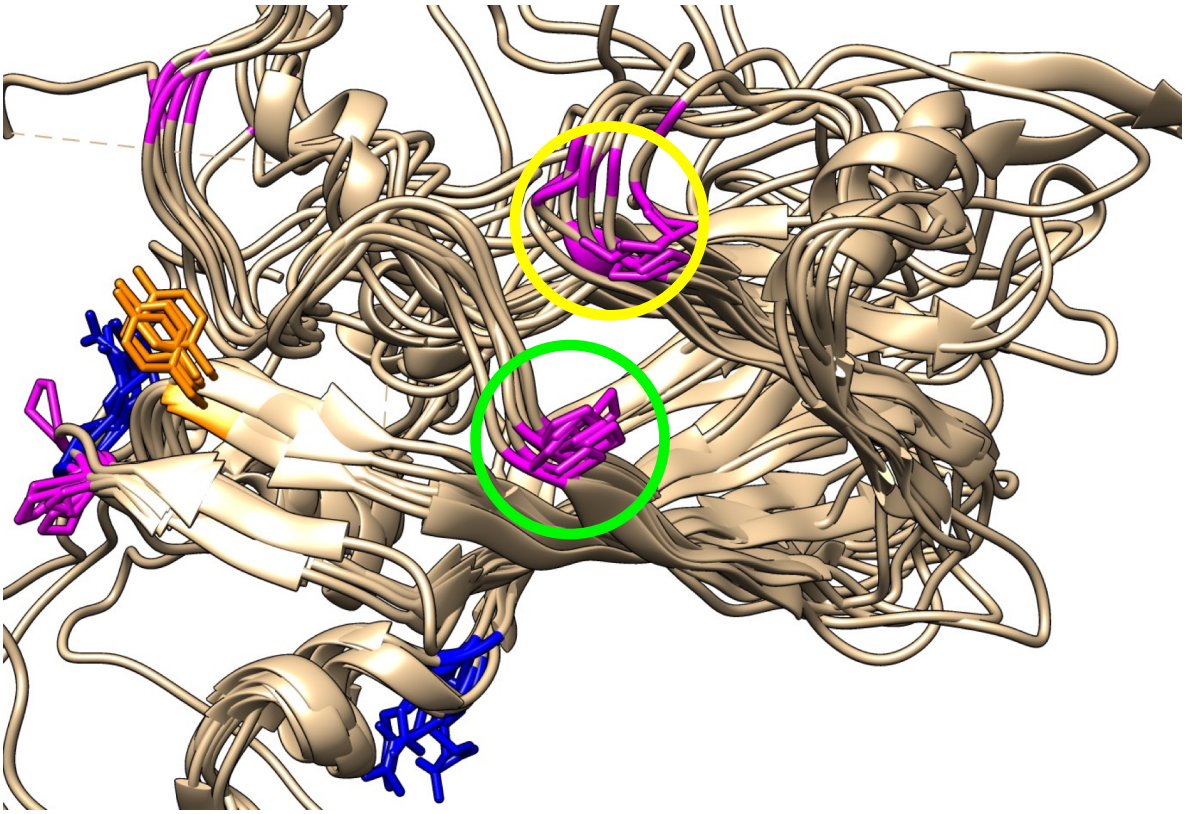


Figura 8. Residuos conservados en VP1 mostrados sobre la superposición de los modelos estructurales representativos de cada género. Las prolinas conservadas en VP1, VP2 y VP3, ubicadas al final de las hebras βE y βG, están señaladas por un circulo amarillo y uno verde, respectivamente. Para VP2 y VP3 ver Anexo 4.

Comparando la ubicación de los residuos conservados con aquellos involucrados en la interacción con el receptor celular de *Enterovirus*, la mayoría de los cuales son de VP1, pero también incluyen a residuos del *loop* FMDV de VP1, *puff* de VP2 y *knob* de VP3 (Strauss et al., 2015), se hace evidente que los residuos conservados parecen tener un rol más bien estructural (Liljas et al., 2002). Mientras que los residuos conservados tienden a

ubicarse en las regiones de contacto entre subunidades, los residuos asociados al reconocimiento del receptor celular se ubican hacia el exterior de la cápside (Figura 9).

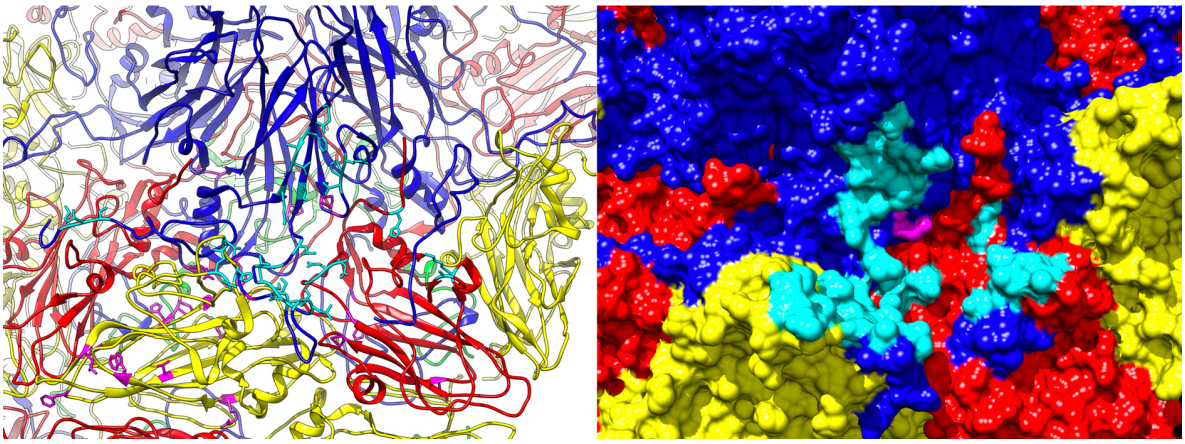


Figura 9. Ubicación de residuos conservados (magenta) y de residuos involucrados en la unión al receptor celular de *Enterovirus* (ciano). A la izquierda se muestran VP1 (azul), VP2 (amarillo) y VP3 (rojo) representadas como caricatura, junto con las cadenas laterales de los diferentes residuos. A la derecha se muestra una representación de superficie. En esta solo se puede observar una glicina conservada de VP1. PDB:3j8f

Para probar el rol de los residuos conservados en las interacciones entre subunidades se realizó un análisis de predicción de puentes de hidrógeno. Este análisis muestra una posible interacción en la interface entre VP1 y VP2 del mismo protómero, en el eje q3, involucrando a un par de residuos conservados en VP1 de *Picornaviridae* (Anexo 5). Otras interacciones involucran, en VP2, a la histidina de βF con la cadena principal, posiblemente estabilizando la vuelta βE-βF, y a la arginina de βB con una asparagina de su propio extremo N-terminal.

Las señales de homología a nivel de estructura primaria y terciaria se complementan para confirmar la ancestral común de estas proteínas. Esto descarta la posibilidad de que algunos residuos, como las prolinas conservadas al final de las hebras βE y βG, se encuentren en sus respectivas ubicaciones debido a causas diferentes a la ancestral común, por ejemplo, convergencias ocasionadas por restricciones estructurales o funcionales. De hecho, la conservación de estos residuos parece responder a presiones de selección que podrían estar relacionadas con la estabilidad de la cápside o con la formación de superficies capaces de interactuar con el hospedero (en el caso de *Enterovirus*, en VP1 una de las prolinas se encuentra al inicio del *loop* FMDV (hebras βG y βH) y en VP2 la otra prolina se encuentra al inicio del *puff* (hebras βE y βF)). La posibilidad de que este par de prolinas se encuentren conservadas debido a presiones de selección natural adquiere soporte adicional del hecho de que la conservación abarca a tres proteínas diferentes en una distribución taxonómica que se extiende desde los *Picornavirales* hasta los virus de la familia *Caliciviridae*.

Las prolinas conservadas en VP1, VP2, VP3 y CP podrían ser residuos presentes en la proteína ancestral antes de los eventos de duplicación, y los residuos característicos de VP1, VP2 o VP3 podrían ser residuos presentes en sus respectivas proteínas ancestrales, después de la duplicación pero antes de la divergencia de los *Picornavirales*. El análisis del contexto estructural en el que se encuentran estos residuos nos puede aportar pistas sobre las circunstancias en las que divergieron o se conservaron estas secuencias. Estas observaciones sugieren que a lo largo de la evolución de estas proteínas de la cápside se

han conservado mejor las características asociadas con la estabilidad de la cápside que aquellas involucradas en la interacción con el hospedero.

5.3 Homólogos lejanos

Para extender la lista de homólogos, un conjunto de los resultados de JACKHMMER agrupados al 30% de identidad de secuencia con PSI-CD-HIT (27 secuencias) se alineó estructuralmente con STAMP. El alineamiento múltiple de secuencias resultante se usó como archivo de entrada para HHpred, con el cual se realizó una búsqueda contra el PDB. El resultado fue una lista de 80 secuencias, de las cuales 57 tienen $P > 95$, 22 tienen $95 > P > 50$, y una $50 > P > 30$. Se podría dudar sobre la homología de aquellos resultados con P menor a 95, sin embargo, todos estos tienen el *jelly roll* y pertenecen a virus de familias que se saben relacionadas. La lista agrega a miembros de las familias *Secoviridae* (*Picornavirales*), *Tombusviridae*, *Solemoviridae* y *Nodaviridae*. Por otro lado, se realizaron tres búsquedas independientes (VP1, VP2 y VP3) con DALI. Del resultado contra el PDB25, se consideraron como homólogos previamente no detectados a aquellos con Z mayor a 2 en por lo menos dos de las búsquedas. Solamente tres modelos estructurales cumplieron con este criterio: CP de un virus de la familia *Hepeviridae*, y CP de los virus satélite *Satellite panicum mosaic virus* (SPMV) y *Satellite tobacco necrosis virus* (STNV). Estos virus satélite solo codifican su CP y dependen del *Panicum mosaic virus* (PMV) y el *Tobacco necrosis virus* (TNV), respectivamente, ambos de la familia *Tombusviridae*, para su replicación. De cada género viral encontrado en las búsquedas con

HHpred (23 géneros y uno no asignado) y DALI (tres géneros) se seleccionó un archivo PDB representante con la mejor resolución (determinado por cristalografía de rayos-x). Esto generó una lista de 53 proteínas o dominios homólogos, cuyos modelos estructurales se volvieron a alinear con STAMP para extraer un alineamiento múltiple de secuencia y usarlo para realizar otra búsqueda en HHpred. En esta búsqueda solamente se encontró una secuencia previamente no identificada (*Astroviridae*). En total, se obtuvieron 54 proteínas o dominios homólogos (Tabla 3). Sorprendentemente, las secuencias de los virus satélite no fueron identificadas por HHpred aún después de incluir a las mismas en el alineamiento múltiple de entrada. Debido a que los virus satélite conservan la hélice αA típica de las proteínas de la cápside de los *Picornavirales* (Ban & McPherson, 1995), es posible que estas proteínas tengan un origen común y que la divergencia a nivel de secuencia haya sido tal que se ha perdido toda señal de homología en estructura primaria.

Tabla 3. Lista total de proteínas o dominios homólogos a VP1, VP2 y VP3. La lista se compone de modelos estructurales representativos de las proteínas de la cápside de cada género viral con proteínas homólogas a VP1, VP2 y VP3. Son 54 proteínas o dominios homólogos distribuidos en 28 especies virales. NA= no asignado.

Familia	Género	Virus	Proteínas	Abreviación	PDB	Resolución
Picornaviridae	Enterovirus	Coxsackievirus A24	VP1/VP2/VP3	EnteroVPx	4q4w	1.4
Picornaviridae	Aphthovirus	Foot and mouth disease virus	VP1/VP2/VP3	AphthoVPx	1qqp	1.9
Picornaviridae	Cardiovirus	Saffold virus	VP1/VP2/VP3	CardioVPx	5cfd	2.5
Picornaviridae	Senecavirus	Seneca Valley virus	VP1/VP2/VP3	SenecaVPx	3cji	2.3
Picornaviridae	Kobuvirus	Aichi virus A	VP1/VP2/VP3	KobuVPx	5aoo	2.1
Picornaviridae	Hepatovirus	Hepatitis A virus	VP1/VP2/VP3	HepatoVPx	4qpi	3.01
Picornaviridae	Parechovirus	Human parechovirus 1	VP1/VP2/VP3	ParechoVPx	5mjv	3.09
Iflaviridae	Iflavirus	Slow bee paralysis virus	VP1/VP2/VP3	IflaVPx	5j98	2.6
Dicistroviridae	Aparavirus	Israel acute paralysis virus	VP1/VP2/VP3	AparaVPx	5cdd	2.7
Dicistroviridae	Triatovirus	Triatoma virus	VP1/VP2/VP3	TriatoVPx	3nap	2.5
Dicistroviridae	Cripavirus	Cricket paralysis virus	VP1/VP2/VP3	CripaVPx	1b35	2.4
Secoviridae	Nepovirus	Grapevine fanleaf virus	CP(VP1/VP2/VP3)	NepoNTD/NepoMD/NepoCTD	2y26	2.7
Secoviridae	Comovirus	Cowpea mosaic virus	S(VP1)L(VP2/VP3)	ComoLNTD/ComoLCTD/ComoS	5fmo	2.3
Caliciviridae	Vesivirus	San Miguel sea lion virus	CP	VesiCP	2gh8	3.2
Caliciviridae	Norovirus	Norwalk virus	CP	NoroCP	1ihm	3.4
Nodaviridae	Betanodavirus	Epinephelus coioides nervous necrosis virus	CP	BetanodaCP	4rft	3.1
Nodaviridae	NA	Orsay virus	CP	OrsayCP	4nwv	3.25
Tombusviridae	Necrovirus	Tobacco necrosis virus	CP	NecroCP	1c8n	2.25
Tombusviridae	Alphacarmovirus	Carnation mottle virus	CP	AlphacarCP	1opo	3.2
Tombusviridae	Betacarmovirus	Turnip crinkle virus	CP	BetacarmCP	3zxa	3.2
Tombusviridae	Gammacarmovirus	Melon necrotic spot virus	CP	GammacarCP	2zah	2.81
Tombusviridae	Panicovirus	Panicum mosaic virus	CP	PanicoCP	4v99	2.9
Tombusviridae	Tombusvirus	Cucumber necrosis virus	CP	TombusCP	4llf	2.89
Solemoviridae	Sobemovirus	Sesbania mosaic virus	CP	SobemoCP	2wlp	2.65
Hepeviridae	Orthohepevirus	Hepatitis E virus	CP	OrthohepCP	2zzq	3.81
Astroviridae	Mamastrovirus	Human astrovirus-8	VP90	MamastVP90	5ibv	2.15
NA	Albetovirus	Satellite tobacco necrosis virus	CP	STNVCP	4bcu	2.29
NA	Papanivirus	Satellite panicum mosaic virus	CP	SPMVCP	1stm	1.9

La similitud estructural entre las CP de *Tombusviridae*, *Solemoviridae*, *Nodaviridae*, *Astroviridae*, *Hepeviridae* y *Caliciviridae*, y con las proteínas de la cápside de los *Picornavirales*, ya se ha reportado con anterioridad (Guu et al., 2009; Guo et al., 2014; Chen et al., 2015; Toh et al., 2016). Además, algunas de estas similitudes eran de esperarse debido a que familias como *Caliciviridae*, *Solemoviridae*, *Nodaviridae* y *Astroviridae* usualmente son consideradas dentro de la súper familia picorna-like (Koonin et al., 2008). Por otro lado, Krupovic y Koonin (2017), muestran que VP1 de *Picornaviridae*, *Iflaviridae* y *Dicistroviridae* se encuentra estructuralmente relacionada con las S de *Comovirus* (*Secoviridae*, del orden *Picornavirales*), CP de *Caliciviridae*, *Solemoviridae*, *Tombusviridae*, *Hepeviridae* y *Astroviridae* (También *Tymoviridae* y *Circoviridae* (ssDNA)

que no son identificadas como homólogas en este estudio). En menor medida, también muestran similitud estructural con las CP de la familia *Nodaviridae* (género *Alphanodavirus* que en este estudio no son identificadas como homólogas) y de los virus satélite, entre otros. Las CP de la familia *Nodaviridae* que se identificaron en el presente estudio son del género *Betanodavirus* y las CP de *Orsay virus* que se saben más relacionadas con las de *Betanodavirus* que con las de *Alphanodavirus* (Guo et al., 2014; Chen et al., 2015). Esto es consistente con un análisis de redes realizado a partir de comparaciones de perfiles de secuencias de proteínas de la cápaide de tipo *jelly roll* de virus de RNA en el cual las CP de *Betanodavirus* se muestran más parecidas a las CP de *Tombusviridae*, reflejando la filogenia de RdRp, mientras que las CP de *Alphanodavirus* se parecen más a las de *Alphatetraviridae*. Esto parece indicar un reemplazo del gen CP en el género *Alphanodavirus* (Wolf et al., 2018). Que algunas CP de virus de otras familias usadas en Krupovic y Koonin (2017) (e. g. *Alphatetraviridae*, *Carmotetraviridae*, *Birnaviridae* (dsRNA), *Parvoviridae* (ssDNA), etc.) no se hayan identificado como homólogas en esta búsqueda de homólogos lejanos no significa de manera definitiva que estas no compartan un ancestro común de origen viral, pues aún cabe la posibilidad de que la divergencia, tanto de estructura primaria como terciaria, sea tal que su detección sea imposible por los criterios establecidos en este estudio.

