UNIVERSIDAD NACIONAL AUTÓNOMA

DE MÉXICO

FACULTAD DE CIENCIAS

ORIGEN DE GENES POR DUPLICACIÓN EN VIRUS DE RNA

T E S I S

QUE PARA OBTENER EL TÍTULO DE:

BIÓLOGO

P R E S E N T A:

ALEJANDRO MIGUEL CISNEROS MARTÍNEZ

DIRECTOR DE TESIS:

DR. ANTONIO EUSEBIO LAZCANO-ARAUJO REYES

Ciudad Universitaria, Cd. Mx., 2016

HOJA DE DATOS DEL JURADO

1. Datos del alumno
   Cisneros
   Martínez
   Alejandro Miguel
   29 76 11 58
   Universidad Nacional Autónoma de México
   Facultad de Ciencias
   Biología
   310597564

2. Datos del tutor
   Dr
   Antonio Eusebio
   Lazcano-Araujo
   Reyes

3. Datos del sinodal 1
   Dr
   Víctor Manuel
   Valdés
   López

4. Datos del sinodal 2
   Dra
   Beatriz
   Gómez
   García

5. Datos del sinodal 3
   Dr
   Lorenzo Patrick
   Segovia
   Forcella

6. Datos del sinodal 4
   Dr
   León Patricio
   Martínez
   Castilla

7. Datos del trabajo escrito
   Origen de genes por duplicación en virus de RNA
   69p
   2016

La realidad está compuesta de objetos materiales, que coexisten en un espacio-tiempo, de los cuales algunos resultan ser sujetos capaces de percibirla e interpretarla.

Me es inevitable sentir fascinación por todo lo que se puede descubrir con el ejercicio de la ciencia. Sin embargo, no estoy convencido de que ciencia sea lo que el mundo necesita.

Cualquiera podría ser Einstein o Jesús. Lo único que hace falta es asumirse capaz y un par de oportunidades.

Entender para no juzgar. No juzgar para ser y dejar ser libre. Juzgar separa, entender hace que todo parezca más familiar.

“Si no vives como piensas acabarás pensando como vives.” -Mahatma Gandhi

“El precio de desentenderse de la política es el ser gobernado por los hombres peores.” -Platón

AGRADECIMIENTOS

El agradecimiento me da humildad. Es una certeza de que solo, con mi soberbia, mi ambición no habría llegado lejos. Es el reconocimiento del valor de los otros y de la insignificancia de mi ego.

Agradezco profundamente a mis padres, Miguel Cisneros Ramírez y Luz María Martínez Mejía, que son mi primer ejemplo a seguir y que me han dado un apoyo incondicional basado en amor y familia. También han sido guías y consejeros a lo largo de mi vida y durante la carrera.

Agradezco a mi hermano, Héctor Miguel Cisneros Martínez, que ha sido un catalizador de cambios en mi vida y una inspiración para superarme.

Siempre esperé este momento para agradecer a mi tío, José Guadalupe Cisneros Ramírez, a quien, por despertar mi fascinación por el conocimiento, considero mi padre intelectual, y cuya memoria permanecerá siempre en mi mente.

Agradezco al Dr. Antonio Lazcano, cuyo ejemplo de tutor, como fe en mi potencial, han sido una motivación para seguir en el camino de la ciencia. También agradezco a mis sinodales por sus valiosos comentarios para la elaboración de esta tesis.

Agradezco mis compañeros de laboratorio, los macacos, que son un ejemplo de fraternidad y una inspiración académica.

Agradezco mucho a mis amigos de la carrera, especialmente a Carlos Antonio González Palma y Armando Rodríguez Velasco, y a todos los miembros de Esporopo-Lignina.

Finalmente, agradezco a todas las personas que me han compartido su amistad, afecto, sueños, ideas y pensamientos.

ÍNDICE

Resumen…….. 7

1. Introducción……..... 8
   - 1.1. Definición de gen…… 8
   - 1.2. Origen de los genes…….. 9
   - 1.3. Duplicación de genes……. 10
   - 1.4. Mecanismos de duplicación……... 11
   - 1.5. Destinos funcionales y tendencias generales de las duplicaciones……..... 12
   - 1.6. Posibles ventajas de las duplicaciones……... 13
   - 1.7. Virus de RNA……. 14
   - 1.8. Duplicaciones en virus de RNA…….. 19

2. Objetivos e hipótesis…….. 22
   - 2.1. Objetivo general…… 22
   - 2.2. Objetivo particular……...…….. 22
   - 2.3. Hipótesis……. 22

3. Metodología……... 23
   - 3.1. Búsqueda y selección de estructuras cristalográficas de proteínas de virus de RNA……...……...…… 23
   - 3.2. Alineamientos estructurales de las proteínas de virus de RNA……. 23
   - 3.3. Alineamientos de secuencia....…….. 24
   - 3.4. Análisis de escenarios que favorecen la fijación de duplicaciones en virus de RNA……...…..……. 25

4. Resultados y discusión……... 26
   - 4.1. Totiviridae…….. 26
   - 4.2. Cápsides Picornavirales……. 29
   - 4.3. Cisteín-proteasas……..... 37
   - 4.4. Genes homólogos dentro de un mismo genoma y los mecanismos que los generan……...….... 39
   - 4.5. Relación entre la organización y el tamaño del genoma con las duplicaciones…..…..…… 43
   - 4.6. Relación entre ambiente y duplicaciones…….. 46
   - 4.7. Destinos funcionales de los genes duplicados en virus de RNA……...…..... 50
5. Conclusiones…….. 54
6. Agradecimientos…… 56
7. Referencias……. 56

Anexos……..……...…… 63

RESUMEN

La duplicación de genes es un mecanismo importante para el crecimiento de los genomas y el origen de genes en seres vivos. Este evento también se ha descrito con frecuencia en virus de DNA de doble cadena pero no en virus de RNA, cuyo tamaño de genoma podría estar restringido por la alta tasa de mutación. Sin embargo, la búsqueda de genes duplicados en virus de RNA solo se ha llevado a cabo a través de comparaciones de secuencias de aminoácidos. En este trabajo se explora la presencia de genes duplicados en virus de RNA por comparación de estructuras terciarias de proteínas. Los resultados muestran cuatro casos que no habían sido confirmados por comparación de secuencias: i) subunidades α -> β de la proteína KP6 en la familia Totiviridae (dsRNA); ii) proteínas de la cápside VP1 -> VP3 y VP3 -> VP2 en las familias Dicistroviridae, Picornaviridae y Secoviridae, del orden de los Picornavirales [(+)ssRNA]; y iii) proteasas 3C -> 2A en la familia Picornaviridae [(+)ssRNA]. Además, se discuten los posibles mecanismos que generan duplicaciones en virus de RNA y los escenarios en los cuales las duplicaciones pueden ser retenidas.

1. INTRODUCCIÓN

1.1 Definición de gen

El concepto de gen ha cambiado varias veces a lo largo de la historia. Las primeras descripciones concebían a una entidad abstracta responsable de las características de los organismos y su herencia (Gerstein *et al.* 2007). En 1866 Gregor Mendel demostró que, al hibridar plantas, algunas características como el color y la forma de las semillas eran transmitidos de una generación a otra de manera independiente a través de factores de herencia a los que llamó "determinantes ocultos". Sin embargo, sus descubrimientos no fueron debidamente reconocidos hasta que en 1900 Carl Correns, Erich von Tschermak-Seysenegg y Hugo De Vries redescubrieron y repitieron el trabajo de Mendel (Gerstein *et al.* 2007; Rose, 2016). En 1909, Wilhelm Johannsen, basado en el concepto generado por Mendel, acuñó por primera vez la palabra "gen" que deriva del griego *genesis* (nacimiento) o *genos* (origen). Décadas más tarde, gracias a los estudios de Muller, Griffith y McClintock, se pudo demostrar que la herencia tiene una base física y molecular, y que un gen se encuentra asociado a un locus reconocible dentro de un cromosoma. En los años 40's del siglo pasado, Beadle y Tatum descubrieron que las mutaciones podían generar defectos en las rutas metabólicas, con lo cual pudieron hacer la asociación "un gen, una enzima" (Gerstein *et al.* 2007). A mediados del siglo XX, ya establecido el DNA como el material portador de la información hereditaria en la célula y dilucidada la estructura de la doble hélice, se propuso la hipótesis de secuencia y el dogma central de la biología molecular, que en conjunto establecen que la secuencia de un ácido nucleico es un código para la secuencia de aminoácidos de una proteína en particular y que una vez que la información ha pasado a la proteína ya no puede volver a salir (Crick, 1958). Estos enunciados dejaron claro la posibilidad de transferencia de información entre ácidos nucléicos y de ácidos nucléicos a proteínas pero no entre proteínas y de proteínas a ácidos nucléicos (Crick, 1970). El casi inmediato descubrimiento sobre genes que no codifican para proteínas sino para moléculas de RNA funcionales permitió la concepción de un gen como un código basado en ácido nucléico que da lugar a un producto funcional. A finales del siglo XX, con el desarrollo de técnicas de secuenciación y el conocimiento del código genético, el concepto de gen se definió basado en las características de las secuencias genéticas que codifican para una proteína, identificadas

como marcos de lectura abiertos u ORFs, por sus siglas en inglés (open reading frames). En años más recientes, conforme ha aumentado el conocimiento sobre la presencia de secuencias reguladoras, el sobrelapamiento de algunos genes, las variantes de splicing alternativo y trans-splicing (unión de moléculas de mRNA de origen diferente), y los RNA no codificantes, la definición de gen se ha complicado. Algunas intentos de hacer una definición integral abarcan a toda secuencia de ácido nucléico que permite la síntesis de un polipéptido funcional (o RNA), incluyendo promotores y potenciadores. Sin embargo la ocurrencia de genes sobrelapados y de diversos productos de splicing generan la necesidad de definir a un gen como una "unión de secuencias genómicas que codifican para un conjunto de productos funcionales potencialmente sobrelapados". Bajo esta definición, las secuencias que codifican para productos provenientes de un mismo transcrito y que comparten por lo menos una región traducida después del splicing alternativo son considerados como variantes de un mismo gen (Gerstein *et al*. 2007).

1.2 Origen de los genes

Existen diferentes mecanismos por los cuales se pueden originar nuevos genes. Por ejemplo, los genes se pueden formar a partir de genes preexistentes o bien originarse *de novo*. El origen de genes a partir de otros preexistentes implica un aumento en el número de genes homólogos, que son genes originados a partir de un gen ancestral común. Dentro de los mecanismos de origen de genes homólogos se encuentra: (a) la duplicación de secuencias, que es uno de los mecanismos de origen de genes más reportados en seres vivos (Neme & Tautz 2014) y que genera genes parálogos (Olendzenski *et al*. 2005); (b) la transferencia horizontal, que implica un intercambio de genes entre organismos cuya relación no es de progenitor-progenie (Soucy *et al.* 2015) y cuyos homólogos son llamados xenólogos (Olendzenski *et al*. 2005); (c) la fusión de genes, que se puede dar por yuxtaposición de dos genes aledaños por eventos de recombinación, translocación o por inserción de retrotranscritos en la ubicación de otro gen (Kaessmann, 2010), y que genera genes sinólogos (Olendzenski *et al*. 2005); y (d) el splicing alternativo, reportado en virus de RNA (Holmes, 2009), que involucra un procesamiento diferente de los mRNA dando lugar a diversos productos (Matlin *et al.* 2005), aunque algunos autores prefieren considerar a los diversos productos como variantes del mismo gen (Gerstein *et al.* 2007). Adicionalmente, los genes

ortólogos, que son genes derivados de eventos de especiación, pueden llegar a adquirir una nueva función (Olendzenski *et al.* 2005). Los mecanismos *de novo,* que podrían generar genes 'huérfanos', que son aquellos que no contienen homólogos en otras especies (Neme & Tautz, 2014; Andersson *et al.* 2015), incluyen: (i) al sobrelapamiento de genes, que puede darse por un mecanismo llamado "overprinting", el cual involucra la acumulación de mutaciones puntuales que pueden resultar en un gen que cambia el uso de su codón de inicio por el de otro gen adyacente río arriba, la extensión de un gen en otro por la pérdida de un codón de término, o la formación de un nuevo ORF dentro de otro preexistente (Delaye *et al*. 2008); (ii) el "slippage" que implica un desplazamiento de las cadenas complementarias en regiones repetitivas, permitiendo que se agreguen más nucleótidos complementarios (Levinson & Gutman, 1987); y (iii) el origen de genes a partir de secuencias no codificantes como regiones intergénicas o pseudogenes (Neme & Tautz, 2014). Se sabe que los genes huérfanos son comunes en virus (Andersson *et al*. 2015). En cuanto a los genes sobrelapados, es posible que se hayan seleccionado ya que optimizan el uso del material genético disponible en genomas pequeños (Holmes, 2009).