5.4 Relación evolutiva

Para saber si la señal de homología a nivel de secuencia es suficiente para resolver la relación evolutiva de estas proteínas se realizó un árbol de máxima verosimilitud con PhyML. El alineamiento de secuencia fue generado a partir de una superposición estructural, con el programa STAMP, de los 54 modelos estructurales derivados de la búsqueda de homólogos lejanos. En esta filogenia hay pocos clados con soporte de ramas mayor a 50 (figura 10). Para VP1, VP2 y VP3, se forman dos clados, uno con proteínas de la cápside de virus que infectan vertebrados (*Senecavirus*, *Cardiovirus*, *Aphthovirus* y *Enterovirus* (y *Kobuvirus* en VP3)), y otro con proteínas de la cápside de virus que infectan invertebrados (*Triatovirus*, *Cripavirus* y *Aparavirus* (e *Iflavirus* en VP1 y VP3)). Los únicos valores de soporte mayores a 80 corresponden a los clados de *Vesivirus* con *Norovirus* (*Caliciviridae*) (83), *Necrovirus* (*Tombusviridae*) con *Sobemovirus* (*Solemoviridae*) (81), y *Tombusvirus* con *Gammacarmovirus* (*Tombusviridae*) (85).

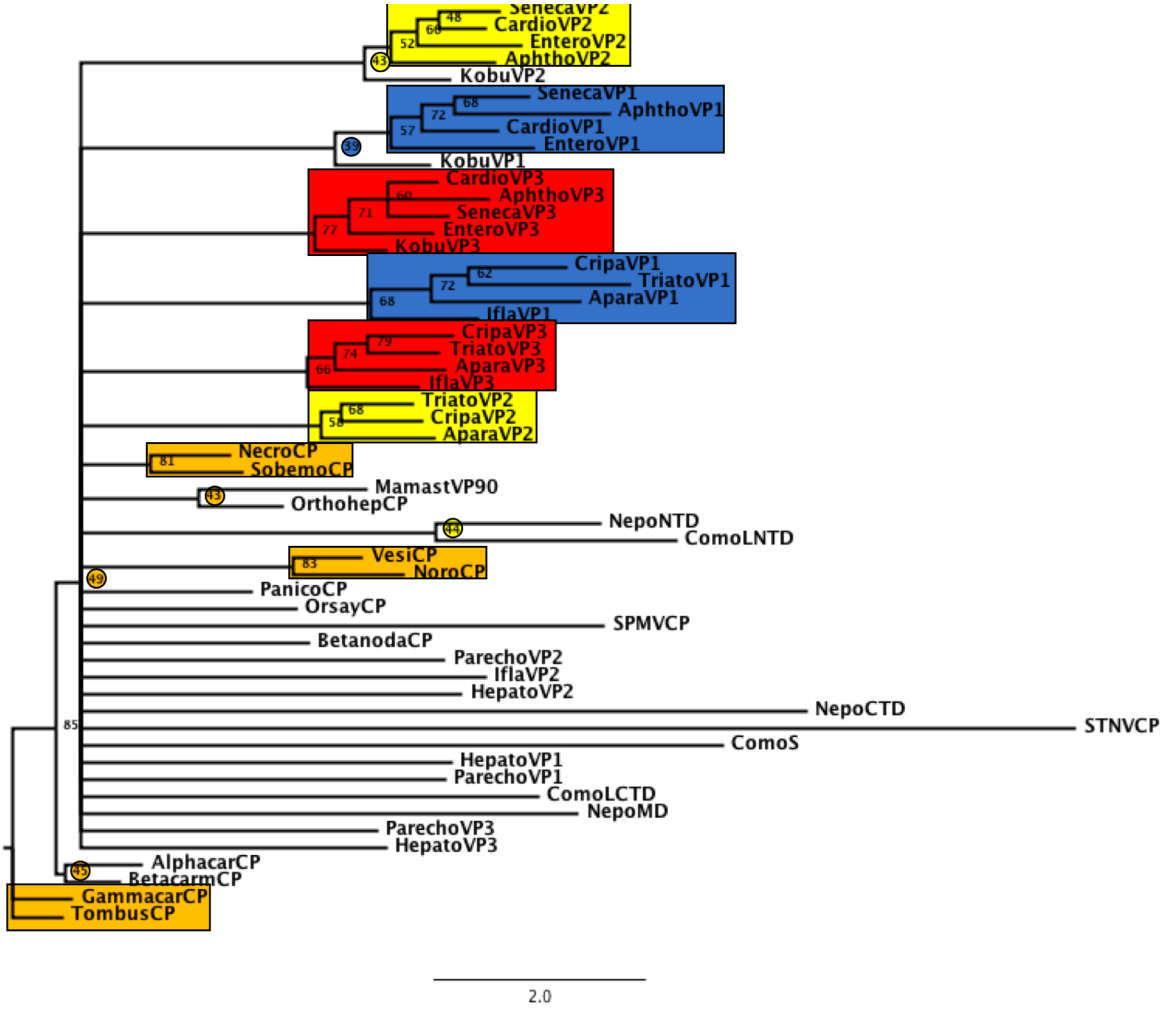


Figura 10. Filogenia de máxima verosimilitud de las secuencias de los 54 modelos estructurales de proteínas identificadas como homólogas. Clados con bootstrap mayor a 50 están resaltados con rectángulos de diferentes colores: VP1 = azul, VP2 = amarillo, VP3 = rojo, y CP = anaranjado. Algunos valores de bootstrap menores a 50 son resaltados por círculos. Todas las ramas con soporte menor a 35 están colapsadas.

Esta filogenia indica que hay suficiente señal filogenética a nivel de secuencia como para resolver la relación entre las VP1, entre las VP2 y entre las VP3 de las familias *Picornaviridae* (salvo *Hepatovirus* y *Parechovirus*) y *Dicistroviridae*. Sin embargo, no es suficiente para resolver las relaciones entre VP1, VP2 y VP3. Este resultado motiva el

análisis de la relación evolutiva de estas proteínas a través de la comparación de sus modelos estructurales.

Se realizaron dos árboles de similitud estructural, uno derivado de una matriz de distancias Q_H, generada a partir de la misma superposición estructural de la que se obtuvo el alineamiento múltiple de secuencias que se usó para construir la filogenia con PhyML, y otro derivado de una matriz de distancias SAS, generada a partir de superposiciones pareadas con el programa SSM. Ambos árboles presentan prácticamente la misma topología (Figura 11). En general, VP1, VP2, VP3 y las CP forman clados independientes. Esta topología es consistente con un origen de los *Picornavirales* a partir de un virus con cápside T=3 que sufrió dos eventos de duplicación (Liljas et al., 2002). La cercanía de las VP2 con el clado de las CP sugiere que del primer evento de duplicación surgió VP2 y una proteína ancestral que, tras un segundo evento de duplicación, habría originado a VP1 y VP3.

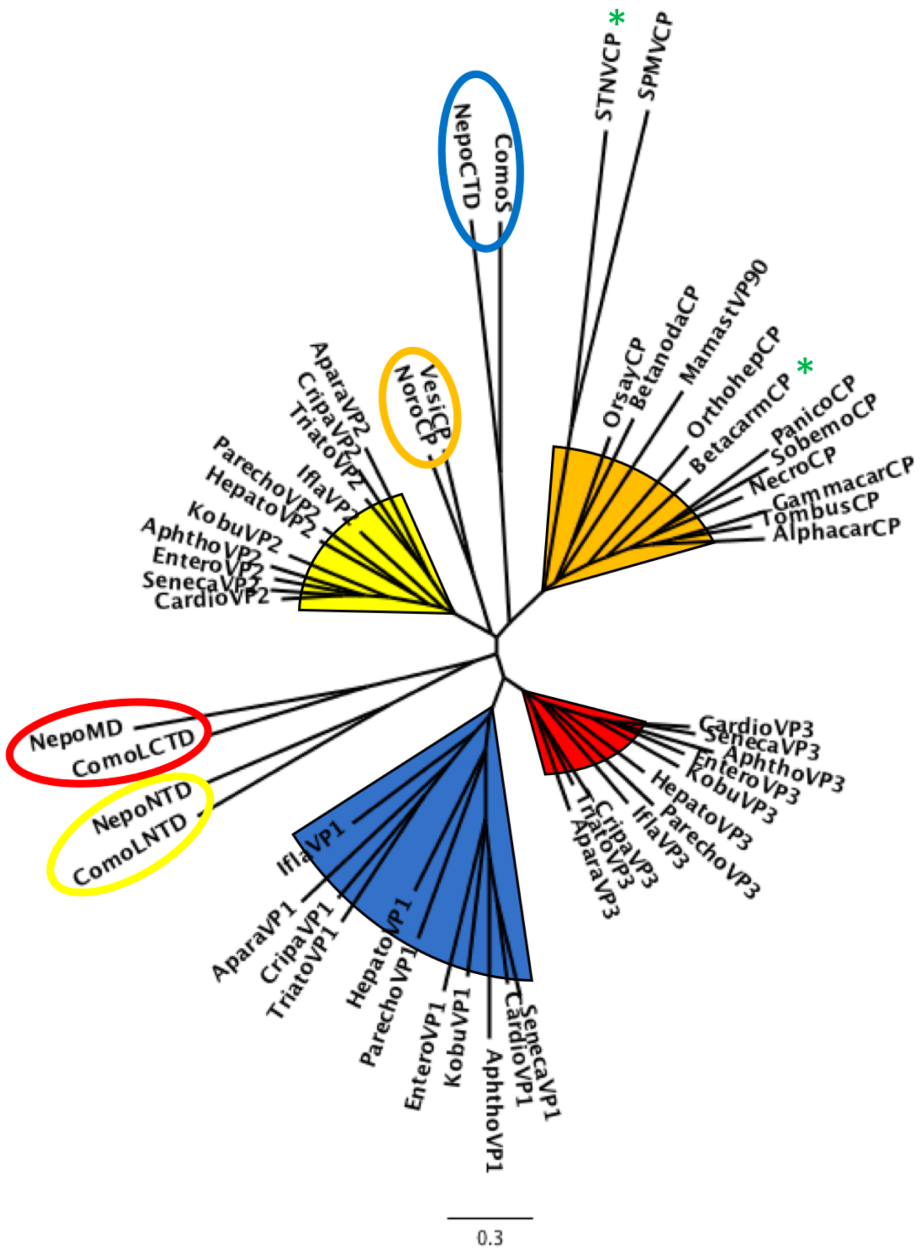


Figura 11. Dendrograma de similitud estructural evaluada a partir de las superposiciones pareadas con SSM y la matriz de distancias SAS. Seed=37, suma de cuadrados (SS) = 17.16969, porcentaje promedio de la desviación estándar (APSD) = 7.74815. Los cuatro clados principales están resaltados por fondos de diferente color: VP1 = azul, VP2 = amarillo, VP3 = rojo, y CP = anaranjado. Las ramas con ubicaciones no esperadas están resaltadas por elipses del color del clado esperado. Para árbol de distancias Q_H ver Anexo 6.

Los patrones de ramificación dentro de los clados también son similares en los dos árboles. En general, en los clados de VP1, VP2 y VP3, se forma una topología consistente con la de los árboles realizados con los protómeros completos (Spurny et al., 2017). Lo que se observa es la formación de dos subclados, uno de virus que infectan vertebrados (*Picornaviridae*) y otro de virus que infectan invertebrados (*Dicistroviridae* e *Iflaviridae*). Solamente en el árbol de distancias SAS (Figura 11), VP2 y VP3 de *Iflavirus* parecen más relacionadas, respectivamente, con las VP2 y VP3 de *Picornaviridae* que con las de *Dicistroviridae*. Esto no es así con VP1. En cambio, en el árbol de distancias Q_H (Anexo 6), VP1, VP2 y VP3 de *Iflavirus* agrupan con sus homólogos de *Dicistroviridae*. Si además tomamos en cuenta la filogenia realizada con secuencias (Figura 10), es probable que las proteínas de *Iflaviridae* realmente se encuentren más relacionadas con las de *Dicistroviridae* que con las de *Picornaviridae*.

Es importante notar que en estos árboles no solo se confirma una estrecha relación entre los *Caliciviridae* y los *Picornavirales* sino que, además, queda establecida una relación cercana entre las CP de *Caliciviridae* y las VP2 de los *Picornavirales*. Algunas filogenias realizadas con secuencias de regiones no estructurales indican que los *Caliciviridae* se encuentran particularmente relacionados con los virus de la familia *Picornaviridae* (King et al., 2012). Sin embargo, un análisis reciente sobre la evolución de los virus de RNA en el que se realizaron filogenias con las secuencias de las RdRp muestra a los virus de la familia *Caliciviridae* más cercanamente relacionados con virus que infectan invertebrados o plantas que con los virus de la familia *Picornaviridae* (Wolf et al.,

2018). Ninguna de estas relaciones se ve reflejada en los árboles de similitud estructural, pues las CP de *Norovirus* y *Vesivirus* no forman un grupo hermano a las VP2 de *Picornaviridae* o de virus que infectan invertebrados, sino que agrupan en la base del clado VP2. Dado que no parece haber un consenso sobre el origen de la familia *Caliciviridae* es posible que las proteínas de la cápisde de esta familia se hayan originado a partir de la VP2 del ancestro de los *Picornavirales*, tal y como lo sugiere la ubicación de la rama de las CP de *Vesivirus* y *Norovirus* en los dendrogramas de similitud estructural mostrados en el presente estudio (Figura 11).

Por otro lado, hay algunas ramas que dificultan más la interpretación de ambos árboles. Unas de estas son las que corresponden a los dominios equivalentes a VP1, VP2 y VP3 (A, C y B, respectivamente) de los *Comovirus* y *Nepovirus* (*Secovirudae*). Estas no agrupan en los clados esperados, y la longitud y ubicación de sus ramas sugieren que no se parecen lo suficiente a las proteínas de ninguno de los cuatro clados (Anexo 9 y 10). En el árbol de distancias Q_H, el clado de VP1 se ve interrumpido por las ramas de los dominios A y B de *Secoviridae*, y por las CP de los virus satélite. En el caso de los dominios A de *Secoviridae*, es interesante notar que estos aparentan estar relacionados con la VP1 de *Aparavirus* (*Dicistroviridae*), pues hay estudios que implican que los virus de la familia *Secoviridae* se originaron a partir de virus que infectan invertebrados (Gorbalenya et al., 2002; Koonin et al., 2008; King et al., 2012; Thompson et al., 2014). Sin embargo, la asociación de estas ramas separa al subclado de los virus que infectan invertebrados, el cual tiene cierto soporte incluso a nivel de secuencia (Figura 10). Por lo tanto, es posible que la

inserción de estas ramas en el clado de VP1 se deba a atracción de ramas largas. De hecho, al colapsar todas las ramas cuya longitud es menor al promedio (0.27), se puede observar que las ramas del clado de VP1 son las únicas que no colapsan (Anexo 7). Al retirar estas ramas largas del árbol, el subclado de los virus que infectan invertebrados se vuelve a formar (Anexo 8). La longitud de las ramas dentro de este clado también nos habla de la divergencia estructural de las VP1 que, de acuerdo con Liljas et al. (2002), se encuentran menos conservadas, tanto a nivel de estructura primaria como a nivel de estructura terciaria, que las VP2 y las VP3. Las ramas de las proteínas de la cápside de *Secoviridae* y de virus satélite no generan este problema dentro del clado de VP1 en el árbol de distancias SAS. En cuanto a la matriz de distancias SAS, solamente los dominios B de los *Secoviridae*, y las CP de los virus satélite tienden a parecerse más a sus homólogos VP3 y CP, respectivamente (Anexo 10). Esta tendencia se ve reflejada en el árbol de distancias SAS en el caso de las CP de virus satélite, que agrupan en la base del clado CP (Figura 11), pero no en el caso de los dominios B de los *Secoviridae*. Esto podría deberse a las limitaciones propias de los métodos de distancia para construir árboles, que no logran ajustar a la perfección la longitud de las ramas a partir de las distancias observadas en la matriz.