1.3 Duplicación de genes

Como lo ha resumido Gregory (2005), los primeros estudios relacionados con la duplicación de genes se dieron a partir de fenómenos de poliploídia vegetal. Así, en 1929 Stadler reportó que las plantas poliploides eran menos sensibles a las radiaciones X que plantas diploides. Basado en esta evidencia, en 1933 Haldane propuso que la poliploidía podría funcionar como un mecanismo de amortiguación del efecto deletéreo de las mutaciones. En 1934, Blakeslee observó que en plantas, había una asociación entre cambios morfológicos y un aumento en el número de copias de algunas regiones cromosómicas, y consideró que este fenómeno podría ser un mecanismo utilizado por la naturaleza para generar nuevas especies. En 1936 Bridges encontró una asociación similar entre cambios morfológicos y la duplicación de un locus particular en la mosca de la fruta. Con el mismo modelo de estudio, Hermann Muller propuso que las copias redundantes generadas por duplicación podían sufrir mutaciones divergentes que las llevarían a parecer genes sin relación alguna entre sí. En 1938 Serebrovsky propuso que un gen determinante en diferentes aspectos del fenotipo podría ser duplicado y perder funciones o especializarse para una sola función. En la década de los 40's Goldschmidt,

Gulick y Metz realizaron asociaciones explícitas entre la complejidad organísmica y la duplicación de genes, argumentando que las diferencias entre organismos simples y complejos no podían deberse sólo a las mutaciones del mismo conjunto de genes, sino que debía haber adición de nuevos elementos por duplicación de regiones de los cromosomas. De manera similar, en 1951 Stephens sugirió que era necesaria la aparición de nuevos loci, ya sea a partir de regiones no génicas o por duplicación de loci preexistentes (Gregory, 2005). A pesar de que varios autores ya habían destacado el papel de las duplicaciones en los procesos evolutivos, su relevancia quedo mejor establecida a partir de que Ohno afirmara que las duplicaciones son el factor más importante en la evolución (Ohno, 1970). Finalmente, en la era postgenómica surgió el estudio del "paranoma", que consiste en caracterizar por completo al conjunto de genes duplicados (parálogos) de una especie. Si se comparan los paranomas de un conjunto de especies cercanas, se puede datar el origen de las duplicaciones, lo cual permite, en principio, reconstruir un parte de la historia de sus genomas. Al estimar el tiempo desde una duplicación, también se puede estimar la tasa de nacimiento y muerte de genes. Estas nuevas aproximaciones nos permiten un mejor entendimiento de la ocurrencia y consecuencias funcionales de las duplicaciones (Gregory, 2005).

1.4 Mecanismos de duplicación

Existen diferentes mecanismos por los cuales se puede dar la duplicación de uno o varios genes, o de regiones del genoma. En eucariontes, los eventos de duplicación que involucran a una mayor cantidad de genes son las duplicaciones de genoma completo, que suelen darse por poliploidía autopoliploide (no disyunción) o alopoliploide (hibridación). Las duplicaciones también pueden darse de manera similar pero a nivel de cromosoma, en donde la no disyunción de los cromosomas homólogos durante la meiosis genera cambios (en este caso aumento) en el número de cromosomas, que constituye el fenómeno conocido como aneuploidía. A nivel de gen pueden darse duplicaciones de dos maneras diferentes, ya sea por transposición replicativa o por retrotransposición, que colocan copias de un gen en diferentes regiones del genoma, o por recombinación desigual, que genera duplicaciones contiguas o en tándem (Gregory, 2005). Por último, existen las duplicaciones internas, en donde una región, motivo o dominio de una proteína se duplica y termina estando aledaño a su origen por un proceso de recombinación, lo que resulta en una proteína elongada (Barker *et al*. 1978;

Nacher *et al*. 2010). En bacterias, se han propuesto otros mecanismos como: i) la recombinación desigual durante la replicación (Romero & Palacios, 1997; Andersson & Hughes, 2009; Andersson *et al.* 2015), que genera duplicados en tándem; ii) la escisión y reinserción de fragmentos circulares, que puede agregar duplicados en diferentes regiones del genoma (Romero & Palacios, 1997; Johnson & Grossman, 2015); y iii) la replicación por círculo rodante, que puede aumentar drásticamente el número de copias aledañas (Romero & Palacios, 1997; Andersson & Hughes, 2009; Andersson *et al.* 2015).

1.5 Destinos funcionales y tendencias generales de las duplicaciones

Luego de la duplicación, pueden darse diferentes desenlaces funcionales como (Figura 1): (a) la pseudogenización, en donde el duplicado pierde su función por completo; (b) la subfuncionalización, en la cual después de la duplicación, el gen va sufriendo mutaciones hasta quedarse con un subconjunto de sus funciones originales; (c) la neofuncionalización, que implica la adquisición de una función completamente nueva por parte del gen duplicado; y por último, (d) la redundancia, en donde el duplicado conserva la función original del gen a partir del cual se duplicó (Gregory, 2005). Una de las tendencias más generales sobre las duplicaciones es la correlación que existe entre el tamaño del genoma y el número de duplicados (Hughes *et al.* 2005). Particularmente se ha observado que el tamaño de los genomas, como el número de genes y la vida media de los duplicados incrementa de procariontes a eucariontes, y de organismos unicelulares a multicelulares (Lynch & Conery, 2003). Una de las explicaciones tiene que ver con el tamaño efectivo de las poblaciones, que disminuye conforme aumenta la masa de los individuos. En este sentido, los organismos más complejos forman poblaciones pequeñas y con poca variación (Lynch & Conery, 2003), en donde un amplio repertorio funcional les permitiría una mejor resistencia a los cambios ambientales. Por otro lado, se ha sugerido que en eucariontes, la multicelularidad permite la retención de genes duplicados de manera subfuncional para su expresión diferencial en tejidos específicos (Zhang, 2003) o que la diploidía podría facilitar eventos de recombinación desigual (Xue *et al.* 2010). Adicionalmente, se ha sugerido que en bacterias, el genoma circular podría imponer un límite superior para el crecimiento de los mismos (Hughes & Friedman, 2010) o que el incremento en el tamaño del genoma podría reducir la velocidad de replicación en genomas con un solo inicio de la replicación (Wagner, 2010). Otra hipótesis

que podría explicar algunas tendencias sobre las duplicaciones es la de balance de dosis, que postula que la duplicación de genes con poca importancia en redes de interacciones y sin efecto deletéreo al desactivarlos genera un menor efecto de dosis que si se duplica un gen esencial (Conant & Wolfe, 2008).

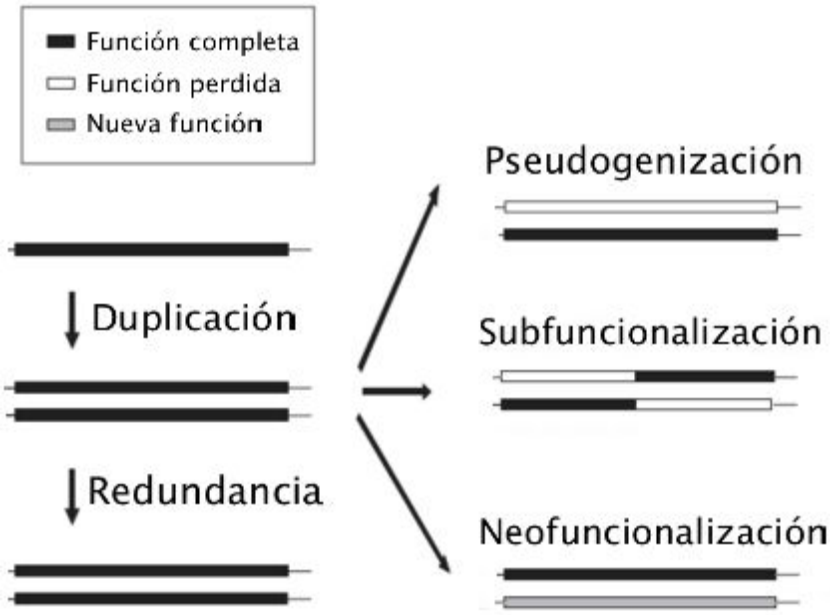


Figura 1. Destinos funcionales de las duplicaciones. Modificada de Gregory (2005).

1.6 Posibles ventajas de las duplicaciones

Las duplicaciones podrían conservarse por diferentes razones. De manera similar a como lo sugirió Haldane en 1933 respecto a la poliploidía, es posible que las duplicaciones que resultan en redundancia sirvan como amortiguador de los efectos deletéreos de las mutaciones (Gregory, 2005). En este caso, la duplicación redundante permitiría la acumulación de mutaciones neutrales en una de las copias o incluso en ambas. Una de las consecuencias más comunes suele ser la pseudogenización (Anderson *et al.* 2015). Sin embargo, también se ha propuesto que sólo gracias a que la selección natural no actúa sobre las mutaciones que se acumulan en el gen duplicado, es que este puede adquirir una nueva función (Ohno, 1970). Otra posibilidad viene dada por el modelo de duplicación-degeneración-complementación (DDC), que sugiere que ambas copias pueden sufrir mutaciones degenerativas hasta lograr una especialización y complementación

funcional (Anderson *et al.* 2015). Otros modelos también proponen que la redundancia puede ser conservada por selección del incremento en la tasa de expresión de algún producto ventajoso (Gregory, 2005; Anderson *et al.* 2015), o como mecanismo de respaldo en circuitos de regulación (Kafri & Pilpel, 2010). Por otro lado, es posible que las duplicaciones contribuyan para la rápida especiación. Un ejemplo son los peces cíclidos africanos, cuya rápida divergencia se atribuye a la gran cantidad de duplicaciones presentes en sus genomas. Estas duplicaciones pudieron aumentar el repertorio funcional de los organismos, permitiéndoles adaptarse rápidamente en ambientes cambiantes. Finalmente, en el caso de las duplicaciones internas, el incremento en el número de dominios duplicados puede mejorar la afinidad y especificidad de la proteína o crear una nueva función (Chen *et al*. 2007).

1.7 Virus de RNA

La clasificación de Baltimore agrupa a los virus, tanto de DNA como de RNA, en siete grupos diferentes basados en la composición y mecanismos de procesamiento del genoma para la replicación del virus (Holmes, 2009). Junto con los viroides, los virus de RNA son los únicos entes biológicos que poseen genomas de RNA. Los genomas de virus de RNA pueden ser de cadena doble (grupo III), cadena sencilla, positiva o negativa (grupo IV y V, respectivamente), o de cadena sencilla retrotranscrita (grupo VI) (Figura 2). 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 luego es integrado en el genoma del hospedero para ser finalmente transcrito por la maquinaria celular (Holmes, 2009). En el caso de los virus 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 doble cadena (Reanney, 1982).

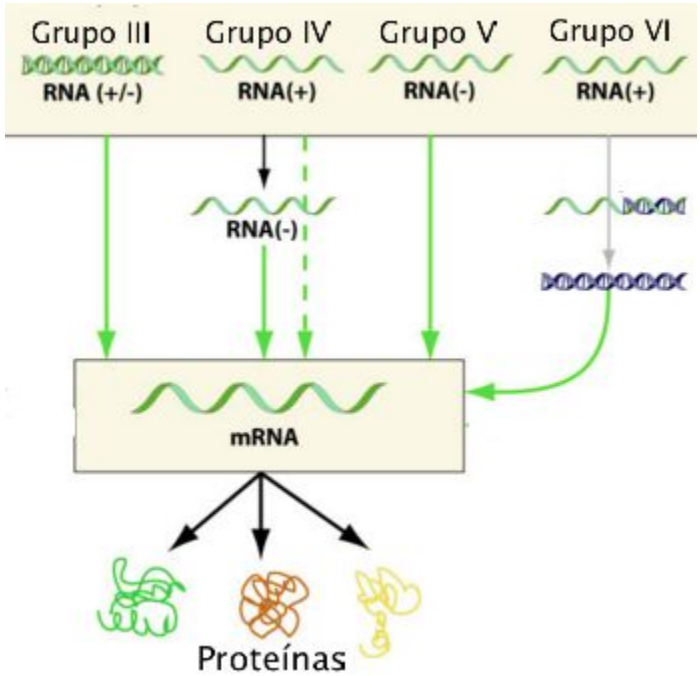


Figura 2. Virus de RNA en la clasificación de Baltimore. Las flechas continuas de color verde indican que el genoma debe ser transcrito a mRNA para que pueda ser traducido. La flecha continua de color gris indica que el genoma es retrotranscrito. La flecha punteada de color verde indica que el genoma de RNA puede funcionar como mRNA. Modificado de ViralZone www.expasy.org/viralzone, Swiss Institute of Bioinformatics.

Los virus de RNA presentan los genomas más acotados de tamaño de todos los virus, y están en un rango de longitud de 1.8kb hasta 33.452kb (Campillo-Balderas *et al.* 2015), y con un promedio de 10kb. La selección de genomas cortos podría estar relacionada con la alta tasa de mutación, de aproximadamente una sustitución cada mil nucleótidos, pues en ausencia de mecanismos de edición, entre más largo sea un genoma, es más propenso a acumular mutaciones deletéreas que uno corto (Holmes, 2009). La presencia de genomas cortos también podría contribuir a una rápida replicación, la cual, aunque existe una relación inversamente proporcional entre la tasa de replicación y la fidelidad de la misma, podría funcionar como un mecanismo para reducir el efecto deletéreo de las mutaciones, pues es posible que los virus sean capaces de generar más descendencia de la que es eliminada por selección purificadora (Holmes, 2009). Consecuentemente, la rápida replicación produce genomas cortos, tal y como sucedió en el experimento clásico de Spiegelman con la polimeraza del bacteriofago Qβ. En este experimento se realizaron transferencias seriales en las cuales los intervalos de síntesis entre transferencias fueron reducidos para seleccionar a las primeras moléculas completadas. El resultado fue un aumento en la tasa de multiplicación

y un producto que para la septuagésima cuarta transferencia había perdido el 83% de su genoma original (Mills *et al.* 1967). Sumado a esto, se ha sugerido que el tamaño reducido de estos genomas no les permitiría albergar genes que codifiquen para proteínas de corrección del material genético o de apertura de dobles hélices. La incapacidad de abrir dobles hélices podría actuar como otra presión a favor de genomas cortos (Reanney, 1982). Dadas las características de los genomas de RNA, es tentador sugerir que el tamaño de estos se encuentra limitado por su alta tasa de mutación. Sin embargo, existen otras estrategias, además de los genomas cortos, que permiten reducir el efecto deletéreo de las mutaciones.

Algunas de estas estrategias, que además podrían promover el aumento en el tamaño del genoma, son la recombinación por "copy-choice" y la ganancia de segmentos homólogos durante el reordenamiento (Figura 3). Por un lado, el reordenamiento sucede exclusivamente en genomas segmentados (Figura 3C). Este proceso implica la coinfección de dos virus y el posterior empacamiento de segmentos de diferente ancestría en una misma cápside. Se ha sugerido que este proceso puede reducir los efectos deletéreos de la acumulación de mutaciones a través del recambio de segmentos mutados por otros no mutados. Es posible que los genomas segmentados se hayan originado por un mecanismo similar, a partir de la coinfección de dos virus con segmentos individuales que se empaquetaron en una misma cápside, formando una relación de complementación funcional (Holmes, 2009). Incluso, algunos autores proponen que la segmentación de genomas pudo haber surgido por eventos de duplicación de genoma completo y la encapsidación de segmentos similares (Shirogane *et al.* 2013).

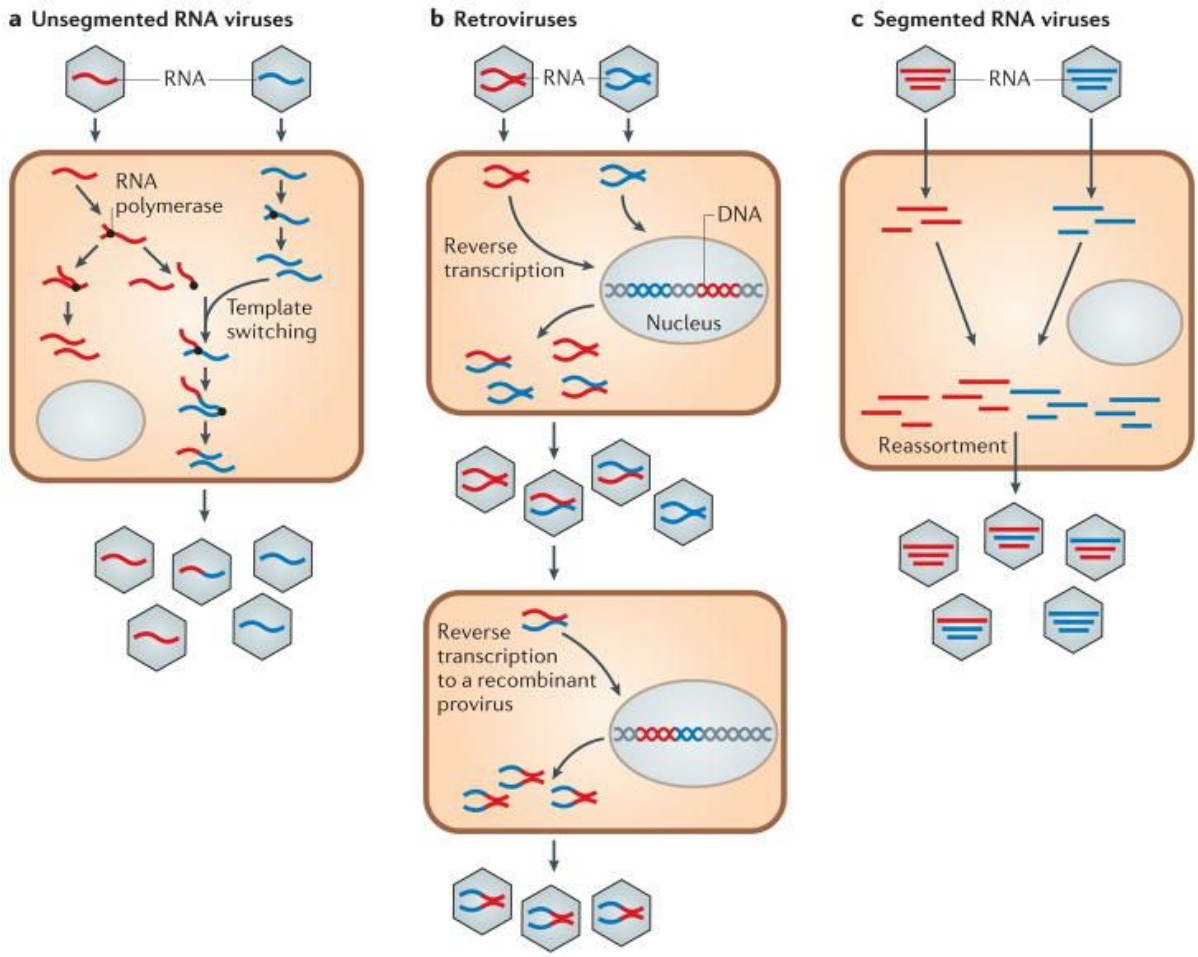


Figura 3. (a) Recombinación por 'copy choice' en virus de RNA no segmentados. (b) Recombinación por 'copy choice' mediado por la transcriptasa reversa en retrovirus 'heterocigos'. (c) Reordenamiento de segmentos. Tomada de Simon-Loriere & Holmes (2012).

Por otro lado, la recombinación por "copy-choice" es común a las cuatro clases de virus de RNA, con genomas segmentados o no segmentados, siendo más frecuente en retrovirus y menos en virus (-)ssRNA (Holmes, 2009). La recombinación por 'copy choice' se da cuando la polimerasa salta de molde durante la síntesis de la cadena complementaria, generando una molécula de RNA con una ancestría combinada, ya sea a partir de secuencias homólogas o no homólogas (Figura 4) (Simon-Loriere & Holmes, 2012). Este proceso puede generar cadenas nacientes con menor cantidad de mutaciones acumuladas. Incluso es posible que, siendo la recombinación uno de los mecanismos más frecuentes para la duplicación de genes, ésta también genere duplicaciones de genes en virus de RNA (Lai, 1992; Agol, 1997). De hecho, se ha observado que después de la recombinación, los genomas virales suelen ser más largos

debido a que se generan duplicaciones en el sitio de la recombinación. Sin embargo, también se ha visto que estos virus pierden adecuación, de modo que las variantes que pierden la región duplicada son seleccionadas (Lowry *et al.* 2014). Cuando la polimerasa y la cadena naciente se disocian de la cadena molde, posiblemente por formación de estructuras secundarias en el RNA, pueden reasociarse sobre moldes provenientes de la misma cepa (intratípicamente), lo cual resultaría en un transcrito ordinario, o en la misma posición sobre moldes provenientes de cepas diferentes (intertípicamente), lo cual resultaría en una recombinación homóloga. Otra posibilidad es la recombinación homóloga aberrante, que sucede cuando la reasociación se da en una posición diferente. Si la recombinación homóloga aberrante se da de manera intratípica podría generarse una duplicación, pero si sucede de manera intertípica el resultado podría ser una transferencia horizontal de genes homólogos. Otro posible resultado de este tipo de recombinación son las deleciones. Además, la reasociación puede darse con moldes no homólogos (recombinación no homóloga o ilegítima) como transcritos celulares, lo cual resultaría en transferencia horizontal de genes no homólogos (Lai, 1992; Simon-Loriere & Holmes, 2012).

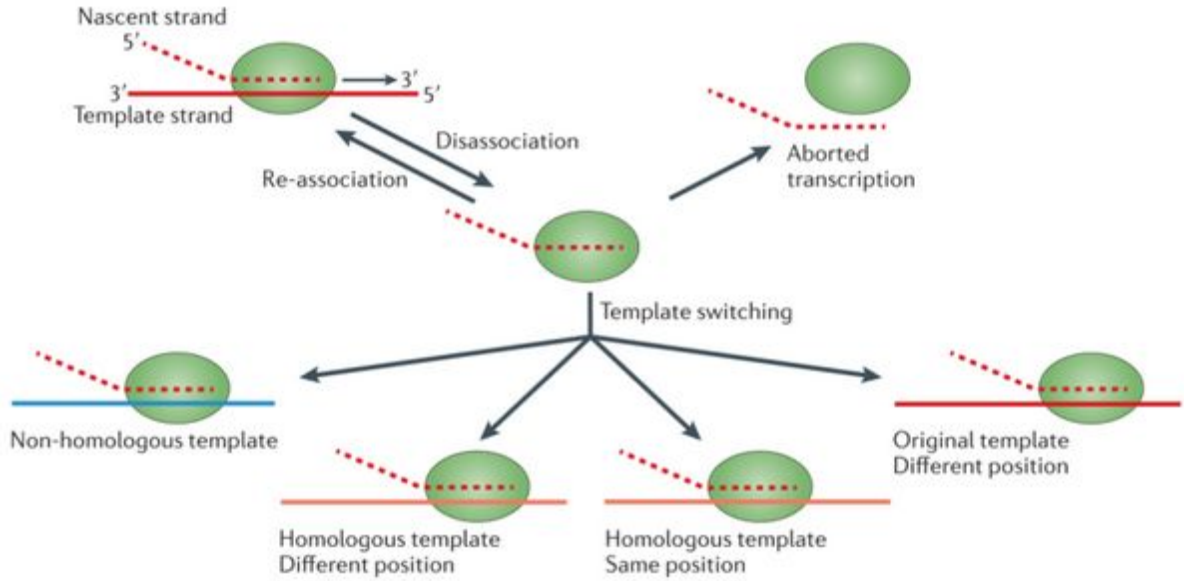


Figura 4. Posibles moldes y posiciones en las cuales se puede reasociar la polimerasa y la cadena naciente durante un proceso recombinación por 'copy choice'. Tomado de Simon-loriere & Holmes (2012).

1.8 Duplicaciones en virus de RNA

A pesar de que las duplicaciones parecen ser un evento frecuente en la evolución de virus de DNA de doble cadena (Shackelton & Holmes, 2004), las duplicaciones han sido poco reportadas en virus de RNA. Simon-Loriere y Holmes (2013) reportaron solo nueve casos significativos a nivel de estructura primaria, de los cuales dos fueron descritos por primera vez (Tabla 1).

Tabla 1. Genes duplicados en virus de RNA. Modificado de Simon-Loriere & Holmes (2013).

Organización del genoma	**Familia**	**Gen duplicado**	**Función**	**Publicación**
(+)ssRNA	Closteroviridae	CP -> CPm	Ensamblaje de la cápside	Boyko *et al*. 1992; Fazeli *et al*. 2000; Kreuze *et al*. 2002; Tzanetakis *et al*. 2005; Tzanetakis *et al*. 2007; Simon-Loriere & Holmes, 2013;
		CPm1 -> CPm2	Ensamblaje de la cápside	Fazeli *et al*. 2000
	Picornaviridae	Vpg -> Vpg	Inicio de la replicación	Forss & Schaller, 1982
	Benyvirus*	**p25 -> p26**	**Factor de patogenicidad**	**Simon-Loriere & Holmes, 2013**
	Flaviviridae^	V3 (NS5A)	Unión a RNA/inhibición de apoptosis	Le Guillou-Guillemette *et al*. 2015
(-)ssRNA	Rhabdoviridae	G -> Gns	Reconocimiento celular y endocitosis	Walker *et al*. 1992; Gubala *et al*. 2010; Blasdel *et al*. 2012
		U1 -> U2	**Desconocida**	**Simon-Loriere & Holmes, 2013**
Retrovirus	Retroviridae	orf A -> orf B	Proliferación celular y regulación de la transcripción viral	LaPierre *et al*. 1999
		orf 1 -> orf 2	Desconocida	Kambol *et al*. 2003
		vpr -> vpx	Supresión del ciclo celular y translocación del complejo de preintgración	Tristem *et al.* 1990

*Género no ha asignado a una familia. ^Posterior al trabajo de Simon-Loriere & Holmes (2013). Los casos descritos por primera vez en Simon-Loriere & Holmes (2013) están resaltados en negritas.

Los dos casos descritos por primera vez en el estudio de Simon-Loriere y Holmes (2013) corresponden al factor de patogenicidad p25 -> p26 de un Benyvirus (grupo IV) y a U1 -> U2, de función desconocida, de la familia Rhabdoviridae (grupo V). Otros estudios también han sugerido que la proteína U3, que es esencial para replicación viral, es paráloga a U1 y U2 (Allison *et al*. 2014). La mayoría de las duplicaciones se ubican en genes adyacentes, lo que sugiere que se trata en efecto de duplicaciones y no de reclutamiento de secuencias. Solamente p25 y p26 de los Benyvirus se encuentran en segmentos diferentes (Simon-Loriere & Holmes, 2013). Entre los casos reportados con anterioridad por otros autores se encuentran: (a) tres casos en virus de cadena sencilla positiva: proteínas de cápside CP -> CPm y CPm1 -> CPm2 de la familia Closteroviridae (Boyko *et al.* 1992; Fazeli & Rezaian, 2000; Kreuze *et al*. 2002; Tzanetakis *et al.* 2005; Tzanetakis & Martin, 2007); y una proteína viral con función de "primer" Vpg -> Vpg de la familia Picornaviridae (Forss & Schaller, 1982); (b) un caso en virus de cadena sencilla negativa: glicoproteína G -> Gns, de la familia Rhabdoviridae (Walker *et al*. 1992; Gubala *et al.* 2010; Blasdell *et al.* 2012); y (c) tres casos en virus de transcripción reversa de la familia Retroviridae: (i) proteína supresora de la fase G2 de la mitosis y promotora de la translocación nuclear del complejo de preintegración, Vpr -> proteina Vpx encargada de la translocación nuclear del complejo de preintegración (Tristem *et al*. 1990; Fletcher *et al.* 1996); (ii) proteína de proliferación celular y reguladora de la transcripción orfA -> orfB, homóloga a ciclina D celular (LaPierre *et al*. 1999); y (iii) orf1 -> orf2 de función desconocida (Kambol *et al*. 2003). Estos casos fueron aceptados con identidades desde 21% siempre y cuando su valor E de BLAST fuese menor a 10^{-5}.

Un caso que no fue detectado por Simon-Loriere y Holmes (2013) es el de las proteasas L1-Pro y L2-Pro del *Citrus tristeza virus* (CTV), de la familia Closteroviridae. Estas son dos cisteín-proteasas del tipo de la papaína, cuya posible relación paráloga ha sido sugerida tanto por análisis de comparación de secuencias (Karasev *et al*. 1995) como por un análisis filogenético (Peng *et al*. 2001). Recientemente se reportó un caso de duplicación del dominio V3 de la proteína NS5A del *Hepatitis C virus*, de la familia Flaviviridae [(+)ssRNA] (Tabla 1). En este se encontraron segmentos de 12 a 31 aminoácidos con identidad de 50 a 91% con el dominio V3, y una posible relación con el desarrollo de cirrosis hepática (Le Guillou-Guillemette *et al*. 2015). Algunos casos de duplicación también han sido reportados en regiones no traducidas o UTR (Holmes, 2009).