Los árboles de similitud estructural no permiten poner a prueba si las cápsides de *Secoviridae* realmente conservan un estado ancestral de la transición de cápsides T=3 a T=p3 (Chandrasekar & Johnson 1997). Mas, si estos virus efectivamente se originaron a partir de virus que infectan invertebrados (Thompson et al., 2014), es posible que sus cápsides representen un caracter derivado que involucra una reducción en la longitud de los

extremos N-terminales (conectores entre dominios en *Secoviridae*) y la falta de procesamiento de la poliproteína estructural por parte de las proteasas virales. Evidencias de esto incluyen que, tanto los virus del género *Waikavirus* y *Sequivirus*, que tienen un genoma monopartito y cuyas proteínas de la cápside son completamente procesadas por la proteasa viral como en el resto de los *Picornavirales*, como los del género *Torradovirus*, que tienen un genoma bipartito como los *Nepovirus* y los *Comovirus*, pero que sus proteínas de la cápside también son completamente procesadas por la proteasa viral, presentan proteínas de la cápside y proteasas con una mayor similitud a nivel de secuencia con virus que infectan animales que con virus que infectan plantas (Thompson et al., 2014). Además, en las filogenias realizadas con secuencias de regiones no estructurales, los *Secoviridae* parecen particularmente cercanos a los virus de la familia *Dicistroviridae* (Gorbalenya et al., 2002; Koonin et al., 2008; King et al., 2012). La posibilidad de que la familia *Secoviridae* se haya originado a partir de la familia *Dicistroviridae*, o de un ancestro, adquiere aún mayor plausibilidad con el ejemplo conocido de un *Cripavirus* que puede transmitirse de un invertebrado a otro a través de plantas (Thompson et al., 2014). Es de notarse que las cápsides de virus que infectan plantas tienden a tener inserciones menos elaboradas (Rossmann & Johnson, 1989). La adaptación a ambientes cambiantes, como cambiar de hospedero o la presión del sistema inmune de los hospederos vertebrados, puede favorecer la diversidad de las inserciones, pero, en los *Secoviridae*, el hecho de que son transmitidos por vectores invertebrados podría explicar la alta presión de selección negativa a la que se encuentran sometidos (Thompson et al., 2014). Analizando la conservación de caracteres estructurales, hay una inserción entre la hebra βC y la hélice αA, primero

identificada en las VP3 de *Iflaviridae* y *Dicistroviridae* (Liljas et al., 2002), que también está presente en todas las VP1 de estas familias. Este podría ser un caracter presente en el ancestro de VP1 y VP3. Este caracter aún está presente en los dominios A y B de los *Comovirus* y *Nepovirus* (Figura 12). En el dominio B de *Nepovirus* se ha propuesto que esta inserción ayuda a formar un sitio de unión a ligando importante en la transmisión del virus vía vector (Schellenberger et al., 2011). Es posible que en virus que infectan invertebrados esta inserción tenga un papel en la unión al receptor celular.

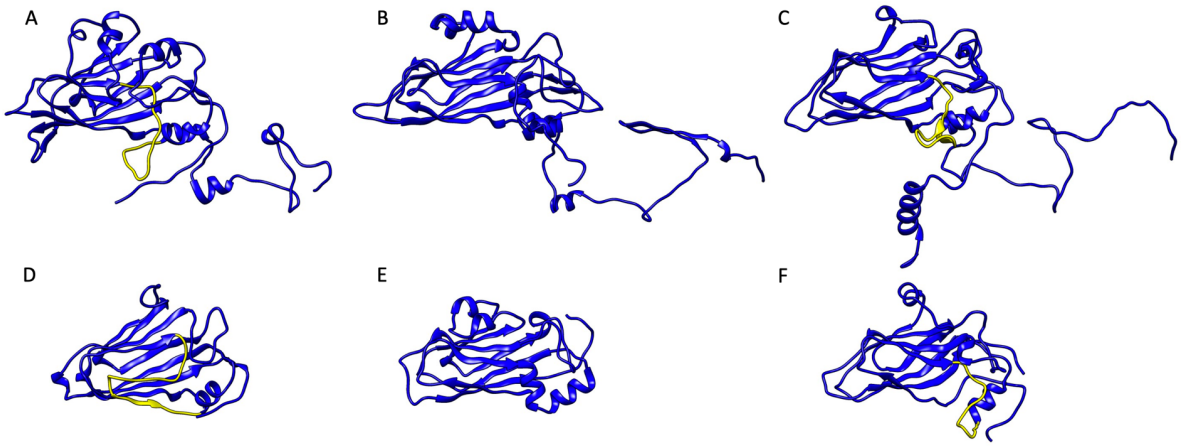


Figura 12. Conservación de caracteres estructurales de virus que infectan invertebrados en cápsides de virus de la familia *Secoviridae*. A, B y C son VP1, VP2 y VP3 de *Cripavirus*, respectivamente. D, E y F son los dominios A, C y B de *Nepovirus*, respectivamente. Las inserciones βC-αA están resaltadas en amarillo.

Debido a la conservación de otros caracteres estructurales (Figura 13), cabe la posibilidad de que las cápsides de los virus de la familia *Picornaviridae* también hayan derivado de cápsides de virus que infectan a invertebrados (Wang et al., 2015). En cuanto a VP1, se puede observar que los virus que infectan invertebrados, además de la inserción βC-αA, tienen una inserción entre las hebras βE y βF de VP1 que genera una hebra β

adicional al formar puentes de hidrógeno con la hebra βC, referida aquí como puente CHEF. Interesantemente, la VP1 de *Hepatovirus*, presenta tanto el puente CHEF como la inserción βC-αA, pero no tiene el *loop* FMDV típico de otros virus de la familia *Picornaviridae*. La VP1 de *Parechovirus* tampoco tiene el *loop* FMDV pero, a diferencia de la VP1 de *Hepatovirus*, no presenta las inserciones típicas de virus que infectan a invertebrados. Estos caracteres conservados en *Hepatovirus* se pudieron haber perdido en las VP1 del resto de la familia *Picornaviridae* (con excepción de *Senecavirus* y *Cardiovirus* que parecen conservar la inserción βC-αA, asociada a otro par de inserciones de la hebra βC), en los cuales evolucionó el *loop* FMDV (salvo en *Parechovirus*). En cuanto a la VP3 de *Hepatovirus*, también se ve conservada la inserción βC-αA, presente en las VP3 de *Ifalviridae* y *Dicistroviridae*, y la ausencia de la inserción característica del resto de los *Picornaviridae*, en este caso el *knob* en βB. En el caso de las VP2 de la familia *Picornaviridae*, estas se caracterizan por la presencia de una inserción grande entre las hebras βE y βF llamada *puff* (salvo en *Aphthovirus* y *Kobuvirus*. Las proteínas de la cápside de *Kobuvirus* tienen características únicas. Quizás, la más relevante es la ausencia de *puff* en VP2, lo cual parece estar compensado por una inserción al final de βC (diferente a la inserción βC-αA) que ocupa un volumen equivalente al del *puff* (Sabin, et al., 2016)), y por un extremo N-terminal retraído que interactúa con su propio barril y forma dos hebras βA_1-βA_2 que extienden la hoja CHEF de una VP3 en un pentámero diferente. En las VP2 de la familia *Dicistroviridae*, además de que carecen del *puff*, el extremo N-terminal se encuentra extendido sobre la VP2 de otro protómero. Esta extensión provoca que el par de hebras βA_1-βA_2 del extremo N-terminal de VP2 interactúe con la hoja CHEF de una VP3 del

mismo pentámero (Liljas et al., 2002). La estructura de los extremos N-terminales es idéntica, y la diferencia entre un extremo retraído y uno extendido es un desplazamiento como de limpiaparabrisas de 180 grados. Por ende, las interacciones que se observan en ambas conformaciones son prácticamente las mismas, pero la conectividad entre subunidades cambia. A este fenómeno se le ha llamado intercambio de dominios. (Liljas et al., 2002; Wang et al., 2015). Las VP2 de *Hepatovirus* y *Parechovirus* son idénticas a las de *Dicistroviridae* (Wang et al., 2015; Kalynych et al., 2016)

.

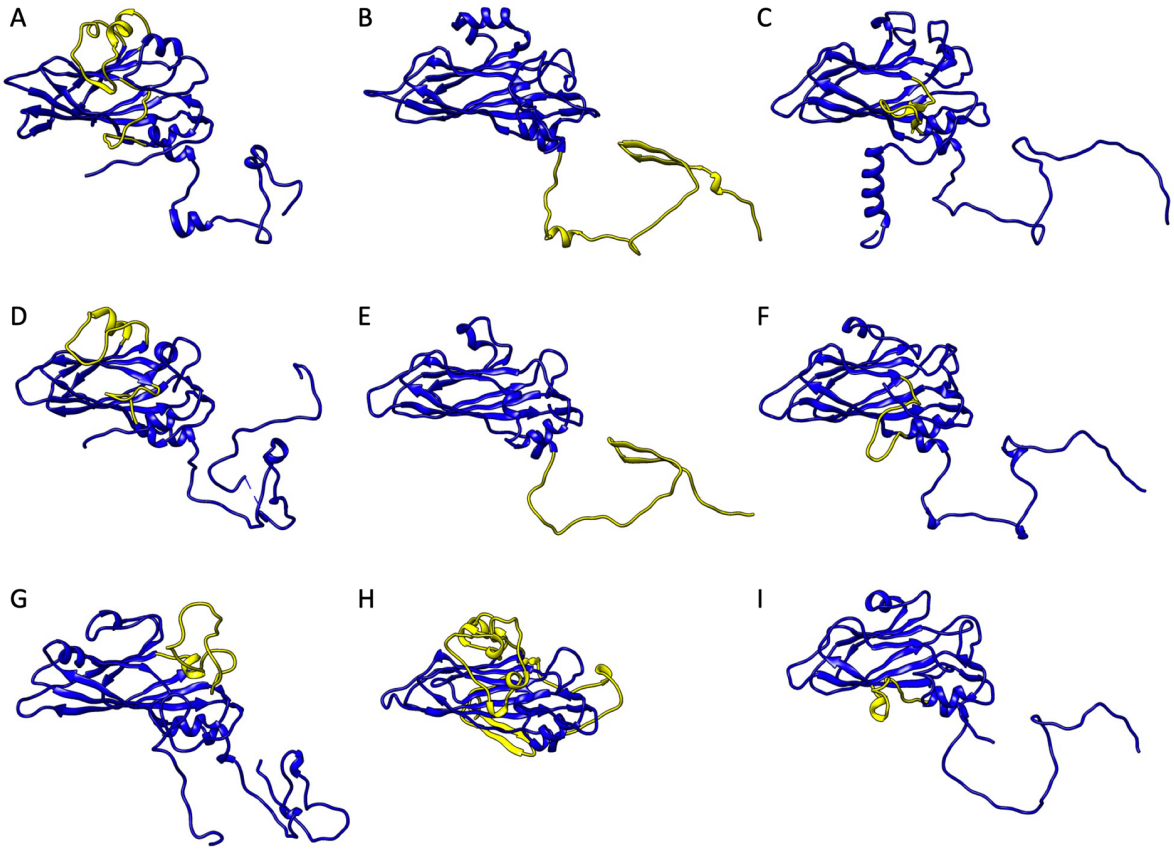


Figura 13. Caracteres típicos de virus que infectan invertebrados conservados en las proteínas de la cápside de *Hepatovirus*, comparados con los caracteres típicos de la familia *Picornaviridae*. A, B y C son VP1, VP2 y VP3 de *Cripavirus*, respectivamente. D, E y F son VP1, VP2 y VP3 de

Hepatovirus, respectivamente. G, H e I son VP1, VP2 y VP3 de *Enterovirus*, respectivamente. El puente CHEF de VP1, la inserción βC-αA de VP1 y VP3, y el extremo N-terminal extendido de *Cripavirus* y *Hepatovirus*, como el *loop* FMDV de VP1, el *puff* y el extremo N-terminal retraído de VP2 y el *knob* de VP3 de *Enterovirus*, están resaltados en amarillo.

Es interesante notar que en *Iflavirus* el extremo N-terminal se encuentra retraído como en *Enterovirus* (Kalynych et al., 2016). Esto podría explicar por qué en el árbol de distancias SAS la VP2 de *Iflavirus* parece más cercana a las VP2 de *Picornaviridae* (Figura 11). Sin embargo, una razón más para pensar que las cápsides de *Iflaviridae* se encuentran más relacionadas con las de *Dicistroviridae*, es la conservación de la organización de los genes estructurales en estas familias. Esta organización implica que el péptido VP4, que en *Picornaviridae* está codificado como una extensión N-terminal de VP2, en *Dicistroviridae* e *Iflaviridae* se encuentra codificado como una extensión N-terminal de VP3 (Liljas et al., 2002; Kalynych et al., 2016). En la familia *Caliciviridae* también hay diferencias en la orientación del extremo N-terminal. En el género *Norovirus*, el extremo N-terminal se encuentra retraído (Prasad et al., 1999) como en VP2 de *Iflaviridae* y algunos *Picornaviridae*, mientras que en el género *Vesivirus* el extremo N-terminal se encuentra extendido (Chen et al., 2006) como en VP2 de *Dicistrovirdae* y los géneros *Hepatovirus* y *Parechovirus* de la familia *Picornaviridae*. Debido a que VP1 y VP3 siempre presentan su extremo N-terminal extendido, es probable que el estado ancestral del extremo N-terminal de VP2 sea extendido, y que el estado retraído haya evolucionado de manera independiente en *Iflaviridae*, *Picornaviridae* y *Caliciviridae*, lo cual solo habría requerido de una rotación de los ángulos dihedros en la posición equivalente al residuo 53 de *Hepatovirus* (Wang et

al., 2015). Los ángulos de las VP2 extendidas (*Cripavirus*, *Hepatovirus*, *Parechovirus* y *Vesivirus*) tienden a asemejarse a los de las hélices α (en promedio phi = -63 y psi = -43), mientras que los ángulos de las VP2 retraídas (*Enterovirus*, *Aphthovirus*, *Iflavirus* y *Norovirus*) tienden a asemejarse a los de las hebras β (en promedio phi = -97 y psi = 65) (Anexo 11). Esto tiene que ver con que los extremos retraídos aumentan el número de contactos con la hebra I de la misma subunidad, generando una extensión de βB (Prasad et al., 1999). En virus T=3, el extremo N-terminal es rico en aminoácidos cargados positivamente, pero este extremo solo se encuentra ordenado en las subunidades C, que ocupan una posición equivalente a VP2, pero no en las subunidades A y B, que ocupan una posición equivalente a VP1 y VP3, respectivamente. Los dímeros C/C forman una superficie plana, mientras que los dímeros A/B forman una superficie curva. Este juego de superficies permite que se formen cápsides T=3 en lugar de cápsides más curvas y pequeñas T=1. Adicionalmente, es posible que, en su interior, la superficie plana permita un mejor alojamiento del RNA (Rossmann & Johnson, 1989; Prasad et al., 1999; Xing et al., 2010; Chen et al., 2015). En la familia *Caliciviridae* el extremo N-terminal tiene la misma conformación y lleva a cabo el mismo tipo de interacciones en las tres subunidades A, B y C, de manera que el juego de superficies es llevado a cabo por el dominio protruyente (P) (Prasad et al., 1999; Chen et al., 2006). En los *Picornavirales* las diferentes conformaciones del extremo N-terminal podrían tener implicaciones en el ensamblaje y estabilidad de la cápside, pues en cápsides con N-terminal de VP2 retraída la unión entre pentámeros parece estar estabilizada por interacciones en el eje triple, mientras que en cápsides con N-terminal de VP2 extendida la unión entre pentámeros parece estar estabilizada por interacciones en

el eje doble (Tate et al., 1999; Liljas et al., 2002; Kalynych et al., 2016). También es posible que tenga implicaciones en el mecanismo de liberación del genoma, que en *Enterovirus* involucra la salida del extremo N-terminal de VP1 a través del eje q3 (Strauss et al., 2015). Además, en *Hepatovirus*, las hélices αA de VP2, que interactúan en el eje doble, se encuentran más cercanas entre sí que en *Enterovirus*, lo cual sugiere un mecanismo de liberación del genoma completamente diferente (Wang et al., 2015).

En cuanto al clado CP, la topología es consistente con un árbol realizado con las secuencias de las CP (King et al., 2012). Lo que se observa es a las CP de *Betanodavirus* y de *Orthohepevirus* relacionadas con un subclado de las CP de *Tombusviridae*. Este subclado además se divide en dos grupos, uno de CP con dominio P que incluye a los géneros *Tombusvirus*, *Alphacarmovirus*, *Betacarmovirus* y *Gammacarmovirus*, y otro de CP sin dominio P, que incluye a los géneros *Panicovirus* y *Necrovirus*, relacionado con los *Sobemovirus*. Solamente en el árbol de distancias SAS, la CP de *Betacarmovirus* no agrupa en el clado de las CP de la familia *Tombusviridae* con domino P. Esto podría deberse a que su modelo estructural no cuenta con asignación de estructura secundaria y por lo tanto los valores de RMSD y Naln se tuvieron que calcular con herramientas diferentes a SSM. En el árbol de distancias Q_H sí se obtiene la agrupación esperada.

Estas CP no tienen inserciones tan elaboradas como en los *Picornavirales*, pero varias presentan uno (*Tombusviridae* y *Nodaviridae*) o dos (*Hepeviridae*, *Astroviridae* y *Caliciviridae*) dominios P en el extremo C-terminal. Hay dos tipos de dominios P, unos son

barriles β de tipo *jelly roll* compuestos de 10 hebras β antiparalelas, presentes en virus de la familia *Tombusviridae*, y otros parecen barriles β torcidos compuestos de seis a ocho hebras β antiparalelas presentes en virus de las familias *Astroviridae*, *Hepeviridae*, *Nodaviridae*, *Caliciviridae* y en VP3 de *Iflaviridae*. Estos barriles tienen un plegamiento que parece derivar del meta-plegamiento *cradle-loop* con topología de monómero RIFT. Este plegamiento se caracteriza por la presencia de un motivo ββαβ duplicado y se encuentra ampliamente distribuido en proteínas antiguas como la proteína ribosomal L3, el dominio N-terminal de la ATPasa tipo-F o el factor de elongación EF-Tu (Alva et al., 2008). Algunas diferencias de los dominios P con respecto a la topología de monómero RIFT es la ausencia de una o ambas hélices α y, en algunos casos, inserciones de hebras β cerca de los *loops* β1-β2/β1'-β2'. La relación con este tipo plegamiento parece confirmarse en el caso del dominio P1 de *Hepeviridae*, que se encontró como homólogo por comparación de estructuras terciarias con el dominio de unión al tRNA del factor de elongación EF-Tu. Sin embargo, la similitud más interesante con una proteína celular es con el dominio barril β de la endo-alfa-sialidasa, lo cual sugiere un mecanismo de unión a las células a través de ácido siálico para este dominio P (Guu et al., 2009). Una hipótesis particularmente interesante sobre la evolución de los dominios P es que los dominios P1 y P2 de *Hepeviridae* son duplicados (Guu et al., 2009). Para poner a prueba esta hipótesis se realizaron búsquedas de DALI partiendo con cada uno de los dominios P de *Astroviridae*, *Hepeviridae*, *Caliciviridae*, *Nodaviridae* y VP3 de *Iflaviridae*. Los valores Z resultantes se ordenaron de manera manual en una red de similitud estructural que confirma un origen común de estos dominios P relacionados con el meta-plegamiento *cradle-loop*, así como la hipótesis de la

duplicación de los dominios P1 y P2 (Anexo 12). Aunque en estas búsquedas no se encontró ninguna similitud con el dominio P2 de *Astroviridae*, posiblemente debido a sus extensas inserciones, se ha reportado que el alineamiento de este con el dominio P2 de *Hepeviridae* genera un valor Z de 3.9 (Dong et al., 2011). Por otro lado, y de manera similar, se ha sugerido que el domino S y P de *Tombusviridae*, ambos de tipo *jelly roll*, son dominios duplicados (Jones et al., 1989). Sin embargo, una búsqueda de DALI con el dominio P de *Tombusviridae* muestra que las proteínas con mayor similitud estructural pertenecen a la familia del factor de necrosis tumoral (TNF) (Anexo 13). Esto descarta la posibilidad de que los dominios S y P de *Tombusviridae* sean dominios duplicados, y sugiere que este dominio P de tipo *jelly roll* se adquirió de manera independiente a partir de un hospedero.

La estrecha relación entre las CP de *Tombusivirdae*, *Solemoviridae*, *Nodaviridae*, *Astroviridae* y *Hepeviridae* también se observa en el análisis de redes realizado por Wolf et al. (2018). La relación entre las CP de *Hepeviridae*, *Nodaviridae* y *Tombusviridae* parece reflejar la filogenia de RdRp presentada en el estudio citado. Sin embargo, la estrecha relación entre las CP de *Astroviridae* y *Hepeviridae*, o las de *Solemoviridae* y *Tombusviridae*, no coincide con las relaciones sugeridas por dicha filogenia, pues las polimerasas de *Astroviridae* y *Solemoviridae* se muestran en un clado que incluye a las proteínas de la cápside de los *Picornavirales* y de *Caliciviridae*. Esto sugiere eventos de reemplazo por HGT entre las CP de *Astroviridae* y *Hepeviridae*, y entre las de CP *Solemoviridae* y *Tombusviridae*. En cuanto a las CP del virus de Orsay y los *Betanodavirus*,

que solo tienen un dominio P, se ha sugerido que sus dominios S se originaron a partir de un ancestro sin domino P similar al de los *Solemoviridae* y algunos *Tombusviridae*. Además, debido a la similitud estructural entre el dominio P del virus de Orsay y el dominio P1 de *Hepeviridae* y *Caliciviridae*, se ha sugerido que las últimas familias adquirieron su CP a partir de un ancestro con una CP de un solo dominio P similar a la del virus de Orsay (Guo et al., 2014). Aunque se ha sugerido que las CP de *Caliciviridae* y *Hepeviridae*, que tienen dos dominios P, se encuentran estrechamente relacionadas, se ha mostrado que sus dominios S no son tan parecidos. Además, la organización de los dominios P es distinta (S-P1-P2 en *Hepeviridae* y S-P1-P2-P1 en *Caliciviridae*) debido a que en *Caliciviridae* P2 parece ser una inserción en P1 (Guu et al., 2009). Por otro lado, la estrecha relación entre las CP de *Astroviridae* y *Hepeviridae* también es sugerida por la similitud estructural de sus dominios P (Dong et al., 2011; Toh et al., 2016). Conciliando esta información con la filogenia de RdRp (Wolf et al. 2018) es posible que el ancestro de *Tombusviridae*, *Nodaviridae* y *Hepeviridae* tuviese una CP sin dominio P. La cápside de *Tombusviridae* sin dominio P pudo haber reemplazado a la cápside original de *Solemoviridae*. Por otro lado, en el ancestro de *Nodaviridae* y *Hepeviridae* pudo haber surgido el dominio P relacionado con el meta-plegamiento *cradle-loop*, el cual se habría duplicado en la familia *Hepeviridae*. Los dominios S, P1 y P2 de *Hepeviridae* pudieron haber reemplazado por completo a la cápisde de *Astroviridae*, y en el caso de *Caliciviridae*, su dominio S se pudo haber originado a partir de VP2 del ancestro de los *Picornavirales*, mientras que sus dominios P1 y P2 se pudieron haber adquirido de manera horizontal a partir de *Hepeviridae* o *Astroviridae*, que también infectan vertebrados. Por último, el

domino P de VP3 de *Iflaviridae*, tiene similitud estructural con los dominios P de *Nodaviridae*, *Hepeviridae*, *Astroviridae* y *Caliciviridae*, pero, contrario a lo que proponen Kalynych et al. (2016), los árboles de similitud estructural sugieren que este fue adquirido de manera horizontal, posiblemente a partir del domino P de un virus similar al virus de Orsay, que también infecta invertebrados.

En síntesis, los arboles de similitud estructural sugieren que hubo dos eventos de duplicación en la evolución de las proteínas de la cápside de virus del orden *Picornavirales*. Estas duplicaciones habrían ocurrido a partir de un ancestro con cápside T=3, dando lugar a un ancestro con cápside T=p3 que divergió en las diferentes familias del orden. Hay ciertas características comunes a todos los virus del orden *Picornavirales*. Por ejemplo, todos codifican para una posible helicasa, una proteasa y una RdRp (Koonin et al., 2008). Además, todos llevan una proteína Vpg en el extremo 5', necesaria para el inicio de la replicación, y una cola de poli(A) en el extremo 3' (King et al., 2012). Es posible que estas características hayan estado presentes en el ancestro de los *Picornavirales* con cápside T=p3 e incluso en el ancestro T=3. La distribución de los caracteres estructurales conservados sugiere que el ancestro T=p3 de los *Picornavirales* era un virus que infectaba invertebrados. Es probable que este ancestro presentara el extremo N-terminal de VP2 extendido, la extensión βC-αA de VP1 y VP3, y el puente CHEF de VP1. La retracción del extremo N-terminal de VP2 es un carácter que pudo haber convergido en *Iflavirus*, *Norovirus* y en varios géneros de la familia *Picornaviridae*. En cuanto a la proteína VP4, no está claro qué características pudieron estar presentes en el ancestro, pues en *Iflaviridae* y

Dicistroviridae es una extensión N-terminal de VP3, y en *Picornaviridae* es una extensión N-terminal de VP2. Además, aunque aparentemente cumplen funciones similares, sus estructuras no son parecidas, lo cual sugiere que surgieron de manera independiente en el linaje de virus que infectan vertebrados y en el linaje de virus que infectan invertebrados (Liljas et al., 2002). Tampoco está claro cuál pudo haber sido la organización de su genoma. En virus con cápside T=3, las proteínas no estructurales están codificadas en el extremo 5', mientras que las proteínas estructurales son codificadas en el extremo 3', salvo en el caso de los nodavirus que codifican sus proteínas en dos segmentos de RNA diferentes (King et al., 2012). Sin embargo, para Llijas et al. (2002), es más probable que su genoma se haya parecido al de la familia *Picornaviridae* o *Iflaviridae*. En estas familias, las proteínas estructurales son codificadas en el extremo 5' y las proteínas no estructurales en el extremo 3', mientras que en la familia *Dicistroviridae* parece haber ocurrido un barajeo del genoma que colocó a las proteínas no estructurales en el extremo 5' y a las proteínas estructurales en el extremo 3'. La conservación de caracteres estructurales, en conjunto con otros estudios evolutivos, sugiere que las cápsides de los virus de la familia *Secoviridae* no presentan caracteres ancestrales de la transición de las cápsides T=3 a T=p3, sino que son derivados a partir de un ancestro cercano a la familia *Dicistroviridae*. En cuanto a las CP de la familia *Caliciviridae*, es posible que estas se hayan originado a partir de VP2 del ancestro del orden *Picornavirales* T=p3 (Figura 14). Estos resultados, en conjunto con el análisis de dominios P, terminan por revelar una historia evolutiva compleja que involucra varios eventos de duplicación, HGT y convergencias funcionales en un grupo de virus que han logrado adaptarse a una gran diversidad de hospederos (Anexo 14). Por último, la determinación de

las estructuras terciarias de las proteínas de la cápside de otros virus relacionados, así como la caracterización de los mecanismos de unión al receptor celular y la formación de superficies antigénicas de las estructuras conocidas, podría ayudar a resolver mejor las relaciones evolutivas entre estas proteínas de la cápside.

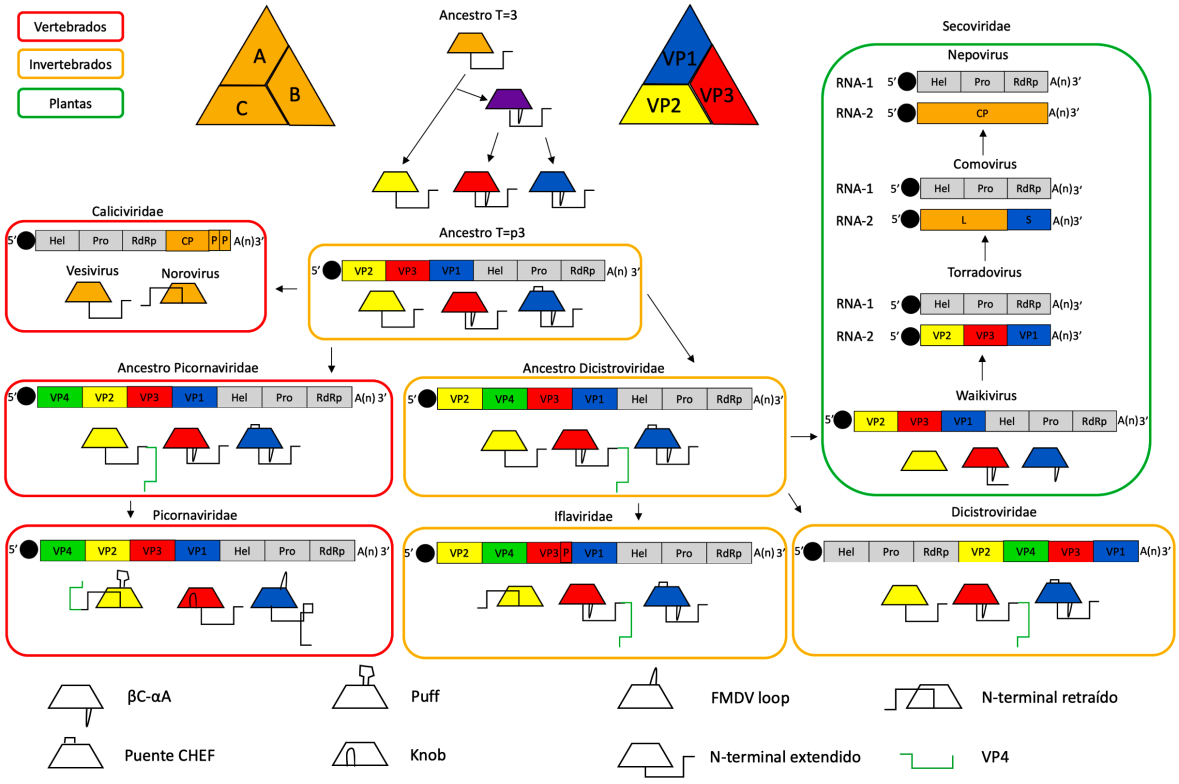


Figura 14. Esquema que sintetiza la hipótesis evolutiva de las cápsides de virus del orden *Picornavirales*. Las proteínas de la cápside son representadas como trapecios sobre los cuales se muestran caricaturas de las diferentes inserciones. Además, se muestra una simplificación de la organización de los genomas mostrando las características conservadas en el orden *Picornavirales* (Vpg (círculo negro), proteínas de la cápside, Hel, Pro, RdRp, cola de poli(A) 3' y en algunos casos dominio P). Los rectángulos que encierran a las representaciones de los genomas y de las proteínas de la cápside indican el tipo de hospedero al que infectan estos virus.

Se pueden vislumbrar dos posibles ventajas de tener duplicadas las proteínas de la cápside. En primer lugar, tener tres proteínas de la cápside en lugar de una podría acelerar el proceso de ensamblaje del virión, pues en lugar de expresar la misma proteína 180 veces, estos virus solo necesitan pasar por 60 rondas de traducción. En segundo lugar, la duplicación de estas proteínas de la cápside parece haber favorecido la especialización de cada subunidad dentro de su ambiente estructural particular. Esto es evidente, ya que en cápsides T=3 cada subunidad puede sufrir ligeros cambios conformacionales dependientes de su entorno estructural para mantener el ensamblaje correcto (Chen et al., 2006). En cápsides T=p3, esta especialización también permite que algunos virus se mantengan en un proceso de constante adaptación a los hospederos con sistemas inmunes más sofisticados. En virus con cápsides T=3 que infectan vertebrados es posible que estas interacciones con el hospedero sean mediadas por los dominios P (Guo et al., 2014). Esto sería consistente con Gao et al. (2017), que proponen que algunas duplicaciones pueden ser adaptativas en términos de una 'carrera armamentista' entre el virus y el sistema inmune del hospedero. Como ejemplo, mencionan el caso de un factor anti hospedero K3L en la familia *Poxviridae* (dsDNA) y el caso de Vpr y Vpx en la familia *Retroviridae*. En el caso de *Retroviridae*, la duplicación mencionada les permite inhibir un mecanismo de defensa intracelular SAMHD1 que bloquea la transcripción reversa.

El análisis de la historia evolutiva de estas proteínas de la cápside nos da indicios sobre los procesos evolutivos que pueden ocurrir en genomas de RNA independientemente de su procedencia (virus de RNA o células del mundo de RNA/proteínas). Estos incluyen

duplicaciones, HGT y convergencias funcionales que permiten un rápido ensamblaje de complejos moleculares, la especialización de proteínas parálogas y una constante adaptación a ambientes cambiantes. En este caso es importante destacar el papel que podrían jugar las helicasas en la apertura del genoma para su replicación, que en los virus estudiados podrían explicar la estabilidad de las diferentes duplicaciones o dominios adquiridos horizontalmente. Tanto los *Picornavirales* como los virus de la familia *Caliciviridae* y *Hepeviridae* tienen secuencias con posible función de helicasa. En estos virus, la energía necesaria para abrir las dobles hélices para la replicación, que incrementa con el tamaño del genoma (Reanney, 1982), no sería un problema mayor. Sin embargo, sus genomas, de entre 7 y 10 kbp, no son especialmente grandes (King et al., 2012). Si bien las helicasas podrían explicar la estabilidad de algunos genes duplicados, es posible que la capacidad de corrección sea más importante para el crecimiento de los genomas. Esto se ve ejemplificado en virus de dsDNA que, entre más grandes, más frecuente es la presencia de polimerasas replicativas y de reparación (Kazalauskas et al., 2016). Un sondeo sobre la presencia de este tipo de enzimas en virus de RNA, así como un análisis sobre la evolución de los virus de RNA con genomas grandes, podría revelar aspectos adicionales sobre el origen de genes en genomas de RNA y por lo tanto sobre los mecanismos que guiaron el crecimiento de este tipo de genomas.