Los análisis que se han hecho sobre duplicación de genes en virus de RNA se han basado únicamente en comparación de estructuras primarias, y la poca frecuencia con la que se reportan estos casos podría ser consistente con el hecho de que el crecimiento de los genomas de RNA se encuentra limitado por su alta tasa de mutación. Sin embargo, es posible que las duplicaciones también se encuentren conservadas en los genomas de RNA y que debido a su rápida divergencia no podamos notar las homologías a nivel de secuencia. De ser así, un análisis basado en comparación de estructuras terciarias, que conserva mejor la señal de homología, podría mostrar casos de duplicación que se han escapado a la resolución de otras metodologías.

2. OBJETIVOS E HIPÓTESIS

2.1 Objetivo general

Investigar si los genomas de virus de RNA son capaces de conservar genes originados por duplicación.

2.2 Objetivo particular

Explorar en qué escenarios se puede favorecer la fijación de las duplicaciones en virus de RNA.

2.3 Hipótesis

Los genomas de RNA son capaces preservar secuencias originadas por duplicación cuando el beneficio de poseer genes duplicados es mayor al costo de tener un genoma largo.

3. METODOLOGÍA

3.1 Búsqueda y selección de estructuras cristalográficas de proteínas de virus de RNA

Para determinar la presencia de genes duplicados en genomas de virus de RNA se buscaron a todas las proteínas de virus de RNA con una estructura determinada en la base de datos de proteínas (PDB) www.rcsb.org, con el fin de compararlas para inferir homología por similitud estructural. Para evitar la comparación de estructuras determinadas por técnicas diferentes, la búsqueda de estructuras se limitó a aquellas determinadas por difracción de rayos-X, que es la metodología que se usó para aproximadamente el 86% de las estructuras en la PDB (Sikic *et al.* 2010). La búsqueda de las estructuras se realizó por grupo taxonómico al nivel de familia viral. Los filtros principales para seleccionar a las estructuras representantes de cada proteína fueron: i) el de resolución máxima de 3 Å (1 Angstrom = 10^{-10} metros. Valores menores de resolución se asocian a una determinación más precisa de la posición de los átomos); y ii) el de reducción de redundancia, que muestra solo a las estructuras representantes (mejor resolución y preferentemente determinadas por rayos-x) de conjuntos de proteínas con 100% de identidad de secuencia. Para reducir la redundancia restante las estructuras fueron clasificadas por tipo viral y se seleccionó a aquellas que presentaron una mayor longitud de cadena de aminoácidos, menor número de mutaciones, menor número de ligandos y mejor resolución.

3.2 Alineamientos estructurales de las proteínas de virus de RNA

Antes de realizar los alineamientos, los archivos en formato PDB de las estructuras que contienen varias cadenas, ya sea de la misma o de diferentes proteínas, fueron editados. En el caso de las proteínas que forman homómeros se conservó una sola cadena. En el caso de las proteínas que forman heterómeros el archivo fue separado en dos para poder comparar a las proteínas. Los alineamientos se realizaron a nivel de familia utilizando el alineador flexible de FATCAT, herramienta para la cual tenemos disponible un procedimiento automatizado. Este programa genera valores de la raíz de la desviación cuadrática media (RMSD) (en Å), entre los carbonos alfa de las estructuras proteicas, similares a los de los métodos rígidos cuando no se requieren rotaciones, pero cuando sí necesitan, genera mejores RMSD entre

estructuras flexibles que los programas de alineamiento rígido, introduciendo menos rotaciones que otros alineadores flexibles (Ye & Godzik, 2003). El criterio principal para seleccionar alineamientos significativos fue el valor P, que es la probabilidad de observar un mejor score (recompensado por una mayor longitud del alineamiento y un menor RMSD, y penalizado por un mayor número de rotaciones para lograr el alineamiento) entre estructuras elegidas al azar. Debido a que P suele tener valores entre 0 y 1e-06 para proteínas ortólogas (e.g. cápsides de los Picornaviridae), y a que algunas proteínas homólogas distantes (e.g. RNasaH e Integrasa de los Retroviridae) tienen valores que oscilan alrededor de 1e-03, se designó como límite para determinar alineamientos significativos a un valor de P = 1e-05, con el fin de evitar falsos positivos. Otros criterios considerados fueron optLength (longitud del alineamiento flexible), LengthA y LengthB (longitud de las proteínas alineadas), TN (número de rotaciones), optRMSD (desviación cuadrática media flexible) y chRMSD (desviación cuadrática media rígida). Para confirmar las relaciones sugeridas por el alineador estructural flexible, se realizaron alineamientos rígidos con DALI (http://ekhidna.biocenter.helsinki.fi/dali_lite/start). En este se tomaron en cuenta el RMSD y Z, que es una estimación de la significancia de S, una puntuación de similitud aditiva basada en el número de residuos alineados (Holm & Rosenström, 2010). A partir de Z = 2 suele considerarse que las proteínas tienen estructuras similares. Todos los alineamientos estructurales fueron visualizados en jmol.

3.3 Alineamientos de secuencia

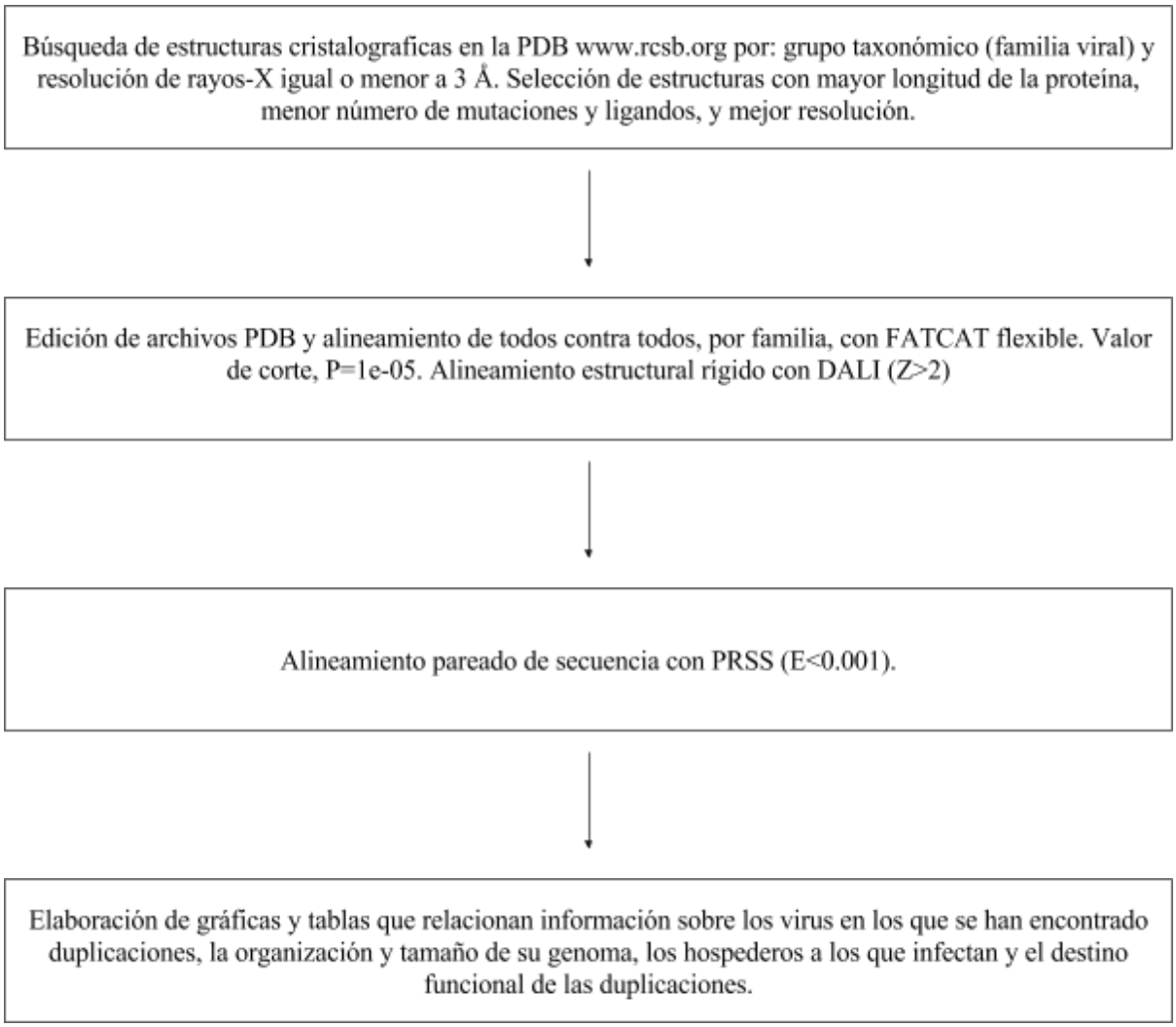
Finalmente, con el programa PRSS del servidor en línea de FASTA (http://fasta.bioch.virginia.edu/fasta_www2/fasta_www.cgi?rm=shuffle&pgm=rss), se realizaron alineamientos de secuencia solamente en los casos de similitud estructural significativa identificados por los alineamientos estructurales. Este programa permite obtener un valor de significancia para un alineamiento pareado sin depender de una base de datos, sirviéndose para ello de secuencias generadas al azar a partir del barajeo de la segunda secuencia. Se calcula así la probabilidad de obtener el mismo score del alineamiento inicial por azar. Para este se reporta identidad, similitud, cobertura y P generada con 200 barajeos por alineamiento. Normalmente se puede inferir homología de secuencia cuando E() < 0.001.

Este análisis sirvió para verificar si a nivel de secuencia se podía reconocer alguna señal de homología.

3.4 Análisis de escenarios que favorecen la fijación de duplicaciones en virus de RNA

Para contestar qué escenarios favorecen la fijación de duplicaciones en virus de RNA, se realizaron gráficas y tablas de datos que relacionan información, basada en Simon-Loriere y Holmes (2013) y este trabajo, sobre los virus en los que se han encontrado duplicaciones, la organización y el tamaño del genoma de estos virus, el hospedero al que infectan y el destino funcional de los genes duplicados.

La metodología se encuentra resumida en el siguiente esquema:



4. RESULTADOS Y DISCUSIÓN

La búsqueda de estructuras cristalográficas dio como resultado un total de 361 entradas de PDB, correspondientes a 170 tipos virales dentro de 30 familias de virus de RNA (Tabla 2). El número de entradas de PDB no equivale al número de proteínas, pues algunas estructuras contienen más de una proteína en forma de heterómeros.

Tabla 2. Descripción de los datos analizados (ver Anexo 1 para información desglosada).

Clasificación	Entradas de PDB	Tipos virales	Familias virales muestreadas	Proporción de familias virales muestreadas	Alineamientos significativos entre estructuras de proteínas virales
(+)ssRNA	174	80	18	0.47	7
(-)ssRNA	84	41	7	0.7	0
(RT)ssRNA	54	19	1	1	0
dsRNA	49	30	4	0.4	1
TOTAL	361	170	30	0.52	8

Los casos reportados por Simon-Loriere y Holmes (2013) no pudieron ser confirmados por comparación de estructura terciaria debido a que sus estructuras no están disponibles en la PDB o no se encontraron bajo los criterios metodológicos de este trabajo.

4.1 Totiviridae

Una pequeña proporción de hongos del maíz de la especie *Ustilago maydis* conviven de manera simbiótica con el virus de doble cadena de RNA *Ustilago maydis virus* (UmV). Este virus simbionte, que solo se transmite de célula a célula por mitosis o meiosis, produce proteínas citotóxicas como la KP6. El hongo hospedero, que posee genes de resistencia específicos para cada toxina, es capaz de secretar proteínas que aniquilan a otras variantes de *U. maydis* que no son resistentes. La proteína KP6 está compuesta de dos subunidades, la KP6α y KP6β. KP6α tiene función de reconocimiento, mientras que KP6β es citotóxica. Allen y colaboradores (2013) reportaron la estructura tridimensional del heterodímero KP6α y KP6β, destacando la similitud estructural entre ambas subunidades (Tabla 3). La estructura de ambas subunidades consta de cuatro hebras β y un par de α hélices antiparalelas, formando un sandwich α/β. La mayor diferencia radica en la presencia de una hélice en el extremo

N-terminal de KP6α, que está ausente en KP6β. Además, en KP6β se puede apreciar una región sin densidad electrónica, que corresponde a un *loop* desordenado de cuatro aminoácidos que se encuentra adelante de la primera α hélice (Figura 5A y B). De acuerdo con Allen *et al*. (2013), la presencia de la hélice N-terminal de KP6α es bastante inusual por estar tan expuesta, ya que se compone principalmente de aminoácidos hidrofóbicos que podrían estar involucrados en el reconocimiento celular a través de interacciones con proteínas de membrana. Es posible que luego de darse el reconocimiento, KP6β ingrese a la célula o interactúe con proteínas de membrana causando lisis celular, ya sea por apertura o cerrado de canales iónicos. La similitud estructural de estas subunidades puede apreciarse a simple vista y es confirmada tanto por el alineamiento flexible como por el rígido (Figura 5C y D).