6 Conclusiones

A pesar de la baja conservación de las secuencias de VP1, VP2 y VP3 hay suficiente señal que permite identificarlas como homólogos lejanos con búsquedas iterativas. Sin embargo, esta señal de homología no es suficiente para resolver su relación filogenética. Los únicos residuos conservados en las tres proteínas son dos prolinas ubicadas al final de las hebras βE y βG. Es posible que estos residuos hayan estado presentes en la proteína ancestral antes de los eventos de duplicación (ancestro con cápside T=3). Otros residuos característicos de VP1, VP2 o VP3 podrían haber estado presentes en las respectivas proteínas ancestrales, despúes de los eventos de duplicación pero antes de la divergencia de los *Picornavirales* (ancestro con cápside T=p3). Los residuos más conservados parecen estabilizar las estructuras de las subunidades y los contactos dentro y entre protómeros, mientras que las regiones más variables parecen tener implicaciones en la unión al receptor celular y la formación de superficies antigénicas. Esto sugiere que a lo largo de la evolución de estas proteínas se han conservado mejor las características asociadas con la estabilidad de la cápside que aquellas involucradas en la interacción con el hospedero. Los dendrogramas de similitud estructural apoyan una hipótesis sobre el origen de la cápside T=p3 de los *Picornavirales* a partir de un cápside T=3 después de dos eventos de duplicación. Del primer evento de duplicación habría surgido VP2 y una proteína ancestral que, tras un segundo evento de duplicación, habría dado origen a VP1 y VP3. Las CP de la familia *Caliciviridae* presentan mayor similitud de estructura primaria y terciaria con VP2 que con VP1, VP3 u otras CP. Es posible que estas CP se hayan originado a partir de VP2

de un ancestro T=p3 del orden *Picornavirales*. La conservación de caracteres estructurales típicos de VP1, VP2 y VP3 de virus que infectan invertebrados, como la inserción βC-αA en VP1 y VP3, el puente CHEF en VP1 y el extremo N-terminal extendido en VP2, en las proteínas de la cápside de *Hepatovirus* sugiere que el ancestro de los *Picornavirales* era un virus que infectaba invertebrados. En la familia *Picornaviridae*, estos caracteres se habrían perdido y, conforme estos virus se adaptaron a sus hospederos vertebrados, habrían surgido otros caracteres como el *loop* FMDV en VP1, el *puff* y el extremo N-terminal retraído en VP2 y el *knob* en VP3. El extremo N-terminal retraído habría evolucionado de manera independiente en *Iflavirus*, *Norovirus* y algunos *Picornaviridae*. La conservación de la inserción βC-αA en los dominios equivalentes a VP1 y VP3 en las proteínas de la cápside de virus de la familia *Secoviridae* sugiere, en conjunto con otras líneas de evidencia, que estas proteínas de la cápside no conservan un estado ancestral de la transición de las cápsides T=p3 a partir de cápsides T=3, sino que son derivaciones de cápsides de virus que infectan invertebrados como los de la familia *Dicistroviridae*. Los dominios P parecen haberse adquirido de manera independiente en virus de la familia *Tombusviridae* y otros virus con cápside T=3. Los dominios P y S de *Tombusviridae* no son parálogos, pero posiblemente sí los dominios P1 y P2 de *Astroviridae* y *Hepeviridae*. La distribución de dominios P en los árboles de similitud estructural sugiere que los dominios P1 y P2 de *Caliciviridae*, como el dominio P de VP3 de *Iflavirus*, se adquirieron de manera horizontal a partir de otros virus con cápside T=3 y dominios P *cradle-loop*. Finalmente, este análisis revela que, en genomas de RNA, la evolución de una familia de proteínas, puede involucrar eventos de duplicación, HGT y convergencias funcionales que, más que provocar una

pérdida en la adecuación debido al incremento en el tamaño del genoma, pueden generar ventajas como un rápido ensamblaje de complejos moleculares, la especialización funcional de proteínas parálogas y una fuente de constante adaptación a ambientes cambiantes.

7 Literatura citada

- Alifano, P., Fani, R., Liò, P., Lazcano, A., Bazzicalupo, M., Carlomagno, M. S., & Bruni, C. B. (1996). Histidine Biosynthetic Pathway and Genes: Structure, Regulation, and Evolution. *Microbiological Reviews, 60*(1), 44-69.
- Alva, V., Koretke, K. K., Coles, M., & Lupas, A. N. (2008). Cradle-loop barrels and the concept of metafolds in protein classification by natural descent. *Current Opinion in Structural Biology, 18*, 358-365.
- Ban, N., & McPherson, A. (1995). The structure of satellite panicum mosaic virus at 1.9 Å resolution. *Nature Structural Biology, 2*(10), 882-890.
- 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. *Annu. Rev. Ecol. Syst., 38*, 361–379.
- Belozerskii, A. N. (1959). On species specificity of nucleic acids in bacteria. In Oparin, A. I., Pasynskii, A. G., Braunshetin, A. E., & Pavloskaya, T. E. (Eds.), *The origin of life on earth* (pp. 313-321). New York: Pergamon Press/McMillan Company.
- Brachet, J. (1959). Les acides nucléiques et l'origine des proteins. In Oparin, A. I., Pasynskii, A. G., Braunshetin, A. E., & Pavloskaya, T. E. (Eds.), *The origin of life on earth* (pp. 313-321). New York: Pergamon Press/McMillan Company.
- Campillo-Balderas, J. A., Lazcano, A., & Becerra, A. (2015). Viral Genome Size Distribution Does not Correlate with the Antiquity of the Host Lineages. *Front. Ecol. Evol., 3*.

- Chandrasekar, V., & Johnson, J. E. (1997). The structure of tobacco ringspot virus: a link in the evolution of icosahedral capsids in the picornavirus superfamily. *Structure, 6*, 157-171.
- Chen, N., Yoshimura, M., Guan, H., Wang, T., Misumi, Y., Lin, C., Chuankhayan, P., Nakagawa, A., Chan, S. I., Tsukihara, T., Chen, T., & Chen, C. (2015). Crystal Structures of a Piscine Betanodavirus: Mechanisms of Capsid Assembly and Viral Infection. *PLoS Pathog., 11*(10).
- Chen, R., Neill, J. D., Estes, M. K., & Prasad, B. V. V. (2006). X-ray structure of a native calicivirus: Structural insights into antigenic diversity and host specificity. *PNAS, 103*(21), 8048-8053.
- Cisneros-Martínez, A. M. (2016). *Origen de genes por duplicación en virus de RNA* (Tesis de pregrado). Universidad Nacional Autónoma de México, Ciudad de México, México.
- Clayton, R. A., White, O., Ketchum, K. A., & Venter, J. C. (1997). The first genome from the third domain of life. *Nature, 387*, 459-462.
- Crick, F. H. C. (1968). The origin of the genetic code. *Journal of Molecular Biology, 39*, 367-379.
- Cui, J., Schlub, T. E., & Holmes, E. C. (2014). An Allometric Relationship between the Genome Length and Virion Volume of Viruses. *Journal of Virology, 88*(11), 6403–6410.
- Darriba, D., Taboada, G. L., Doallo, R., & Posada, D. (2011). ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics, 27*, 1164-1165.

- Davison, A. J., Benkő, M., & Harrach, B. (2003). Genetic content and evolution of adenoviruses. *Journal of General Virology, 84*, 2894-2908.
- Delaye, L., DeLuna, A., Lazcano, A., & Becerra, A. (2008). The origin of a novel gene through overprinting in Escherichia coli. *BMC Evolutionary Biology, 8*(31).
- Dong, J., Dong, L., Méndez, E., & Tao, Y. (2011). Crystal structure of the human astrovirus capsid spike. *PNAS, 108*(31), 12681-12686.
- Eakin, R. E. (1963). An approach to the evolution of metabolism. *PNAS, 49*, 360-366.
- Felsenstein, J. (1989). PHYLIP - Phylogeny Inference Package (Version 3.2). *Cladistics, 5*, 164-166.
- Finn, R. D., Clements, J., & Eddy, S. R. (2011). HMMER web server: interactive sequence similarity searching. *Nucleic Acids Research, 39*(2), W29-W37.
- Flores, R., Serra, P., Minoia, S., Di Serio, F., & Navarro, B. (2012). Viroids: from genotype to phenotype just relying on RNA sequence and structural motifs. *Frontiers in Microbiology, 3*.
- Fraenkel-Conrat, H., Singer, B., & Williams, R. C. (1957). Infectivity of viral nucleic acid. *Biochemica et Biophysica Acta, 25*, 87-96.
- Fraenkel-Conrat, H., & Singer, B. (1959). The infective nucleic acid of tobacco mosaic virus. In Oparin, A. I., Pasynskii, A. G., Braunshetin, A. E., & Pavloskaya, T. E. (Eds.), *The origin of life on earth* (pp. 313-321). New York: Pergamon Press/McMillan Company.
- Fuu, 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.

- Gago, S., Elena, S. F., Flores, R., & Sanjuán, R. (2009). Extremely High Mutation Rate of a Hammerhead Viroid. *Science, 323*, 1308.
- Gao, Y., Zhao, H., Jin, Y., Xu, X., & Han, G. (2017). Extent and evolution of gene duplication in DNA viruses. *Virus Research, 240*, 161-165.
- García-Meza, V., González-Rodríguez, A., & Lazcano, A. (1994). Ancient paralogous duplications and the search for Archean cells. In Fleischaker, G. R., Colonna, S., & Luisi, P. L. (Eds.), *Self-Reproduction of Supramolecular Structures: from Synthetic Structures to Models of Minimal Living Systems* (pp. 231-246). Dordrecht, Netherlands: Kliiwer Academic Press.
- Gevers, D., Vandepoele, K., Simillion, C., & Van de Peer, Y. (2004). Gene duplication and biased functional retention of paralogs in bacterial genomes. *TRENDS in Microbiology, 12*(4), 148-154.
- Gilbert, W. (1986). The RNA world. *Nature, 319*, 618.
- Gogarten, J. P., Kibak, H., Dittrich, P., Taiz, L., Bowman, E. J., Bowman, B. J., Manolson, M. F., Poole, R. J., Date, T., Oshima, T., Konishi, J., Denda, K., & Yoshida, M. (1989). Evolution of the vacuolar H+-ATPase: Implication for the origin of eukaryotes. *Proc. Natl. Acad. Sci., 86*, 6661-6665.
- Gorbalenya, A. E., Pringle, F. M., Zeddam, J., Luke, B. T. L., Cameron, C. E., Kalmakoff, J., Hanzlik, T. N., Gordon, K. H. J., & Ward, V. K. (2002). The Palm Subdomain-based Active Site is Internally Permuted in Viral RNA-dependent RNA Polymerases of an Ancient Lineage. *J. Mol. Biol., 324*, 47-62.

- Guerrier-Takada, C., Gardiner, K., Marsh, T., Pace, N., & Altman, S. (1983). The RNA Moiety of Ribonuclease P is the Catalytic Subunit of the Enzyme. *Cell, 35*, 849-857.
- Guindon, S., Dufayard, J. F., Lefort, V., Anisimova, M., Hordijk, W., & Gascuel, O. (2010). New Algorithms and Methods to estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Systematic Biology, 59*(3), 307-321.
- Guo, Y. R., Hryc, C. F., Jakana, J., Jiang, H., Wang, D., Chiu, W., Zhong, W., & Tao, Y. J. (2014). Crystal structure of a nematode-infecting virus. *PNAS, 111*(35), 12781-12786.
- Guu, T. S. Y., Liu, Z., Ye, Q., Mata, D. A., Li, K., Yin, C., Zhang, J., & Tao, Y. J. (2009). Structure of the hepatitis E virus-like particle suggests mechanisms for virus assembly and receptor binding. *PNAS, 106*(31), 12992-12997.
- Haldane, J. B. S. (1954). The origins of life. *New Biology, 16,* 12-27.
- Handler, P. (1963). Evolution of the coenzymes. In Oparin, A. I. (Ed.), *Proceedings of the Fifth International Congress of Biochemistry, Vol. III. Biochemistry* (pp. 149-157). New York: Pergamon Press/McMillan Company.
- Hartman, H. (1975). Speculations on the origin and evolution of metabolism. *Journal of Molecular Evolution, 4*, 359-370.
- Holm, L., & Laakso, L. M. (2016). Dali server update. *Nucleic Acids Research, 44*, W351-W355.
- Holmes, E. C. (2009). *The Evolution and Emergence of RNA Viruses*. Oxford: Oxford University Press.
- Holmes, E. C. (2011). What Does Virus Evolution Tell Us about Virus Origins? *Journal of Virology, 85*(11), 5247-5251.

- Hughes, A. L., & Friedman, R. (2005). Poxvirus genome evolution by gene gain and loss. *Molecular Phylogenetics and Evolution, 35*, 186-195.
- Hulo, C., de Castro, E., Masson, P., Bougueleret, L., Bairoch, A., Xenarios, I., & Le Mercier, P. (2011). ViralZone: a knowledge resource to understand virus diversity. *Nucleic Acids Res., 39*, D576-82.
- Humphrey, W., Dalke, A., & Schulten, K. (1996). VMD: Visual Molecular Dynamics. *Journal of Molecular Graphics, 14*, 33-38.
- Iwabe, N., Kuma, K., Hasegawa, M., Osawa, S., & Miyata, T. (1989). Evolutionary relationship of archaebacteria, eubacteria and eukaryotes inferred from phylogenetic tres of duplicated genes. *Proc. Natl. Acad. Sci., 86*, 9355-9359.
- Jones, E. Y., Stuart, D. I., & Walker, N. P. C. (1989). Structure of tumor necrosis factor. *Nature, 338*, 225-228.
- Kaessmann, H. (2010). Origins, evolution, and phenotypic impact of new genes. *Genome Research, 20*, 1313–1326.
- Kalynych, S., Pálková L. & Plevka, P. (2016). The Structure of human Parechovirus 1 Reveals an Association of the RNA Genome with the Capsid. *Journal of Virology, 90*, 1377-1386.
- Kalynych, S., Přidal, A., Pálková, L., Levdansky, Y., de Miranda, J. R., & Plevka, P. (2016). Virion Structure of Iflavirus Slow Bee Paralysis Virus at 2.6-Angstrom Resolution. *Journal of Virology, 90*, 7444-7455.

- Kazlauskas, D., Krupovic, M., and Venclovas, C. (2016). The logic of DNA replication in double-stranded DNA viruses: insights from global analysis of viral genomes. *Nucleic Acids Research, 44*(10), 4551–4564.
- King, A. M. Q., Adams, M. J., Carstens, E. B., & Lefkowitz, E. J. (2012). *Virus Taxonomy. Classification and Nomenclature of Viruses: Ninth Report of the International Committee on Taxonomy of Viruses*. Cambridge, Massachusetts: Elsevier Academic Press.
- Koonin, E. V., Wolf, Y. I., Nagasaki, K., & Dolja, V. V. (2008). The Big Bang of picorna-like virus evolution antedates the radiation of eukaryotic supergroups. *Nature, 6*, 925-939.
- Krissinel, E., & Hendrick, K. (2004). Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Cryst. D60*, 2256-2268.
- Kruger, K., Grabowski, P. J., Zaug, A. J., Sands, J., Gottschling, D. E., & Cech, T. R. (1982). Self-Splicing RNA: Autoexcision and Autocylization of the Ribosomal RNA Intervening Sequence of Tetrahymena. *Cell, 31*, 147-157.
- Krupkin, M., Matzov, D., Tang, H., Metz, M., Kalaora, R., Belousoff, M. J., Zimmerman, E., Bashan, A., & Yonarh, A. (2011). A vestige of a prebiotic bonding machine is functioning within the contemporary ribosome. *Phil. Trans. Ro. Soc. B., 366*, 2972-2978.
- Krupovic, M., & Koonin, E. V. (2017). Multiple origins of viral capsid proteins from cellular ancestors. *PNAS*, 1-10.

- Kunin, V., Cases, I., Enright, A. J., de Lorenzo, V., & Ouzounis, C. A. (2003). Myriads of protein families, and still counting. *Genome Biology, 4*(2), 401.
- Lazcano, A., Fox, G. E., & Oró, J. F. (1992). Life Before DNA: The Origin and Evolution of Early Archean Cells. In Robert P. Mortlock (Ed.), *The Evolution of Metabolic Function* (pp. 237-295). Boca Raton, Florida: CRC.
- Lazcano, A. (1995). Cellular evolution during the early Archean: what happened between the progenote and the cenancestor?. *Microbiología SEM, 11*, 185-198.
- Lazcano, A. (2012). The Biochemical Roots of the RNA World: from Zymonucleic Acid to Ribozymes. *Hist. Phil. Life Sci., 34*, 407-424.
- Lazcano, A. (2014). The RNA World: stepping out of the shadows. In Gabriel Trueba (Ed.), *Why does evolution matters? The importance of understanding evolution* (pp. 101-119). Newcastle upon Tyne: Cambridge Scholars Publishing.
- Li, W., & Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics, 22*, 1658-1659.
- Liljas, L., Tate, J., Lin, T., Christian, P., & Johnson, J. E. (2002). Evolutionary and taxonomic implications of conserved structural motifs between picornaviruses. *Archives of Virology, 157*, 59-84.
- Magadum, S., Banerjee, U., Murugan, P., Gangapur, D., & Ravikesavan, R. (2013). Gene duplication as a major force in evolution. *Journal of Genetics, 92*(1), 155-161.
- McGeoch, D., & Davison, J. (1999). Molecular evolutionary history of the herpesviruses. In Domingo, E., Webster, R. G., & Holland, H. F. (Eds.), *Origin and Evolution of Viruses* (pp. 441-465). London, UK: Academic Press.

- Ohno, S. (1970). *Evolution by Gene Duplication*. New York, NY: Springer-Verlag.
- Orgel, L. E. (1968). Evolution of the genetic apparatus. *Journal of Molecular Biology, 38*, 381-393.
- Pei, J., Kim, B., & Grishin, N. V. (2008). PROMALS3D: a tool for multiple sequence and structure alignment. *Nucleic Acids Res., 36*(7), 2295-2300.
- 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. *J. Comput. Chem., 25*(13), 1605-1612.
- Prasad, B. V. V., Hardy, M. E., Dokland, T., Bella, J., Rossmann, M. G., & Estes, M. K., (1999). X-ray Crystallographic Structure of the Norwalk Virus Capsid. *Science, 286*, 287-290.
- Reanney, D. C. (1982). The evolution of RNA viruses. *Annu. Rev. Microbiol., 36*, 47–73.
- Reyes-Prieto, F., Hernández-Morales, R., Jácome, R., Becerra, A., & Lazcano, A. (2012). Coenzymes, viruses and the RNA world. *Biochimie, 94*, 1467-1473.
- Rich, A. (1962). On the problems of evolution and biochemical information transfer. In Kasha, M. & Pullman, B. (Eds.), *Horizons in Biochemistry* (pp. 103-126). New York: Academic Press.
- Roberts, E., Eargle, J., Wright, D., & Luthey-Schulten, Z. (2006). MultiSeq: unifying sequence and structure data for evolutionary analysis. *BMC Bioinformatics, 7*.
- Rossmann, M. G., Arnold, E., Erickson, J. E., Frankenberger, E. A., Griffith, J. P., Hecht, H., Johnson, J. E., Kamer, G., Luo, M., Mosser, A. G., Rueckert, R. R., Sherry,

B., & Vriend, G. (1985). Structure of a human common cold virus and functional relationship to other picornaviruses. *Nature, 317*, 145-153.

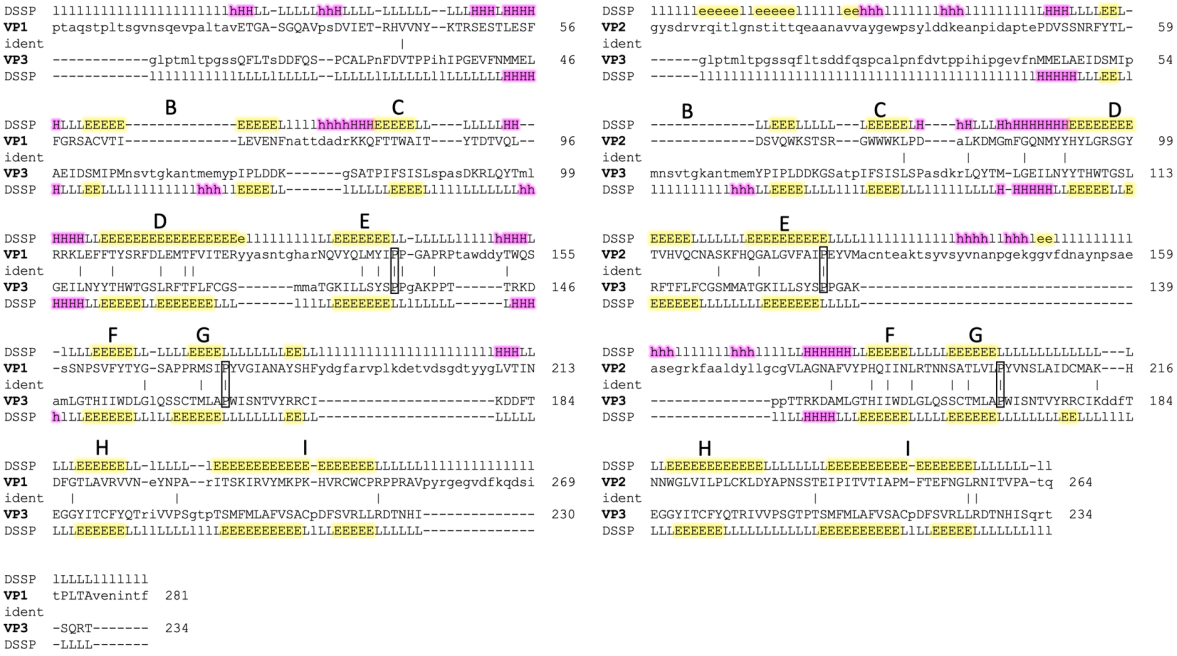
- Rossmann, M. G., & Johnson, J. E. (1989). Icosahedral RNA virus structure. *Annu. Rev. Biochem., 58*, 533-73.
- Russell, R. B., & Barton, G. J. (1992). Multiple Protein Sequence Alignment from Tertiary Structure Comparison: Assignment of Global and Residue Confidence Level. *PROTEINS: Structure, Function and Genetics, 14*, 309-323.
- Sabin, C., Füzik, T., Škubník, K., Pálková, L., Lindberg, A. M. & Plevka, P. (2016). Structure of Aichi Virus 1 and Its Empty Particle: Clues to Kobuvirus Genome Release Mechanism. *Journal of Virology, 90*, 10800-10810.
- Schellenberger, P., Sauter, C., Lorber, B., Bron, P., Trapani, S., Cergdoll, M., Marmonier, A., Schmitt-Keichinger, C., Lemaire, O., Demangeat, G., Ritzenthaler, C. (2011). Structural Insights into Viral Determinants of Nematode Mediated Grapevine fanleaf virus Transmission. *PLoS Pathog., 7*(5).
- Shackelton, L. A., & Holmes, E. C. (2004). The evolution of large DNA viruses: combining genomic information of viruses and their hosts. *TRENDS in Microbiology, 12*(10), 458–465.
- Simon-Loriere, E., & Holmes, E. C. (2012). Why do RNA viruses recombine?. *Nat. Rev. Microbiol., 9*(8), 617-626.
- Simon-Loriere, E., & Holmes, E. C. (2013). Gene Duplication is Infrequent in the Recent Evolutionary History of RNA Viruses. *Mol. Biol. Evol., 30*(6), 1263–1269.

- Spurny, R., Pridal A., Pálková, L., Kiem, H. K. T., de Miranda, J. R., & Plevka, P. (2017). Virion Structure of Black Queen Cell Virus, a Common Honeybee Pathogen. *Journal of Virology, 91*, 1-14.
- Stanley, W. M. (1935). Isolation of a crystalline protein possessing the properties of tobacco-mosaic virus. *Science, 81*, 644-645.
- Stanley, W. M. (1959). On the nature of viruses, genes and life. In Oparin, A. I., Pasynskii, A. G., Braunshetin, A. E., & Pavloskaya, T. E. (Eds.), *The origin of life on earth* (pp. 313-321). New York: Pergamon Press/McMillan Company.
- Strauss, M., Filman, D. J., Belnap, D. M., Cheng, N., Noel, R. T., & Hogle, J. M. (2015). Nectin-Like interactions between Poliovirus and Its Receptor Trigger Conformational Changes Associated with Cell entry. *Journal of Virology, 89*(8), 4143-4157.
- Subbiah, S., Laurents, D. V., & Levitt, M. (1993). Structural similarity of DNA-binding domains of bacteriophage repressors and the globin core. *Current Biology, 3*, 141-148.
- Tate, J., Liljas, L., Scotti, P., Christian, P., Lin, T., Johnson, J. E. (1999). The crystal structure of cricket paralysis virus: the first view of a new virus family. *Nat. Struct. Biol., 6*(8), 765-774.
- Thompson, J. R., Kamath, N., & Perry, K. L. (2014). An Evolutionary Analysis of the Secoviridae Family of Viruses. *PLoS ONE, 9*(9).
- Toh, Y., Harper, J., Dryden, K. A., Yeager, M., Arias, C. F., Méndez, E., & Tao, Y. J. (2016). Crystal Structure of the Human Astrovirus Capsid Protein. *Journal of Virology, 90*(20), 9008-9017.

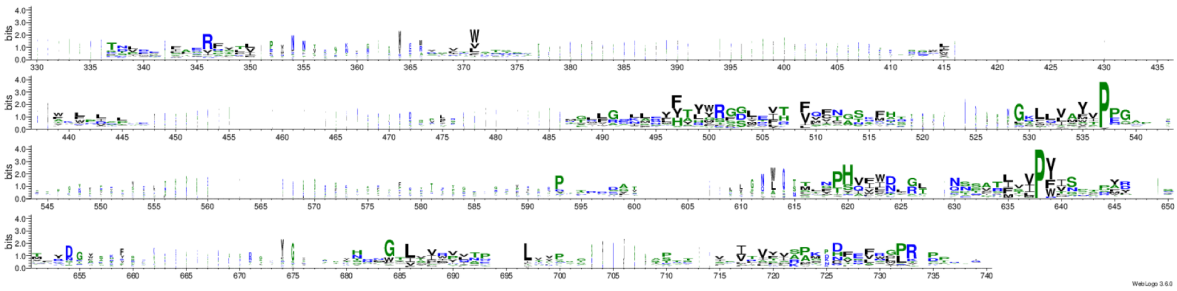
- Tully, J. G., Cole, R. M., Taylor-Robinson, D., & Rose, D. L. (1981). A newly discovered Mycoplasma in the human urogenital tract. *The Lancet, 317*(8233), 1288-1291.
- Tuthill, T. J., Harlos, K., Walter, T. S., Knowles, N. J., Groppelli, E., Rowlands, D. J., Stuart, D. I., & Fry, E. E. (2009) Equine Rhinitis A Virus and Its Low pH Empty Particle: Clues Towards an Aphthovirus Entry Mechanism?. *PLoS Pathog, 5*(10), 1-11.
- Velasco, A. M., Becerra, A., Hernández-Morales, R., Delaye, L., Jiménez-Corona, M. E., Ponce-de-Leon, S., & Lazcano, A. (2013). Low complexity regions (LCRs) contribute to the hypervariability of the HIV-1 gp120 protein. *Journal of Theoretical Biology, 338*, 80-86.
- Wang, X., Ren J., Gao, Q., Hu, Z., Sun, Y., Li, X., Rowlands, D. J., Yin, W., Wang, J., Stuart, D. I., Rao, Z., & Fry, E. E. (2015). Hepatitis A virus and the origin of picornaviruses. *Nature, 517*(7532), 85-88.
- White III, H. B. (1976). Coenzymes as fossils of an earlier metabolic state. *Journal of Molecular Evolution, 7*, 101-104.
- Willemsen, A., Zwart, M. P., Higueras, P., Sardanyés, J., & Elena, S. F. (2016). Predicting the stability of homologous gene duplications in a plant RNA virus. *Genome Biol. Evol., 8*(9), 3065–3082.
- Woese, C. R. (1967). *The Genetic Code: the molecular basis for gene expression*. New York: Harper and Row.

- Wolf, Y. I., Kazlauskas, D., Iranzo, J., Lucía-Sanz, A., Kuhn, J. H., Krupovic, Mñ, Dolja, V. V., & Koonin, E. V. (2018). Origins and Evolution of the Global RNA Virome. *mBio, 9*(6), e02329-18.
- Xing, L., Li, T., Mayazaki, N., Simon, M. N., Wall, J. S., Moore, M., Wang, C., Takeda, N., Wakita, T., Miyamura, T., & Cheng, R. H. (2010). Structure of Hepatitis E Virion-sized Particle Reveals an RNA-dependent Viral Assembly Pathway. *The Journal of Biological Chemestry, 285*(43), 33175-33183.
- Zimmermann, L., Stephens, A., Nam, S., Rau, D., Kübler, J., Lozajic, M., Söding, J., Lupas, A. N., & Alva, V. (2018). A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. *Journal of Molecular Biology, 430*(15), 2237-2243.
- Zocher, G., Mistry, N., Frank, M., Hähnlein-Schick, I., Arnberg, N., & Stehle, T. (2014). A Sialic Acid Binding Site in a Human Picornavirus. *PLoS Pathog., 10*(10).

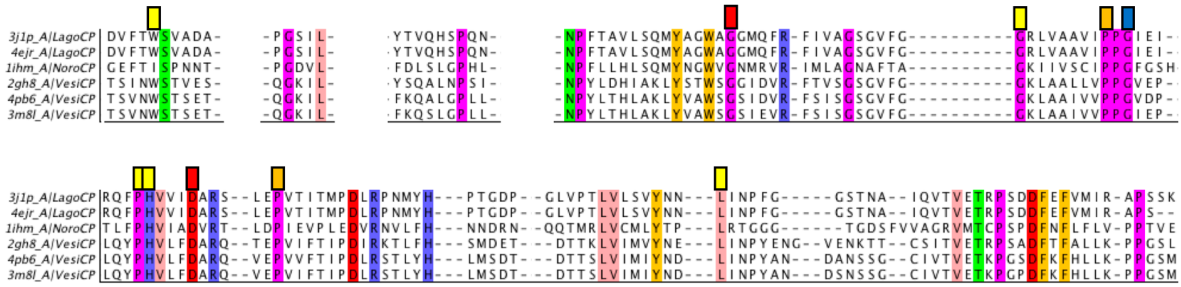
8 Anexos



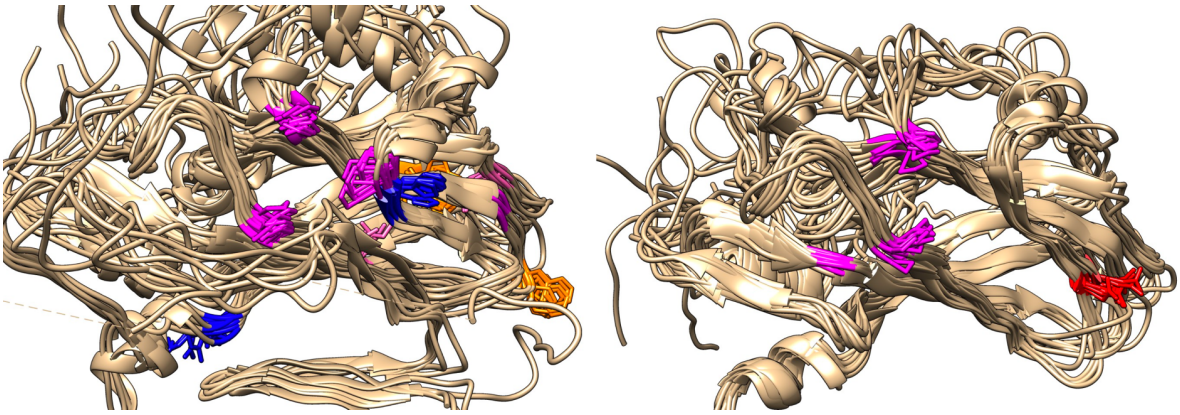
Anexo 1. Alineamientos pareados de estructura primaria y secundaria derivados de las superposiciones estructurales con DALI. A la izquierda el alineamiento entre VP1 y VP3. A la derecha el alineamiento entre VP2 y VP3. Las indicaciones son las mismas que las de la figura 5.