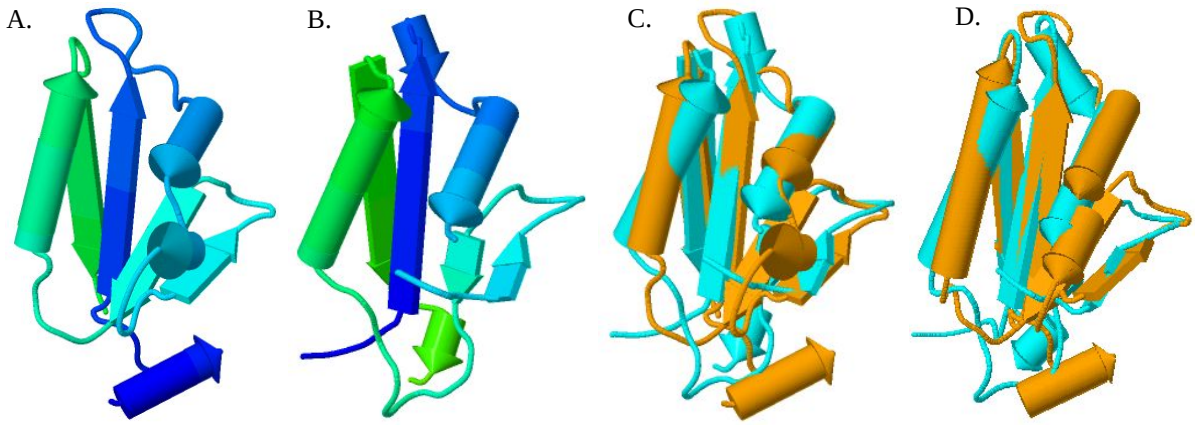


Figura 5. Alineamiento estructural de la subunidad alfa beta de la proteína KP6 (4GVB A y B). (A) KP6α coloreada del N-terminal (azul) al C-terminal (verde). (B) KP6β coloreada del N-terminal (azul) al C-terminal (verde). (C) Alineamiento flexible entre KP6α (anaranjado) y KP6β (cyan). P = 2.16e-07; OptLength = 62; LengthA = 77; LengthB = 74; TN = 0; OptRMSD = 1.52; chRMSD = 1.41. (D) Alineamiento rígido entre KP6α (anaranjado) y KP6β (cyan). RMSD = 1.6; Z = 8.5.

Los valores generados por ambos métodos de alineamiento estructural son significativos (P = 2.16e-07 y Z = 8.5). Por otro lado, a nivel de secuencia se obtuvo una identidad no significativa del 36.8%, similitud de 57.9%, cobertura de 70.37% y E(200) = 1.1 (ver Anexo 2 para las regiones alineadas por comparación de secuencia). Los valores de RMSD obtenidos en este estudio son similares a los reportados por Allen *et al*. (2013) (Tabla 3).

Tabla 3. Comparación con los valores obtenidos por Allen *et al.* (2013) para el alineamiento estructural y de secuencia entre KPα y KP6β.

Trabajo	Programa	RMSD	Identidad
Allen *et al.* 2013	Chimera	1.5	25%
Este estudio	FATCAT/DALI	1.52/1.6	36.8%

Estas subunidades son codificadas a partir de un mismo transcrito que da lugar a un polipéptido que presenta una secuencia de 31 aminoácidos que une el extremo C-terminal de KP6α con el N-terminal de KP6β. Después de ser traducida, una proteasa remueve estos 31 aminoácidos, de modo tal que se separan las subunidades, quedando los extremos C-terminal de KP6α y el N-terminal de KP6β en posiciones opuestas (Allen *et al.* 2013). Es interesante que la secuencia que codifica para estas subunidades no forma parte del genoma que codifica para las principales proteínas virales, sino que se ubica en un dsRNA satélite.

Es difícil saber cuál es la procedencia de estas proteínas y en qué momento se dio su duplicación. A nivel de estructura terciaria se han encontrado similitudes con pequeños dominios con función de interacción proteína-proteína dentro de proteínas eucariontes. En ninguna de las estructuras similares se ha encontrado un heterodímero como el de KP6 (Allen *et al.* 2013). Debido ello, es posible que un fragmento de algún transcrito celular haya sido encapsidado en forma de segmento de RNA al interior de un Totivirus, y que luego se haya dado la duplicación adyacente, formando el heterodímero KP6. Este parece ser un ejemplo de duplicación de dominios. Aunque las poliproteínas en virus son comunes, esto parece más bien una proteína compuesta de dos dominios homólogos que después son separados por una proteasa. El resultado estructural de esta duplicación en tandem es parecida a la formación de "dominios pseudodiméricos", en donde una proteína que normalmente forma dímeros es duplicada, dando lugar a un dímero fusionado que puede ser funcional si la región que conecta a ambos dominios es lo suficientemente larga para que las subunidades alcancen la orientación correcta (Taylor & Sadowski, 2010). Es posible que la remoción de la región que conecta a estas subunidades resuelva el impedimento estructural que impone su longitud para la formación del dímero, que en este caso es formado por dos macromoléculas distintas (heterodímero).

A pesar de que Allen y colaboradores (2013) concluyeron de manera textual que “las estructuras de ambas subunidades son ‘más homólogas’ de lo implicado por su similitud de secuencia” (Allen *et al*. 2013), no se hizo explícito que esta similitud estructural puede representar un caso de duplicación.

4.2 Cápsides Picornavirales

Las cápsides de la mayoría de los virus pertenecientes al orden de los Picornavirales se ensamblan a partir de heterotrímeros de proteínas VP1, VP2 y VP3. Estas proteínas se pudieron haber originado a partir de dos duplicaciones consecutivas ocurridas en el ancestro y conservadas en todas las familias pertenecientes a este orden. Esta relación ha sido propuesta (Liljas *et al.* 2002) pero no confirmada por alineamientos estructurales. El análisis realizado confirma esta relación en el orden de los Picornavirales a partir de los alineamientos generados entre las proteínas de la cápside de las familias Dicistroviridae, Picornaviridae y Secoviridae (Figura 6, 7 y 8).

En los Dicistroviridae, tanto el alineamiento flexible como el rígido arrojan valores típicos de un alineamiento significativo. Además, la topología de las proteínas es muy parecida, y consisten principalmente en un barril beta de ocho hebras antiparalelas de tipo *jelly roll* y una cola N-terminal casi idéntica que forma contactos entre subunidades y protómeros (Liljas *et al.* 2002). Por otro lado, VP3 contiene una hélice alfa adicional en el extremo C-terminal (Figura 6C), cuya conexión se entrelaza con la conección entre el barril beta y el extremo C-terminal de VP1. Además, esta hélice podría tener interacciones polares con VP2 (Liljas *et al*. 2002).

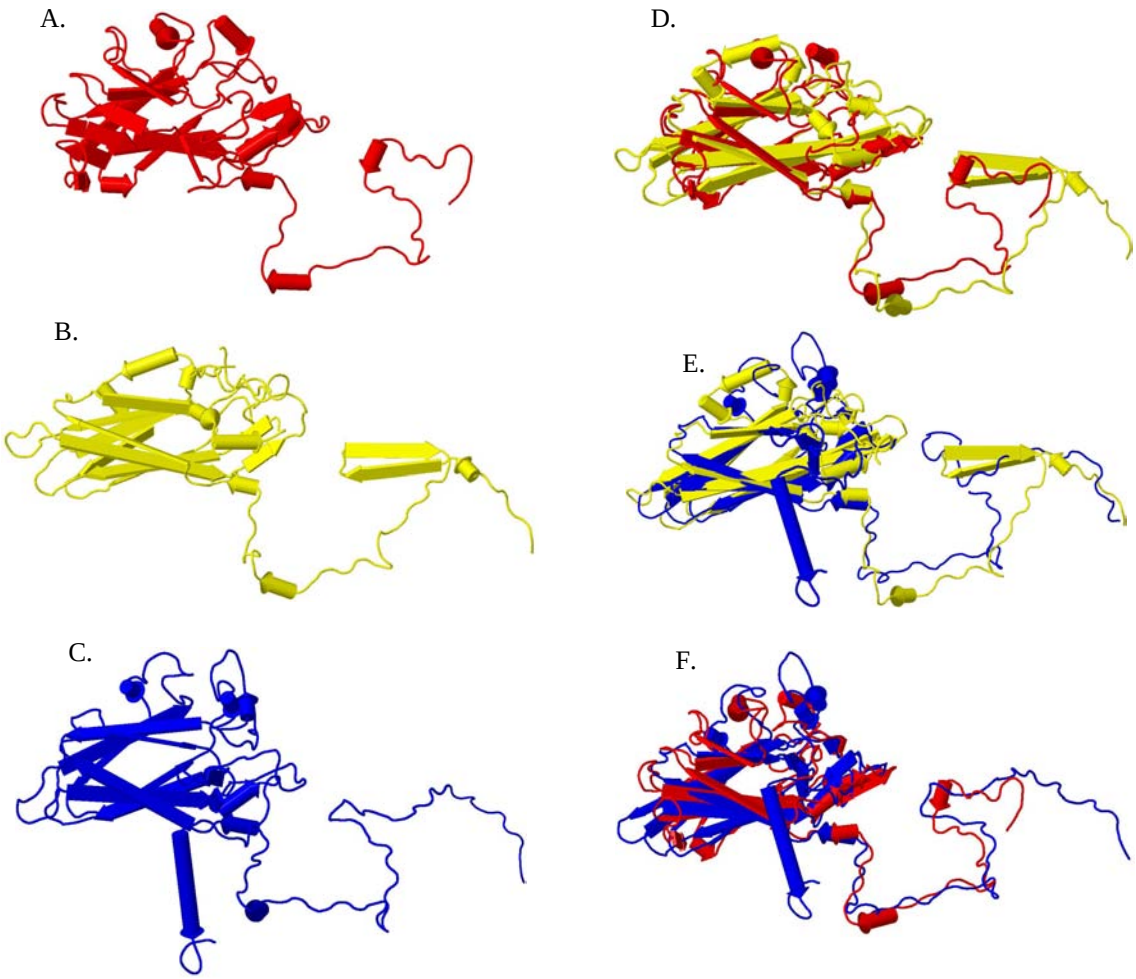


Figura 6. Proteínas de la cápside VP1, VP2 y VP3 de Dicistroviridae (1B35 A, B y C). (A, B y C) VP1, VP2 y VP3 sin alinear. (D) Alineamiento flexible entre VP1 (rojo) y VP2 (amarillo). optLength=164; LengthA=260; LengthB=255; TN=0. (E) Alineamiento flexible entre VP2 (amarillo) y VP3 (azul). optLength=210; LengthA=260; LengthC=282; TN=1. (F) Alineamiento flexible entre VP1 (rojo) y VP3 (azul). optLength=208; LengthB=255; LengthC=282; TN=0.

El RMSD promedio entre las proteínas de la cápside de los Dicistroviridae, obtenido con FATCAT, fue de 3.12 con un rango de 3.09 a 3.18, introduciendo solo una rotación en el alineamiento entre VP2 y VP3, específicamente en la región de la cola N-terminal de VP3 que alinea con el par de hebras beta de la cola N-terminal de VP2. Esta rotación pudo haber sido un artefacto del algoritmo para lograr un mejor alineamiento entre las estructuras. El RMSD y valor Z promedio obtenidos con DALI fueron de 4.33 y 12.66, con rangos de 3.6 a 4.8 y de 10 a 15.2, respectivamente. De acuerdo con el alineamiento flexible, la mayor similitud estructural corresponde a VP1 y VP3 (valor P) o, a VP2 y VP3 (RMSD). En cuanto

al alineamiento rígido, el valor Z más alto se da entre VP2 y VP3, mientras que le RMSD más bajo se da entre VP1 y VP3 (Tabla 4). En general, VP1 y VP2 son las proteínas que comparten menos similitud estructural. A nivel de secuencia VP1 no alinea con VP2 (Tabla 6).

Tabla 4. Valores de alineamiento flexible y rígido de las proteínas de cápside de Dicistroviridae.

Alineamiento	Valores de alineamiento flexible			Valores de alineamiento rígido	
	P	chRMSD	optRMSD	Z	RMSD
VP1-VP2	4.13e-06	3.76	3.1	10	4.8
VP2-VP3	5.59e-06	2.67	3.09	15.2	4.6
VP1-VP3	2.52e-08	4.31	3.18	12.8	3.6

En los Picornaviridae la cola N-terminal de VP2, en lugar de extenderse hacia otras subunidades, se encuentra retraída interactuando con la parte principal de la misma subunidad. La diferencia en la orientación se ha atribuído a un intercambio de dominio ocurrido cerca de la Ile72 de Dicistroviridae (Liljas *et al*. 2002). Sin embargo, independientemente de la diferencia en la orientación, las colas N-terminales de VP2 de los Dicistroviridae y los Picornaviridae conservan un par de hebras beta que no se encuentra en otras subunidades. Debido a esto era factible esperar una rotación en dicha región para alinear la cola N-terminal de VP2 con la de VP1 y VP3.