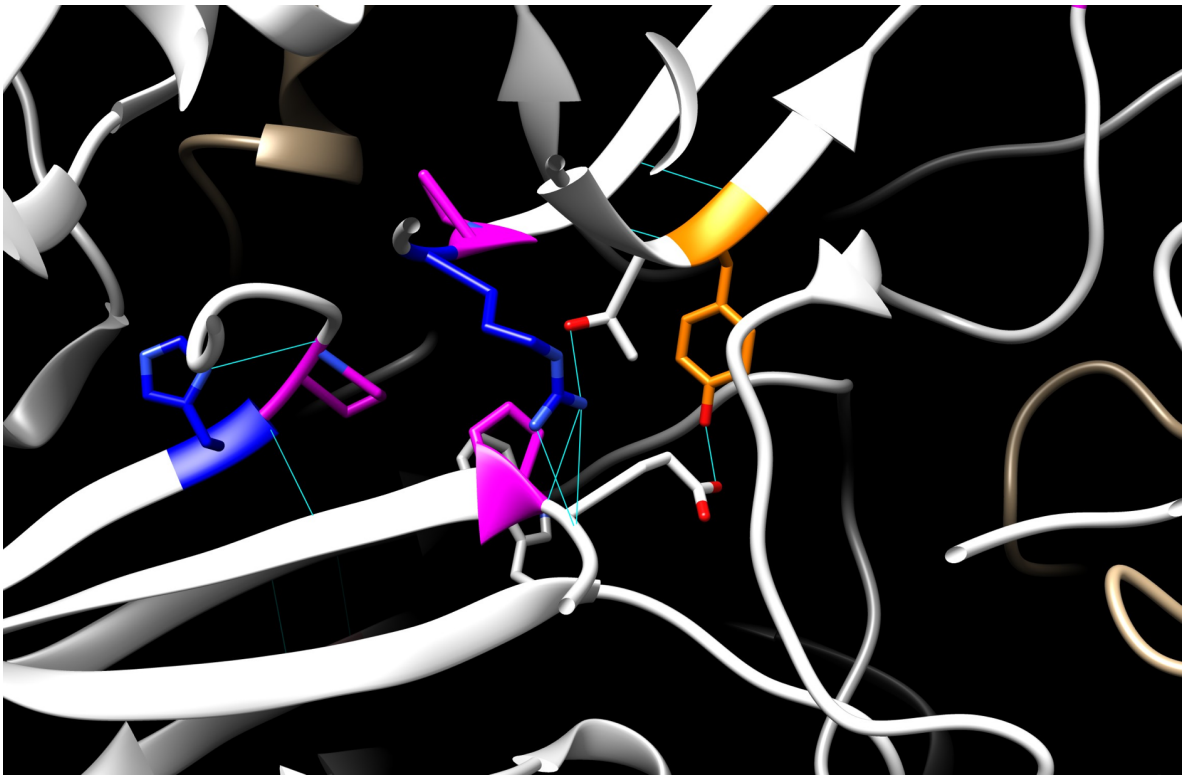
Anexo 2. Logo mostrando los patrones de conservación de las 138 secuencias alineadas con PROMALS3D. Realizado con WebLogo 3.6.0 en Mayo de 2018.



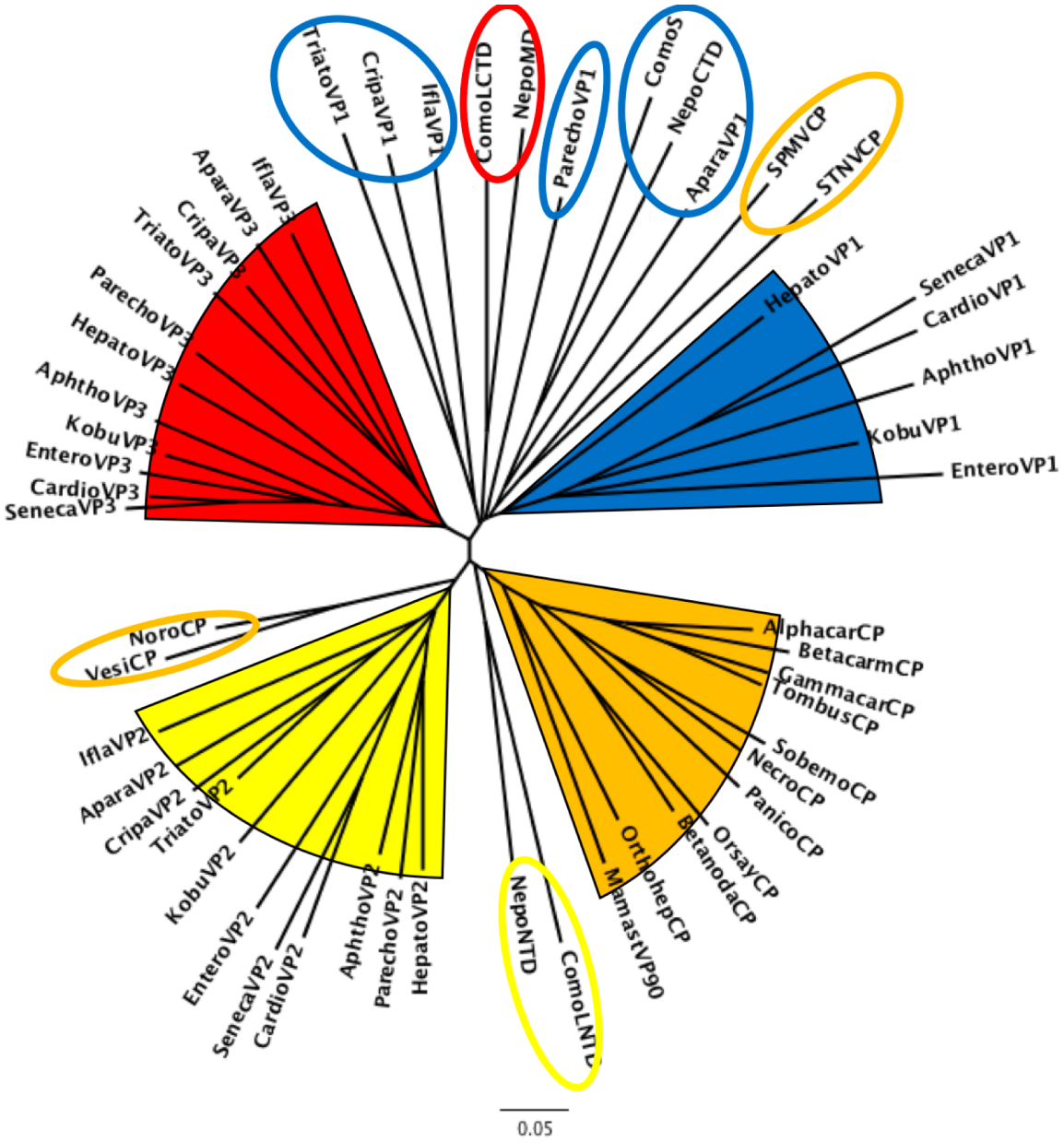
Anexo 3. Residuos conservados en CP de *Caliciviridae*. En la parte superior se indican los residuos conservados entre CP y VP1 (rectángulo azul), CP y VP2 (rectángulo amarillo) y CP y VP3 (rectángulo rojo). Los rectángulos anaranjados indican los residuos conservados en VP1, VP2, VP3 y CP.



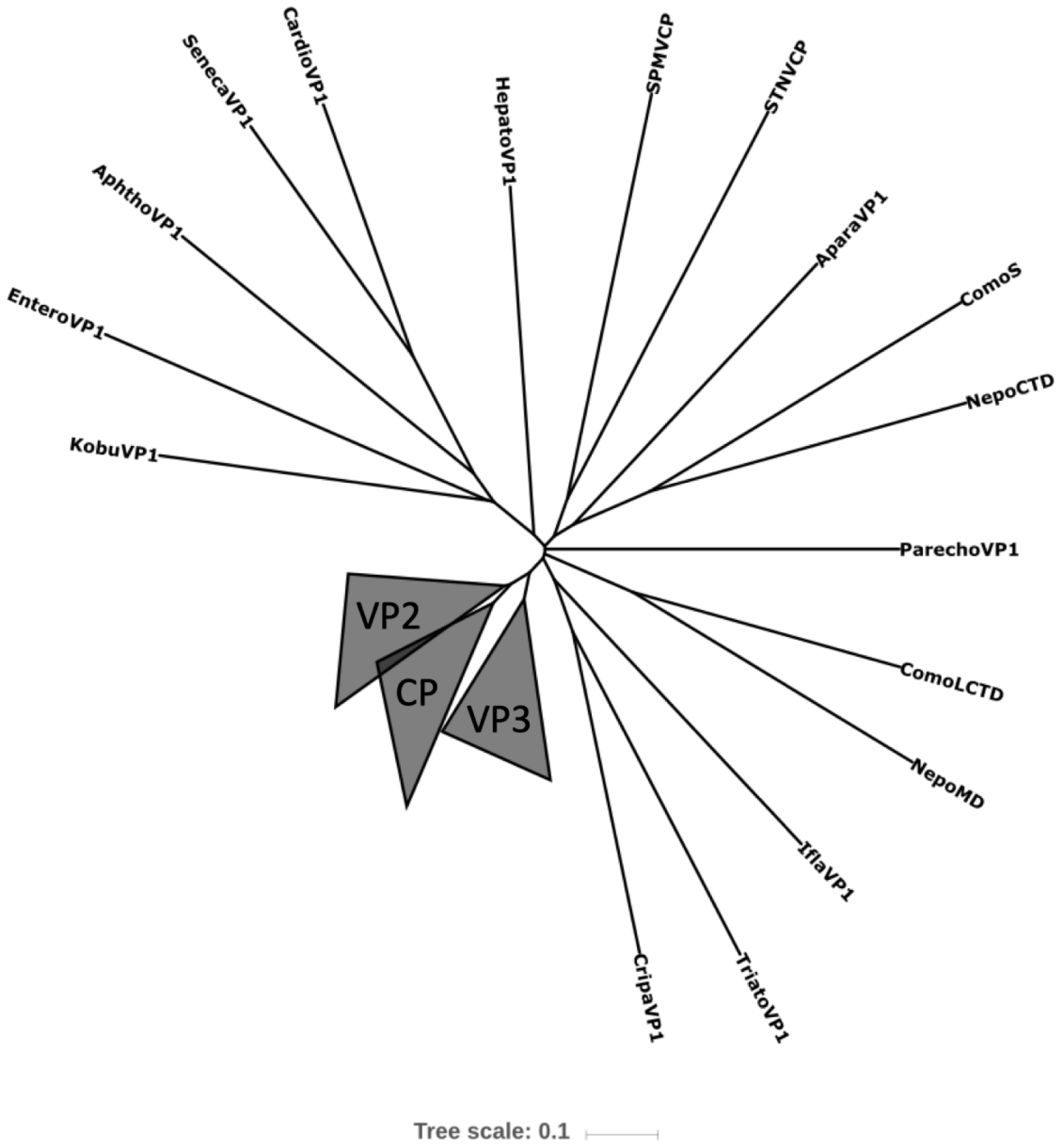
Anexo 4. Residuos conservados de VP2 (izquierda) y VP3 (derecha) mostrados sobre la superposición de modelos estructurales representativos de cada género.



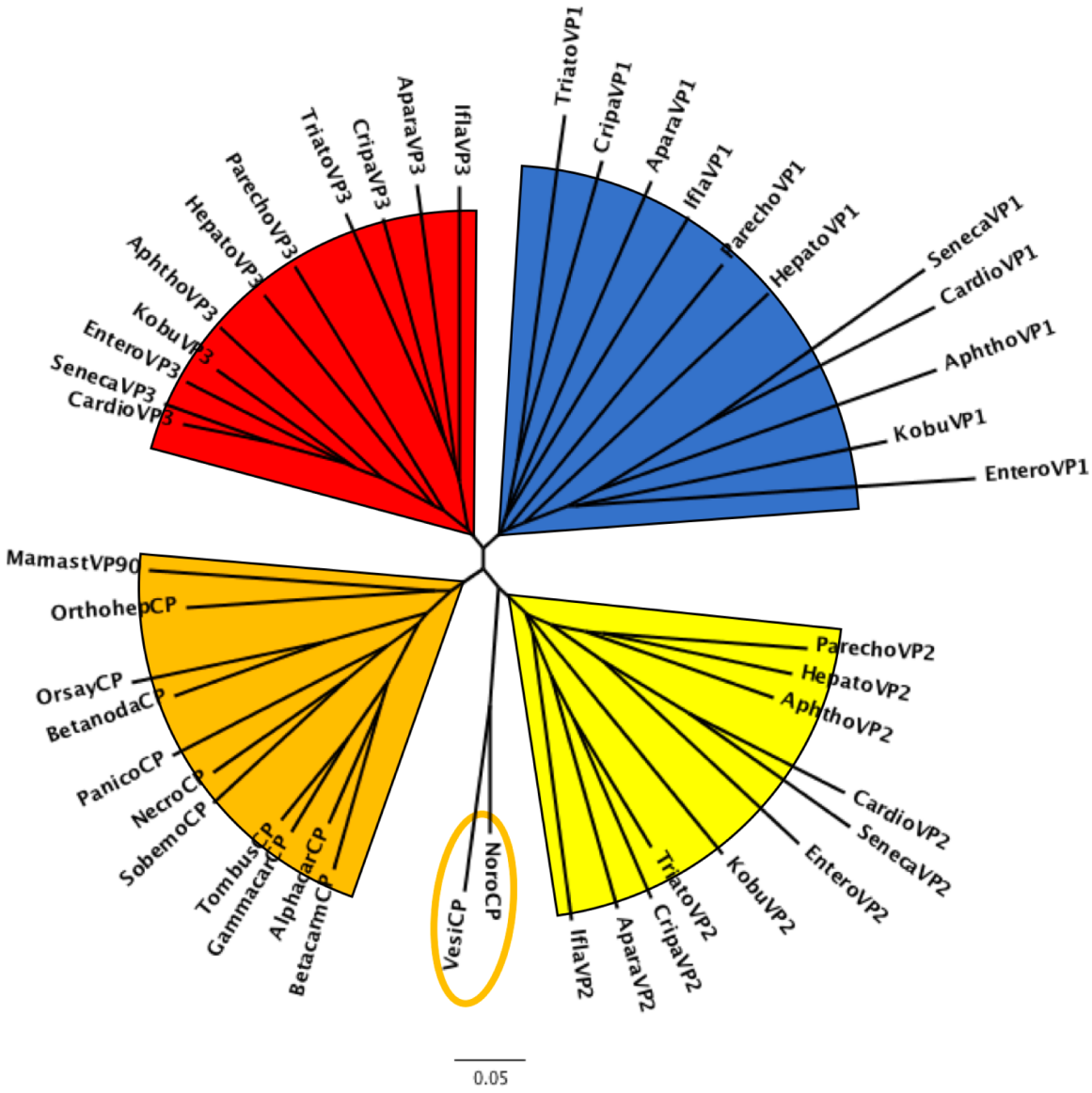
Anexo 5. Predicción de puentes de hidrógeno entre residuos conservados de VP1 y VP2, realizado con Chimera. PDB: 3j8f. Las interacciones predichas son entre Tyr127 (inicio de D) de VP1 y a Glu129 de VP2, y Arg272 (final de I) de VP1 con la cadena principal de Glu129 y de Pro128 de VP2.



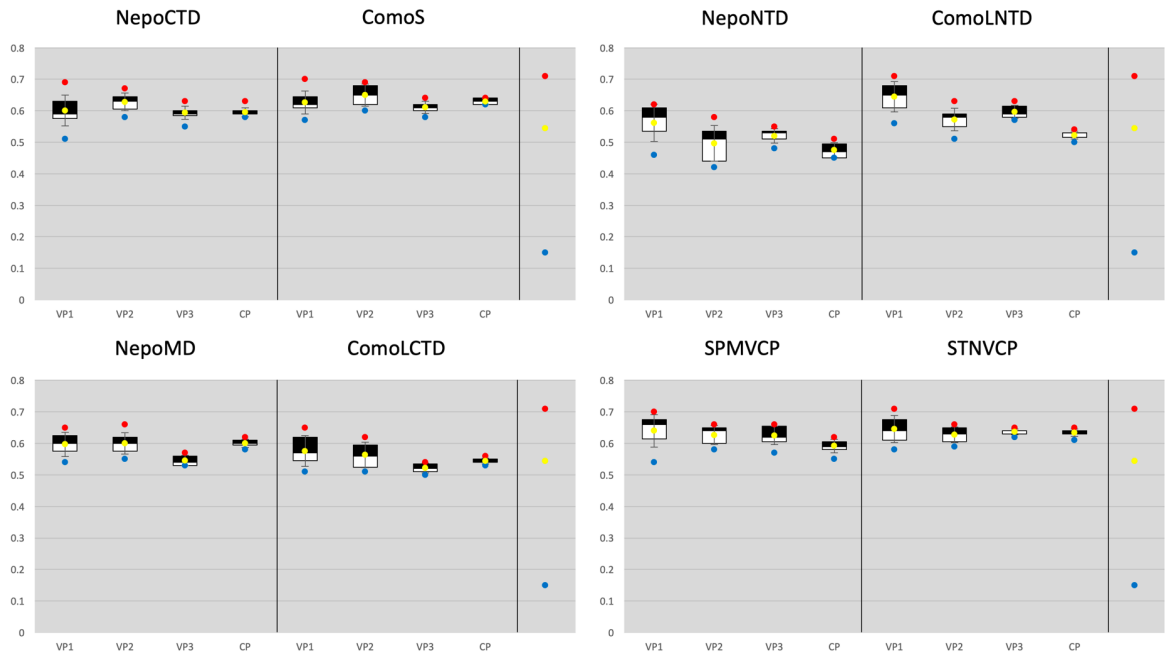
Anexo 6. Dendrograma de similitud estructural evaluada por superposición con STAMP y matriz de distancias Q_H. Seed = 41; SS = 5.79223; APSD = 4.50028.