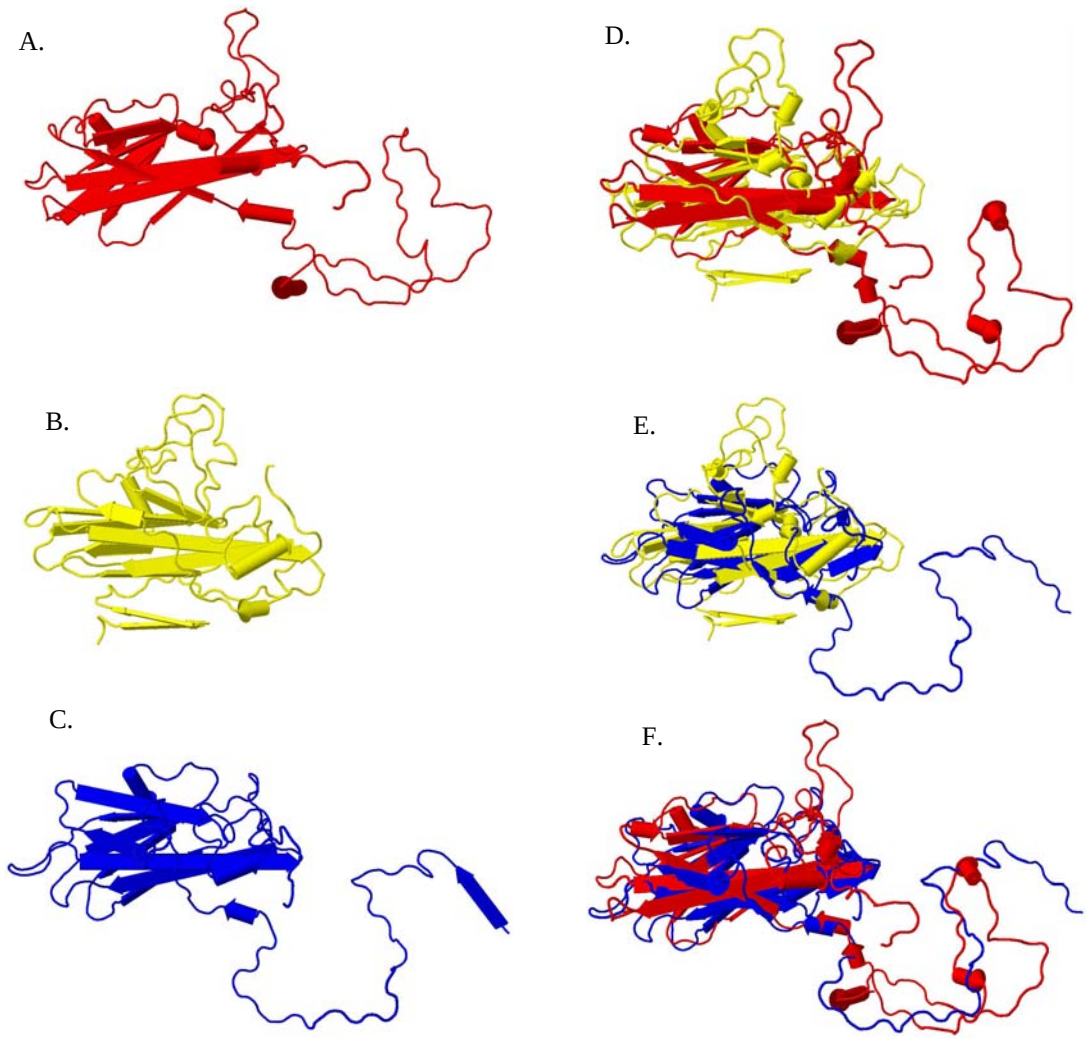


Figura 7. Proteínas de la cápside VP1, VP2 y VP3 de Picornaviridae (1AYM 1, 2 y 3). (A, B y C) VP1, VP2 y VP3 sin alinear. (D) Alineamiento flexible entre VP1 (rojo) y VP2 (amarillo). optLength=154; LengthA=285; LengthB=252; TN=1. (E) Alineamiento flexible entre VP2 (amarillo) y VP3 (azul). optLength=176; LengthA=285; LengthC=238; TN=0. (F) Alineamiento flexible entre VP1 (rojo) y VP3 (azul). optLength=162; LengthB=252; LengthC=238; TN=0.

El RMSD promedio entre las proteínas de la cápside de los Picornaviridae, obtenido con FATCAT, fue de 3.01 con un rango de 2.86 a 3.1, introduciendo solo una rotación en el alineamiento entre VP1 y VP2 en una región no relacionada con las colas N-terminal. La ausencia de la rotación esperada en la cola N-terminal de VP2 podría deberse al hecho de que, a diferencia de la similitud estructural que se observa entre las colas N-terminal de las subunidades de Dicistroviridae, las colas N-terminal de los Picornaviridae presentan conformaciones diferentes (Liljas *et al.* 2002) (Figura 6 y 7). El RMSD y valor Z promedio

obtenidos con DALI fueron de 3.53 y 10.1, con rangos de 3.1 a 3.6 y de 8.8 a 11.6, respectivamente (Tabla 5).

Tabla 5. Valores de alineamiento flexible y rígido de las proteínas de cápside de Picornaviridae.

Alineamiento	Valores de alineamiento flexible			Valores de alineamiento rígido	
	P	chRMSD	optRMSD	Z	RMSD
VP1-VP2	1.72e-03	4.01	3.06	8.8	3.6
VP2-VP3	1.25e-06	4.1	3.1	11.6	3.1
VP1-VP3	2.24e-06	4.47	2.86	9.9	3.8

En los Picornaviridae se volvió a encontrar que VP3 es la más parecida a las otras dos proteínas tanto a nivel de estructura, flexible y rígida (Tabla 5), como a nivel de secuencia (Tabla 6). A pesar de que VP1 y VP2 no alinean de manera significativa con FATCAT, probablemente por la cantidad de *loops* y/o por la posición de la cola N-terminal de VP2, se puede inferir que si VP3 es homóloga a VP2 y a VP1, entonces VP1 y VP2 también son homólogas.

El alineamiento de estructura terciaria hace más evidente la relación que existe entre estas proteínas de la cápside que los alineamientos de secuencia, cuyas identidades yacen en la zona de penumbra de entre 20 y 35% (Rost, 1999). El caso más significativo a nivel de secuencia podría ser el de VP2 y VP3 de Dicistroviridae, con identidad del 22.3% y E(200) = 0.066 (Tabla 6). De acuerdo con Liljas y colaboradores (2002), mientras que VP1 es la proteína menos conservada entre los Picornavirales, VP2 y VP3 tienen un nivel de conservación similar. Esto, en conjunto con el hecho de que VP3 es la subunidad que guarda mayor similitud estructural con VP1 y VP2, me permite suponer que VP1 fue la proteína ancestral a partir de la cual se duplicó VP3 y que VP3 dio origen por duplicación a VP2. Otra alternativa podría ser que tanto VP3, que estabiliza la formación de pentámeros, como VP2, que estabiliza la interacción entre pentámeros (Liljas *et al.* 2002), se encuentren mejor conservadas por la importancia de sus respectivos papeles en el ensamblaje de la cápside.

Tabla 6. Valores de alineamientos de secuencia entre proteínas de la cápside de Dicistroviridae y Picornaviridae. En ninguna de las familias se pudieron alinear VP1 y VP2.

Alineamiento	Identidad	Similitud	Query Cover	E(200)
VP1-VP3 Dicis	28.8%	52.5%	45.38%	23
VP2-VP3 Dicis	22.3%	53.6%	65.09%	0.066
VP1-VP3 Picor	28.3%	51.7%	21.05%	91
VP2-VP3 Picor	45.8%	75%	9.19%	3200

En el caso de los Secoviridae, la cápside se ensambla a partir de dos polipéptidos que forman tres sandwiches beta de tipo *jelly roll* (Lin *et al*. 2003). Este es un heterodímero compuesto de una subunidad grande (L) con dos sandwiches beta y una subunidad pequeña (S), con solo uno. En el alineador flexible se sobrepone S con uno de los dos dominios de L sin obtener valores significativos, con RMSD de 3.12 y P = 1.09e-02. Sin embargo, al analizar el alineamiento visualmente se notó una gran similitud estructural entre S y ambos dominios de L, incluso sin alinearlos (Figura 8A, B y C), de modo que se procedió con el análisis. En el alineador rígido, S alinea con ambos dominios de L de manera significativa, con RMSD de 3.6 y 3.8, y Z de 7.4 y 7.2, respectivamente (Figura 8D). A nivel de secuencia no se pudieron alinear.

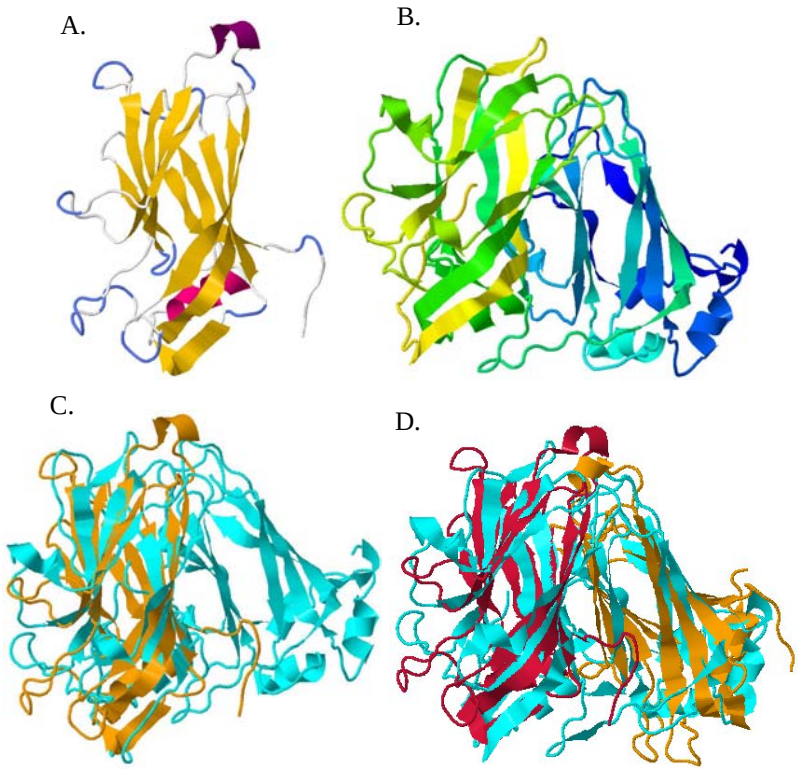


Figura 8. Alineamiento entre la subunidad S y L de la cápside de Secoviridae (1PGW 1 y 2). (A) Subunidad S sin alinear. (B) Subunidad L sin alinear, coloreada de N (azul) a C-terminal (verde) para distinguir sus dos dominios. (C) Alineamiento flexible entre S (anaranjado) y L (cyan). P=1.09e-02; OptLength=137; LengthA=185; LengthB=351; TN=1; OptRMSD=3.12 ; chRMSD=4.68. (D) Alineamiento rígido mostrando cómo S (rojo y anaranjado) alinea con ambos dominios de L (cyan). Z=7.4; RMSD=3.6; Z'=7.3; RMSD'=3.8.

Una peculiaridad de la subunidad S es el número de hebras β que conforman el sandwich β. A diferencia del resto de las proteínas de cápside, que se componen de ocho hebras beta antiparalelas en forma de *jelly roll*, la subunidad S presenta cinco hebras de ambos lados (Lin *et al.* 2003).

Los tres casos mostrados de duplicación de proteínas de la cápside de familias pertenecientes al orden de los Picornavirales parecen representar el mismo caso de duplicaciones ancestrales que a partir de un solo gen dieron origen a tres proteínas de cápside, pues también se ha demostrado que las distintas VP1, VP2 o VP3 de las diferentes familias están estructuralmente relacionadas (Liljas *et al.* 2002). Así, tanto S como los dos dominios de L de Secoviridae deben ser equivalentes a VP1, VP2 y VP3 de otras familias de Picornavirales, con la peculiaridad de que en L los dos dominios no son separados por la proteasa.

De acuerdo con la base de datos de virus ViralZone (www.expasy.org/viralzone), en las tres familias de Picornavirales las proteínas de la cápside se encuentran de manera adyacente, pero su distribución en el genoma es diferente (Figura 9). En los Dicistroviridae el genoma está compuesto de dos marcos de lectura abiertos (ORF), de los cuales el segundo codifica para las proteínas estructurales, que están ubicadas en el extremo 3', en el orden VP1, VP4, VP2 y VP3 (Figura 9A). En Picornaviridae el genoma consiste en un solo ORF que codifica una poliproteína que se divide en tres regiones procesadas por las proteasas. En la región P1 5' se codifican las proteínas estructurales en el orden VP4, VP2, VP3 y VP1 (Figura 9B). En la familia Secoviridae el genoma se divide en dos segmentos de ssRNA, RNA-1 y RNA-2, cada uno con su respectiva Vpg y cola de poli-A 3'. El RNA-1, de 6 a 8 kb, codifica para proteínas no estructurales mientras que el RNA-2, de 4 a 7 kb, codifica para las proteínas estructurales ubicadas en su propio extremo 3' (Figura 9C).

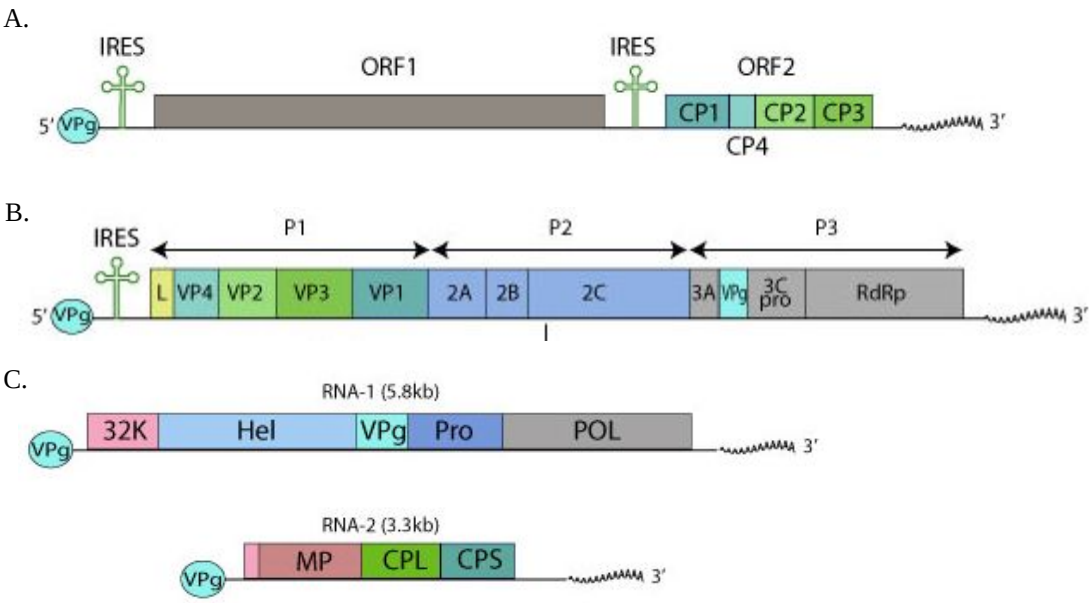


Figura 9. Organización genómica de (A) Dicistroviridae, (B) Picornaviridae y (C) Secoviridae. Tomadas de ViralZone www.expasy.org/viralzone, Swiss Institute of Bioinformatics.

A pesar de que cada subunidad lleva a cabo un papel particular durante el proceso de ensamblaje de la cápside, la notable similitud estructural entre las mismas me permite hipotetizar que sus papeles pueden ser intercambiables. De ser así, estas duplicaciones podrían representar un caso de redundancia funcional cuya relevancia adaptativa estaría

relacionada con en el rápido ensamblaje del virión. La rápida replicación podría ser ventajosa en ambientes cambiantes como el generado por el sistema inmune (Holmes, 2009). Lo primero que salta a la mente para que un virus logre una rápida replicación es la selección de genomas cortos. Sin embargo, también es posible que la duplicación redundante de genes relacionados con el ciclo de replicación viral permita el crecimiento de los genomas sin reducir la velocidad de replicación, amortiguando la presión de la rápida replicación sobre los genomas. En concreto, es posible que la velocidad de ensamblaje del virión sea dependiente de la cantidad de subunidades disponibles, de modo que si se tiene un mayor número de genes que codifican para proteínas que contribuyen al ensamblaje de la cápside, más rápida será su formación.

4.3 Cisteín-proteasas

El último caso que sugiere una duplicación de secuencia en un genoma de RNA se presenta de nueva cuenta en virus del orden de los Picornavirales, esta vez involucrando a la proteasa 3C y a la proteasa 2A de la familia Picornaviridae. Estas proteasas, caracterizadas por la presencia de una cisteína en su sitio catalítico, se distribuyen exclusivamente en virus y se ha demostrado que son homólogas a las serín-proteasas que se encuentran distribuidas en los tres dominios de la vida y en algunos virus de RNA (Bazan & Fletterick, 1988; Gorbalenya *et al.* 1989). Aunque no se ha hecho una filogenia, se ha sugerido que un virus adquirió una serín-proteasa de un hospedero eucarionte y que luego se sustituyó la serina del sitio catalítico por una cisteína (Barrett & Rawlings, 2001). La estructura general de las proteasas relacionadas con la tripsina se compone de dos barriles β homólogos de seis hebras beta cada uno (Lesk & Fordham, 1996). Sin embargo, la proteasa 2A resulta ser el primer caso descrito de una proteasa con un dominio N terminal compuesto de cuatro hojas beta antiparalelas (Petersen *et al.* 1999). A pesar de la diferencia en el número de hojas beta entre sus respectivos dominios N terminal, las estructuras de estas proteínas alinean excepcionalmente bien (Figura 10). El alineador de FATCAT consiguió sobreponerlas con un RMSD de 2.96 y P de 3.72e-08. Los alineamientos fueron mejores cuando se usó el alineador rígido, con RMSD de 2.7 y Z de 14.

A nivel de secuencia se obtiene una identidad del 38.7%, similitud de 61.3%, cobertura de 17.22% y E(200) = 1300 que, de acuerdo con otros estudios, podría corresponder exclusivamente al sitio catalítico. Por ejemplo, ya se había propuesto la relación entre la proteasa 3C y 2A a partir de un análisis de alineamiento de secuencia en el cual, a pesar de conseguir una identidad promedio de apenas 13%, se identificó la conservación absoluta de la triada catalítica His-Asp-Cys, que es homóloga a la triada catalítica de las serín-proteasas His-Asp-Ser (Bazan & Fletterick, 1988). Un estudio que determinó la estructura de la proteasa 2A, la comparó cualitativamente con la proteasa 3C y algunas serín-proteasas, argumentando que la similitud estructural entre 2A y 3C implica que provienen de una duplicación (Petersen *et al.* 1999).

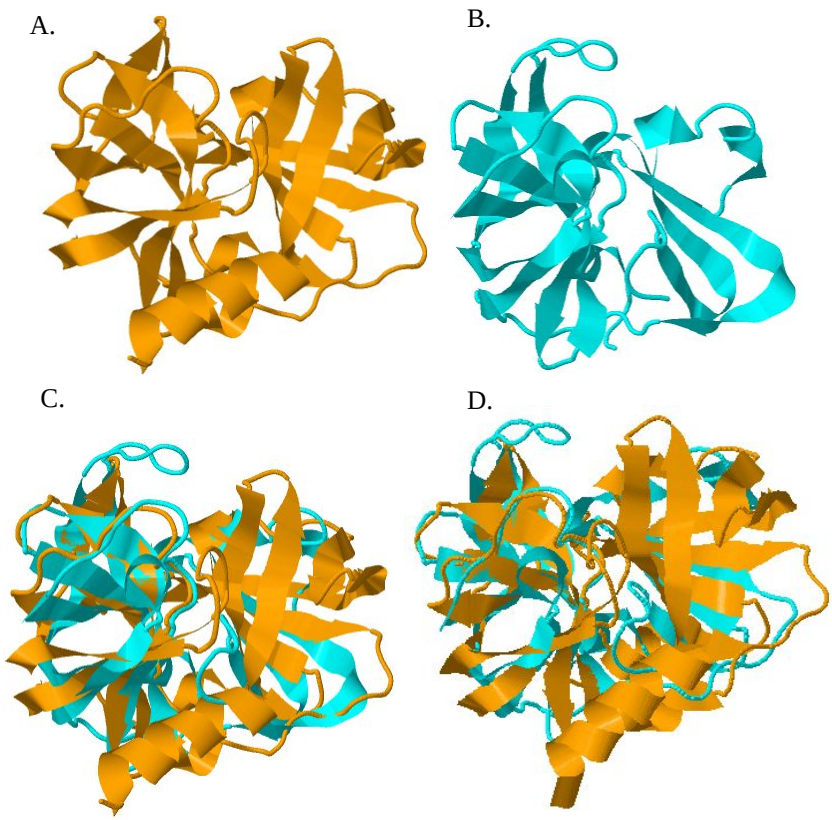


Figura 10. Alineamiento entre la proteasa 3C (1CQQ) y la proteasa 2A (2HRV) de Picornaviridae. (A y B) Estructuras sin alinear de 3C y 2A, respectivamente. (C) Alineamiento flexible. P=3.72e-08; OptLength=133; LengthA=180; LengthB=139; TN=0; OptRMSD=2.96 ; chRMSD=2.11. (D) Alineamiento rígido. RMSD=2.7; Z=14.0.

Aunque las estructuras de estas proteasas son bastante similares, las diferencias en el dominio N-terminal podrían ser suficientes para explicar las diferentes funciones que estas

desempeñan. Por un lado, la proteasa 3C es la principal encargada de dividir a la mayoría de los componentes de la poliproteína (Figura 9B), incluyendo su propia escisión, cortando en sitios Gln-Gly. Además, se sabe que está involucrada en la proteólisis de proteínas como la histona H3 y la proteína de unión a caja TATA (TBP), inhibiendo severamente la transcripción del hospedero. Por otro lado, la proteasa 2A solamente cataliza su propia escisión en sitios Tyr-Gly durante el procesamiento de la poliproteína. Una función adicional de esta proteasa está relacionada con la proteolisis de eIF4F que resulta en la inhibición de la traducción celular (Porter, 1993). Lo que se observa aquí es un caso de duplicación que resultó en neofuncionalización.

4.4 Genes homólogos dentro de un mismo genoma y los mecanismos que los generan

Es interesante hacer notar que el *Duck hepatitis A virus* (DHAV), un picornavirus del genero Avihepatovirus, contiene tres copias del gen 2A (Figura 11). Esto podría ser indicativo de varios eventos de duplicación. Sin embargo se ha notado que cada copia tiene mayor similitud con genes 2A de otros picornavirus, sugiriendo que estas copias no se duplicaron en el DHAV sino que se transfirieron horizontalmente a partir de virus diferentes (Simon-Lorirere & Holmes, 2013). Un caso similar se ha reportado en las serín-proteasas P1a y P1b del género Ipomovirus de la familia Potyviridae (Valli *et al*. 2007)

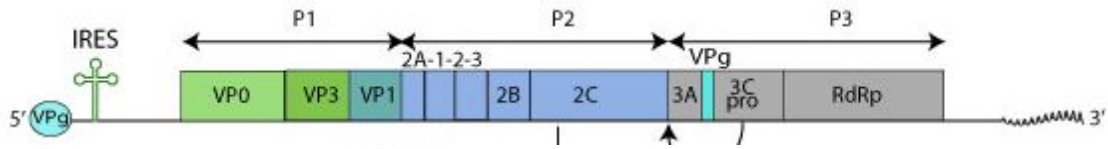


Figura 11. Organizacion del genoma de Avihepatovirus. Tomado de ViralZone www.expasy.org/viralzone, Swiss Institute of Bioinformatics.

Esto podría indicar que la recombinación por "copy choice" podría ser el mismo mecanismo que genera las duplicaciones y la transferencia horizontal de genes. Otro caso similar que resulta interesante es el de la RNasaH de retrovirus. El genoma de retrovirus presenta una secuencia conectora entre la transcriptasa reversa y la RNasaH, la cual tiene una gran similitud estructural con la última, por lo que se ha sugerido que refleja un evento de duplicación (Artymuik *et al.* 1993). Sin embargo, basados en alineamientos de secuencias, Malik y Eickbush (2001) propusieron que la región conectora solía ser la RNasaH original de

retrovirus y que la RNasaH activa de retrovirus provino de una transferencia a partir de un retrotransposon, causando la pseudogenización de la primera. También resulta interesante el que, adyacente a la RNasaH, se encuentra el gen de la integrasa, cuyo sitio catalítico pertenece a la superfamila de proteínas con estructura similar a la RNasaH (CL0219). Sin embargo, un análisis filogenético de las proteínas pertenecientes a la superfamilia RNasaH, muestra que la integrasa se encuentra más relacionada con la transposasa de transposones de DNA, que con las RNasaH (Majorek *et al*. 2014). Incluso, algunos estudios sobre el origen de los retrovirus, sugieren que estos pudieron resultar de la fusión de un retrotransposon sin repetición terminal larga (LTR), que poseía transcriptasa reversa y RNasaH pero no integrasa, con un transposón de DNA, que carece de transcriptasa reversa y RNasaH pero tiene transposasa (Malik & Eickbush, 2001). Esta fusión habría dado lugar a un retrotransposon con LTR, que posteriormente habría adquirido al gen ENV para dar origen a los retrovirus (Malik & Eickbush, 2001). Esto indicaría que la integrasa es ortóloga a la transposa, que su posición en el genoma es resultado de transferencias horizontales y que su relación con la RNasaH es anterior al origen de los retrovirus. No sería raro que estas transferencias se hayan dado por recombinación por 'copy choice', pues se sabe que la tasa de recombinación en retrovirus puede ser más alta que la tasa de mutación (Holmes, 2009). La causa que determina el resultado de la recombinación es parte de las posibilidades del mecanismo de 'copy choice' (Figura 4).

Además de la recombinación por 'copy-choice', el reclutamiento de segmentos homólogos en virus con genomas segmentados podría funcionar como un mecanismo que genera secuencias homólogas en un mismo genoma. Ambos mecanismos tienen el potencial de generar copias homólogas de origen intratípico o intertípico, de modo que es posible que cualquier par de secuencias homólogas consideradas como parálogos sean, en realidad, genes xenólogos. El resultado podría depender de la frecuencia de las coinfecciones. Pero aunque las coinfecciones sean muy frecuentes, el reclutamiento de segmentos podría estar limitado por un reconocimiento de motivos estructurales en el RNA, que descartaría a las secuencias divergentes (Reanney, 1982). Por otro lado, la recombinación por 'copy-choice' es más frecuente entre secuencias similares (Lai, 1992; Simon-Loriere & Holmes, 2012). De manera análoga se ha observado que, en arqueas, conforme aumenta la distancia evolutiva disminuye la frecuencia de transferencia horizontal (Soucy *et al.* 2015). Por lo tanto, es más probable

que las secuencias homólogas ubicadas en el mismo genoma provengan, en primer lugar, de la replicación del genoma original (parálogos) y, en segundo lugar, de genomas de virus emparentados que infectan al mismo hospedero (xenólogos). Entre más relacionados estén los virus coinfectantes, es más probable que intercambien genes. Es por lo tanto factible que las cápsides de los Picornavirales, que están presentes en todos los géneros de las diferentes familias (Tabla 7), sean un rasgo ancestral con una alta probabilidad de sufrir recombinación y posiblemente transferencia horizontal. Aunque se ha visto que en al familia Picornaviridae no ocurre recombinación en las proteínas VP1 y VP3 (Lai, 1992), es posible que las recombinaciones expliquen las diferencias en la organización de los genomas entre las familias de Picornavirales (Reanney, 1982; Lai, 1992; Tate *et al*. 1999) (Figura 9).

Tabla 7. Número de géneros que presentan la duplicación. Basada en Simon-Loriere & Holmes (2013) y ViralZone www.expasy.org/viralzone, Swiss Institute of Bioinformatics.

Clasificación	Familia	Número de géneros	Gen duplicado	Géneros con la duplicación	Géneros con estructuras*
(+)ssRNA	Closteroviridae	4	CP -> CPm	4	0
			CPm1 -> CPm2	1	0
	Picornaviridae	29	Vpg -> Vpg	2	0
			VP1 -> VP3 y VP2	29	4
			3C -> 2A	27	1
	Dicistroviridae	2	VP1 -> VP3 y VP2	2	1
	Secoviridae	8	CPS -> CPL1 y 2	8	2
(-)ssRNA	Benyvirus	1	p25 -> p26	1	0
	Rhabdoviridae	11	G -> Gns	1	0
			U1 -> U2	1	0
(RT)ssRNA	Retroviridae	7	orfA -> orfB	1	0
			orf1 -> orf2	1	0
			vpr -> vpx	2	0
dsRNA	Totiviridae	5	KP6α -> KP6β	2	1

*De acuerdo con los criterios de la metodología empleada en este trabajo.