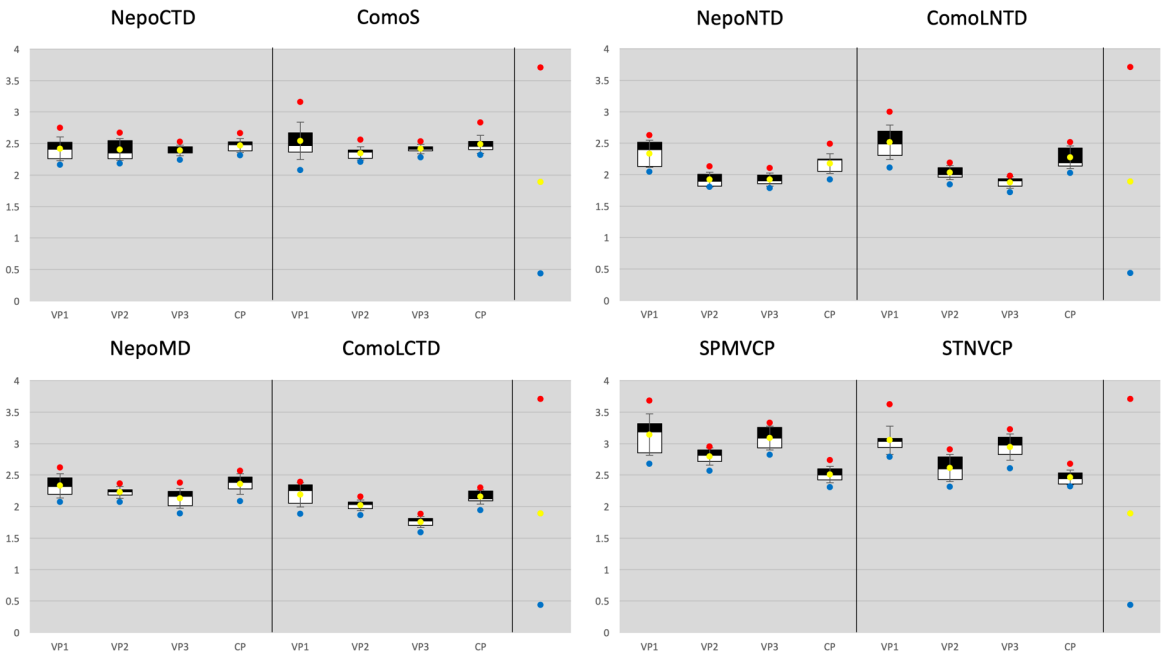
Anexo 7. Árbol de distancias Q_H con ramas colapsadas si su longitud es menor al promedio (0.27).



Anexo 8. Árbol de distancias Q_H sin las ramas de *Secoviridae* y de virus satélite.

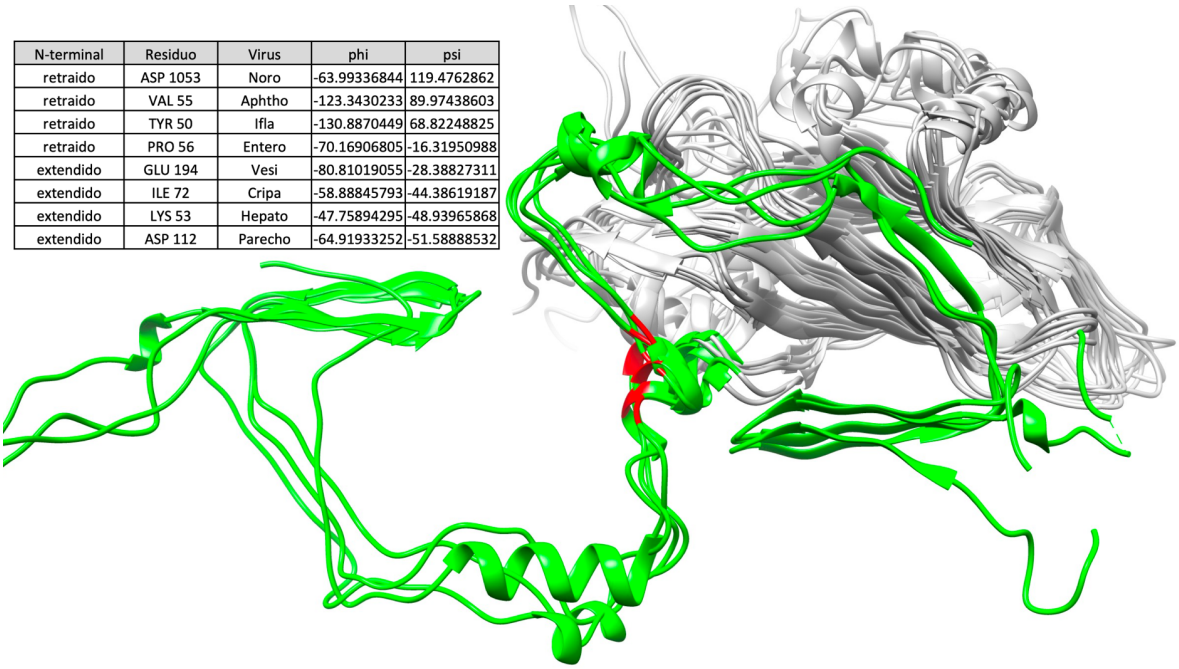


Anexo 9. Distribución de distancias Q_H de las proteínas de la cápside de *Secoviridae* y de virus satélite con respecto a las proteínas de los cuatro clados principales (VP1, VP2, VP3 y CP). Se muestran los valores mínimos (azul), medios (amarillo) y máximos (rojo) en cada caso. En la última columna de cada gráfico se muestra el valor mínimo, medio y máximo de la matriz completa. Las distancias de las CP de *Caliciviridae* y de virus satélite no son incluídas en el clado CP para el análisis de distancias. Los dominios equivalentes a VP1, VP2 y VP3 de los *Secoviridae* tampoco son incluídos en sus respectivos clados para el analisis de distancias.

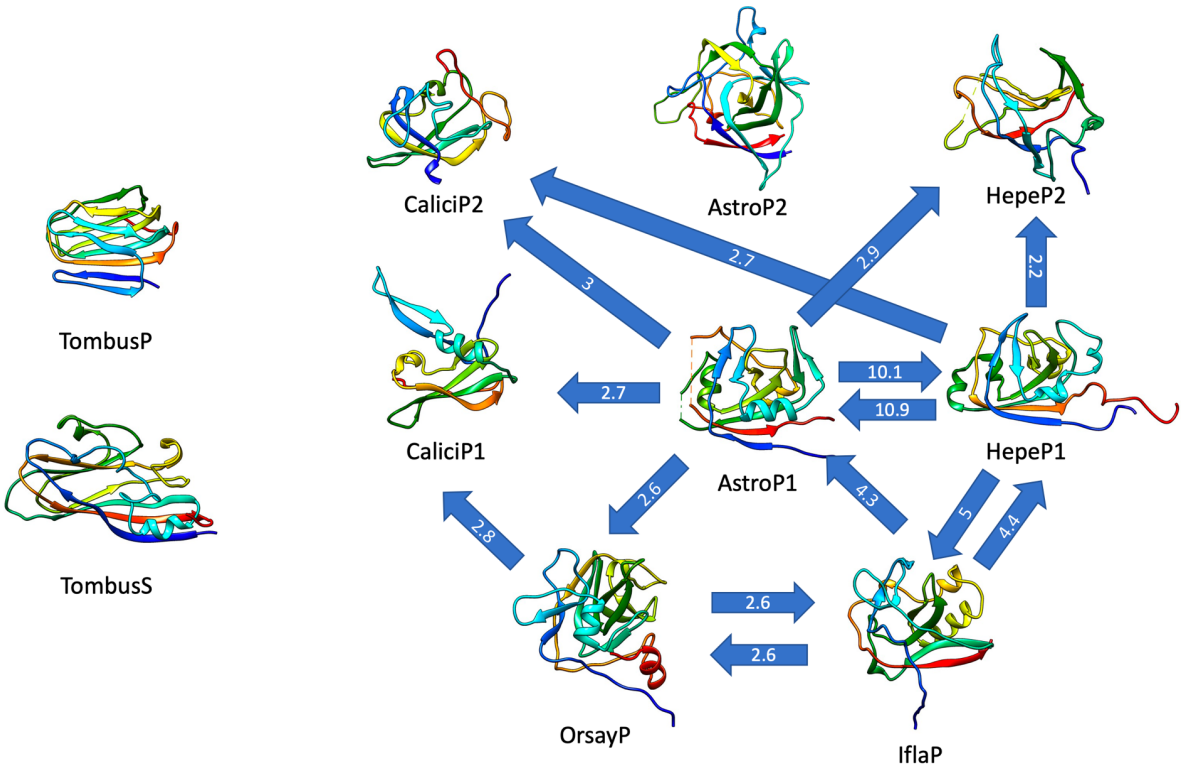


Anexo 10. Distribución de distancias SAS de las proteínas de la cápside de *Secoviridae* y de virus satélite con respecto a las proteínas de los cuatro clados principales (VP1, VP2, VP3 y CP). Se muestran los valores mínimos (azul), medios (amarillo) y máximos (rojo) en cada caso. En la última columna de cada gráfico se muestra el valor mínimo, medio y máximo de la matriz completa. Las distancias de las CP de *Caliciviridae* y de virus satélite no son incluídas en el clado CP para el análisis de distancias. Los dominios equivalentes a VP1, VP2 y VP3 de los *Secoviridae* tampoco son incluídos en sus respectivos clados para el analisis de distancias.

N-terminal	Residuo	Virus	phi	psi
retraido	ASP 1053	Noro	-63.99336844	119.4762862
retraido	VAL 55	Aphtho	-123.3430233	89.97438603
retraido	TYR 50	Ifla	-130.8870449	68.82248825
retraido	PRO 56	Entero	-70.16906805	-16.31950988
extendido	GLU 194	Vesi	-80.81019055	-28.38827311
extendido	ILE 72	Cripa	-58.88845793	-44.38619187
extendido	LYS 53	Hepato	-47.75894295	-48.93965868
extendido	ASP 112	Parecho	-64.91933252	-51.58888532



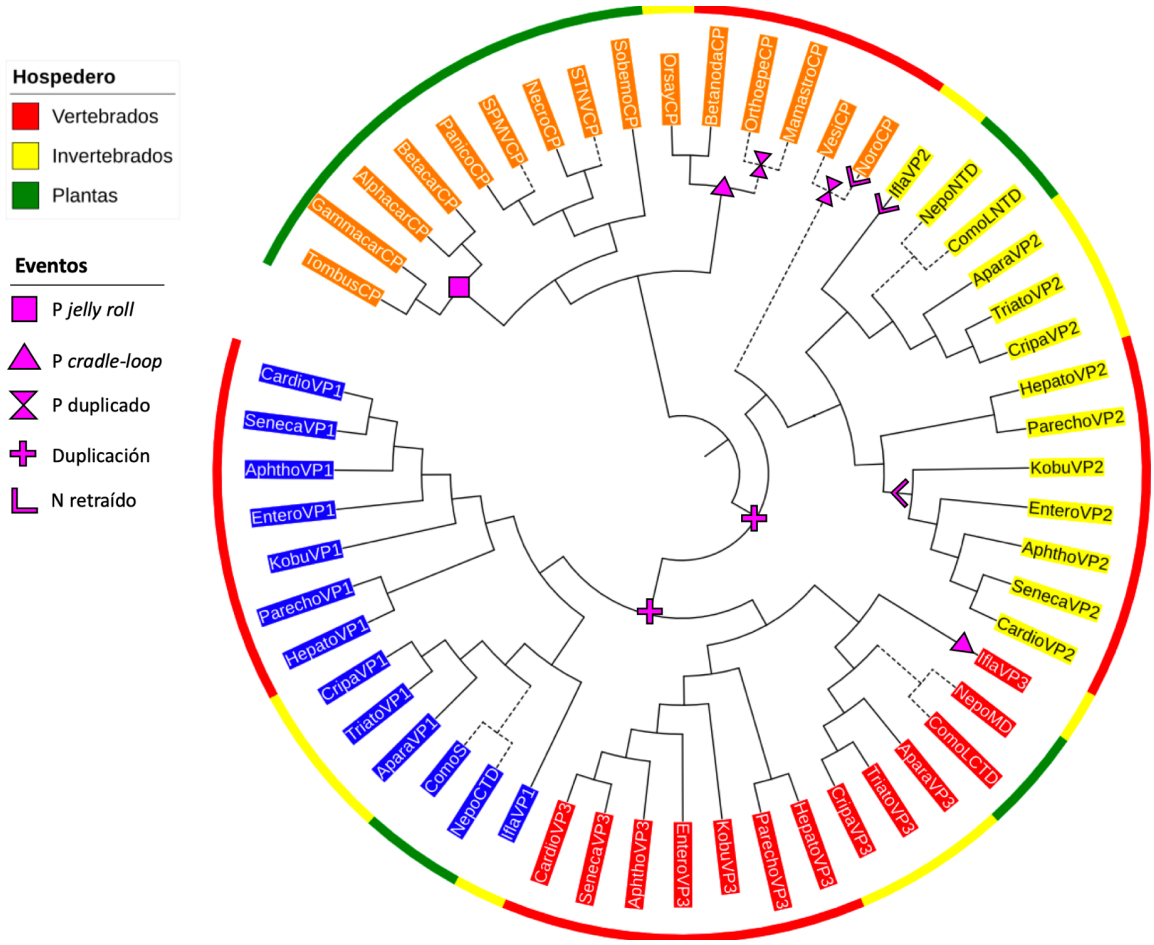
Anexo 11. Ángulos dihedros de los residuos equivalentes a LYS 53 de *Hepatovirus*. En la tabla se muestran los valores y en la imagen se indican estos residuos en color rojo. Los extremos N-terminal están representadas en color verde.



Anexo 12. Valores Z entre los diferentes dominios P. Las estructuras se muestran coloreadas en gradiente de azul a rojo de N-terminal a C-terminal.

No:	Chain	Z	rmsd	lali	nres	%id	Description
1	2zah-C	17.4	1.7	115	331	22	Tombusviridae P domain
2	2tbv-C	16	1.5	113	320	24	Tombusviridae P domain
3	4ht1-T	7.3	2.7	90	131	6	*Homo sapiens* TNF superfamily 12
4	2r32-A	7.3	3.1	91	140	9	*Homo sapiens* TNF superfamily 18
5	2r6q-A	7.2	2.8	98	138	16	*Bacillus anthracis* Bcla
6	1rj7-E	6.8	2.9	94	149	12	*Homo sapiens* Ectodysplasin A (TNF-like)
7	1yq8-A	6.6	2.4	88	191	9	Enterobacteria phage PRD1 (dsDNA Tectiviridae) P5
8	3lkj-B	6.5	3.2	96	132	8	*Homo sapiens* Cytokine CD40 ligand (TNF-like)
9	4i8e-X	6.4	2.9	92	359	10	*Streptococcus gordonii* adhesin GspB
10	5wux-F	6.3	3.3	92	141	9	*Homo sapiens* TNF alpha

Anexo 13. Modelos estructurales con los 10 mejores valores Z de la búsqueda de DALI con el dominio P de *Tombusviridae*.



Anexo 14. Esquema que resume la evolución de las cápsides de virus del orden Picornavirales y sus homólogos con cápsides T=3. Las relaciones poco claras, así como aquellas no observadas en los árboles de similitud estructural pero sugeridas por la conservación de caracteres estructurales y por otros estudios, son indicadas con líneas punteadas. La distribución de los dominios P se indica sobre el esquema de las relaciones evolutivas.