La ubicación de los genes duplicados podría darnos un indicio sobre el mecanismo que los generó. De todos los casos reportados hasta ahora solamente los correspondientes a las proteasas 3C y 2A y a los factores de patogenicidad p25 y p26 no se encuentran de manera adyacente (Tabla 8). El caso de los factores de patogenicidad podría estar explicado por reclutamiento de segmentos homólogos. En cuanto a las proteasas, es posible que estas se hayan duplicado por recombinación de tipo 'copy-choice'. Sin embargo, aunque es común que se utilice el criterio de localización para distinguir entre parálogos y xenólogos (Treangen & Rocha, 2011), debido al contraste que existe entre los mecanismos que los generan en seres vivos, en virus es difícil determinar la naturaleza de las homologías dentro de un mismo

genoma utilizando el mismo criterio, debido a que los mecanismos que generan genes parálogos y xenólogos pueden ser los mismos. Determinar si un par de secuencias homólogas dentro de un mismo genoma es resultado de la duplicación o de la transferencia horizontal, podría resolverse de mejor manera con análisis filogenéticos. Un aspecto importante sobre la posición relativa de los genes duplicados tiene que ver con la estabilidad de la duplicación. Por ejemplo, se ha sugerido que las duplicaciones adyacentes son menos estables porque aumentan la probabilidad de recombinación y pérdida segregacional (Anderson *et al.* 2015). De hecho, es posible que la deleción de las regiones duplicadas luego de la recombinación, evento que permite que los virus recuperen adecuación, se deba a recombinaciones posteriores (Lowry *et al.* 2014). Entonces, si la generación de duplicaciones adyacentes y su subsecuente pérdida son tan frecuentes, se vuelve evidente que las duplicaciones adyacentes que se han conservado deben conferir una clara ventaja adaptativa.

Tabla 8. Organización, longitud del genoma y la posición relativa de los genes duplicados. Basado en Simon-Loriere & Holmes (2013) y ViralZone www.expasy.org/viralzone, Swiss Institute of Bioinformatics. La longitud promedio del genoma se calculó a partir de los genomas de los tipos virales en los que se encontraron las duplicaciones (ver Anexo 4). Los casos encontrados en este estudio están resaltados en negritas.

Clasificación	Familia	Organización del genoma	Longitud promedio del genoma(kb)	Gen duplicado	Posición relativa
(+)ssRNA	Closteroviridae	mono o bipartito	17.242	CP -> CPm	Adyacente
				CPm1 -> CPm2	Adyacente
	Picornaviridae	monopartito	7.479	Vpg -> Vpg	Adyacente
				VP1 -> VP3 y VP2	Adyacente
				3C -> 2A	Separados
	Dicistroviridae	monopartito	9.185	**VP1 -> VP3 y VP2**	Adyacente
	Secoviridae	bipartito	9.657	**CPS -> CPL1 y 2**	Adyacente
	Benyvirus	pentapartito	15.914	p25 -> p26	Segmentos diferentes
(-)ssRNA	Rhabdoviridae	monopartito	14.916	G -> Gns	Adyacente
				U1 -> U2	Adyacente
(RT)ssRNA	Retroviridae	monopartito, dimérico	10.708	orfA -> orfB	Adyacente
				orf1 -> orf2	Adyacente
				vpr -> vpx	Adyacente
dsRNA	Totiviridae	monopartito	6.099	**KP6α -> KP6β**	Adyacente

Interesantemente, se ha sugerido que tanto la recombinación 'copy-choice' como el reclutamiento de segmentos homólogos pueden influir en la emergencia de enfermedades virales. Por ejemplo, es posible que el *Rous sarcoma virus* haya adquirido el gen de la tirosin quinasa, un oncogen, por recombinación por 'copy choice' con un transcrito celular. Por otro lado, el reordenamiento de segmentos, específicamente de los que codifican para hemaglutinina y neuraminidasa en virus de influenza, contribuye para la evasión del sistema inmune y la ocurrencia de epidemias (Simon-Loriere & Holmes, 2012).

4.5 Relación entre la organización y el tamaño del genoma con las duplicaciones

En promedio, la clase de virus de RNA con genomas más grandes es la (-)ssRNA, seguida de dsRNA, (+)ssRNA y (RT)ssRNA (Campillo-Balderas *et al.* 2015). Este patrón podría estar explicado por la proporción de familias con genomas segmentados que sigue una distribución casi idéntica (Tabla 9) pues, en promedio, los genomas segmentados son de mayor tamaño que los no segmentados (Holmes, 2009).

Tabla 9. Número de segmentos y longitud de los genomas por clasificación. Basado en ViralZone www.expasy.org/viralzone, Swiss Institute of Bioinformatics.

Clasificación	Número de familias	Número de familias con genomas segmentados	Proporción de familias con genomas segmentados	Promedio de longitud de genoma (kb) (Campillo-Balderas *et al.* 2015)
(-)ssRNA	10	6	0.6	13.264
dsRNA	9	6	0.67	11.771
(+)ssRNA	38	11	0.29	9.594
(RT)ssRNA	1	0	0	8.383

Partiendo de que la segmentación de genomas permite el crecimiento de los mismos sin incrementar el efecto deletéreo de las mutaciones, y de que existe una relación directamente proporcional entre el tamaño del genoma y el número de duplicados (Hughes *et al.* 2005), se esperaría que (-)ssRNA y dsRNA, con genomas grandes y segmentados, hubiesen presentado la mayor cantidad de casos de duplicación. Sin embargo, la mayor cantidad de duplicaciones

reportadas hasta ahora en virus de RNA se han encontrado en (+)ssRNA (7), seguido de (RT)ssRNA (3), (-)ssRNA (2) y por último dsRNA (1) (Figura 12).

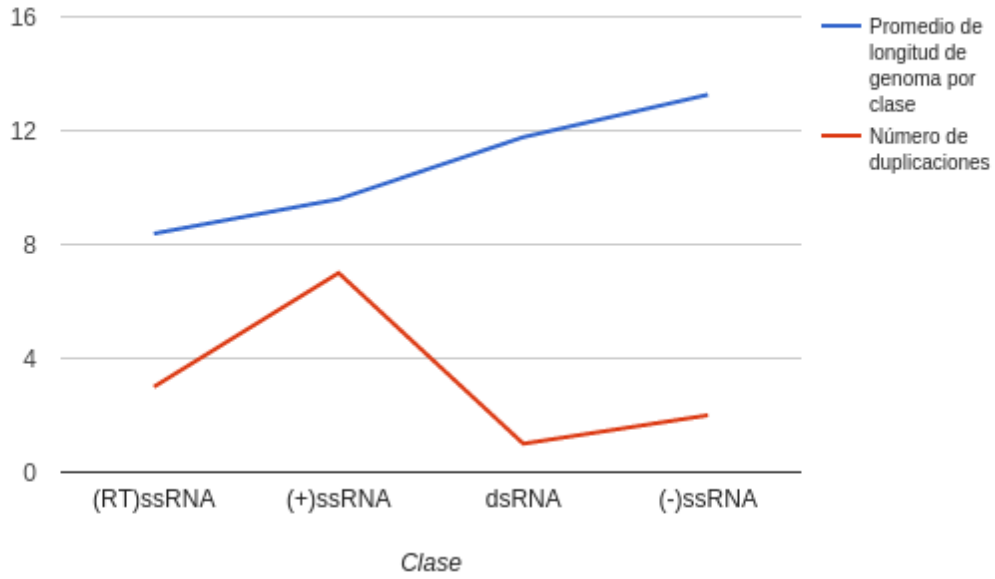


Figura 12. La longitud del genoma no está correlacionada con el número de duplicaciones encontradas hasta ahora en virus de RNA (ver Anexo 3).

Aunque la posibilidad de encontrar más duplicaciones en genomas grandes está limitada por la ausencia de datos suficientes, la familia Totiviridae, que es por ahora la única representante de virus de doble cadena que presenta algún caso de duplicación, contiene los genomas más pequeños en relación al resto de las familias mencionadas en este trabajo y posee un solo segmento (Tabla 8). De manera interesante, el promedio de longitud de genoma de los virus en los que se han encontrado duplicaciones, agrupados por clase, es mayor al promedio de longitud de genoma por clase, y sigue una relación parecida a la esperada excepto por el caso de la familia Totiviridae (Figura 13), cuyo caso podría tener alguna relación con el hecho de que el virus de RNA con el genoma más pequeño, *Saccharomyces cerevisiae killer virus M1,* que también infecta a hongos, forma parte de los virus de dsRNA (Campillo-Balderas *et al*. 2015). El caso de estos virus de dsRNA podría ser resultado de las posibilidades que giran en torno a las tasas de replicación y mutación intrínsecamente altas de los virus de RNA, que no permitirían su crecimiento.

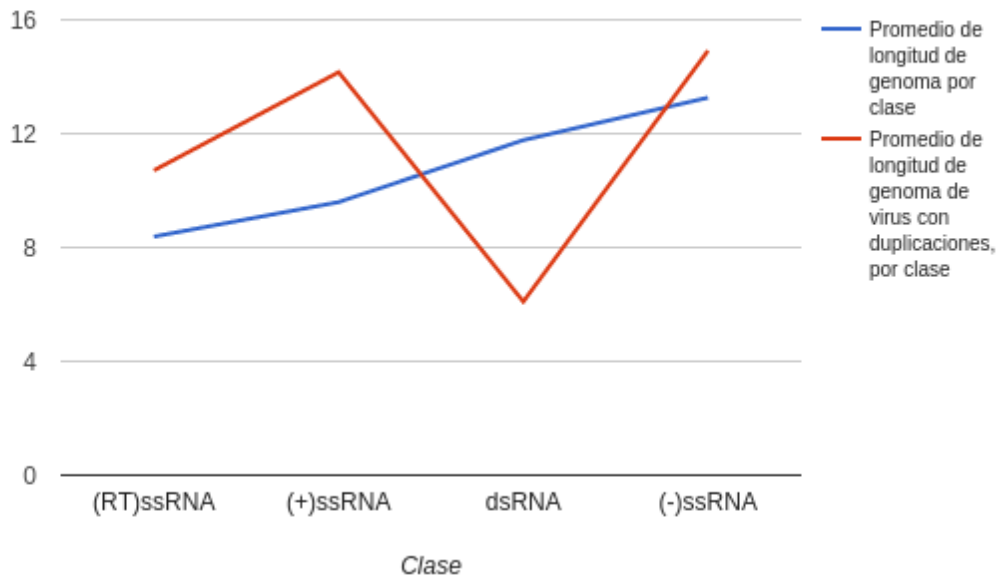


Figura 13. El promedio de longitud de genoma de virus con duplicaciones, agrupados por clase, es mayor al promedio general de cada clase y sigue una relación similar a la esperada, salvo por el caso de la familia Totiviridae de dsRNA.

La siguiente categoría con menor cantidad de duplicaciones reportadas es (-)ssRNA (Figura 12). A pesar de que esta es la segunda categoría con mayor proporción de familias con genomas segmentados (Tabla 9), Rhabodovirdae, que no tiene genomas segmentados, es la única familia de esta categoría que hasta ahora presenta casos de duplicación. Aun así, las duplicaciones ocurridas pudieron haber contribuido para incrementar el tamaño de su genoma que es mayor al del promedio de su clase (Tablas 8 y 9).

En el caso de (RT)ssRNA los genomas no se consideran segmentados pero se encapsidan en pares, formando un virus 'pseudo diploide' que aumenta la probabilidad de recombinación (Holmes, 2009). Esta característica podría explicar la frecuencia de las duplicaciones en la familia Retroviridae (Figura 12). Sin embargo, se ha sugerido que una infección persistente, que aumenta la probabilidad de coinfección, influye más sobre la tasa de recombinación en estos virus (Simon-Loriere & Holmes, 2012).

Finalmente, el caso de (+)ssRNA es interesante porque, a pesar de ser una de las clases con menor proporción de familias con genomas segmentados y con menor longitud promedio del genoma (Tabla 9), presenta la mayor cantidad de casos de duplicación (Figura 12). En esta clase hay tres familias cuyos casos podrían ser fácilmente explicados. Tanto los Closteroviridae, como los Secoviridae y los Benyvirus, representan tres de los pocos casos de genomas segmentados en (+)ssRNA. Consecuentemente, los Closteroviridae y los Benyvirus presentan genomas de mayor tamaño, incluso que el promedio de longitud de (-)ssRNA (Tabla 8 y 9). Por otro lado, las familias Dicistroviridae y Picornaviridae presentan genomas de longitud cercana a la media de virus de RNA (10kb) y sin segmentación.

4.6 Relación entre ambiente y duplicaciones

Otro aspecto interesante que podría explicar la fijación de duplicaciones es la relación del virus con el hospedero. Por ejemplo, se ha sugerido que en ambientes estables, como la carencia de inmunidad adaptativa en plantas, los virus tienen una tasa de replicación y sustitución más baja que virus que infectan animales (Holmes, 2009). Por lo tanto, podría esperarse que en los virus que infectan vertebrados los genomas sean más cortos por selección de una rápida replicación. De manera consistente podemos observar que, de los virus en los que se han reportado duplicaciones, los que infectan plantas tienen genomas más grandes y con más segmentos que los virus que infectan a otros hospederos (Figura 14). De las ocho familias que hasta ahora reportan casos de duplicación, solamente dos (Picornaviridae y Retroviridae) infectan exclusivamente a vertebrados y no presentan segmentación (Tablas 8 y 10). El tamaño del genoma de estas familias se encuentra en un rango de 7.152kb (*Rhinovirus A*) a 13.125kb (*Walleye epidermal hyperplasia virus 2*). En el caso de los Rhabdoviridae, que infectan a vertebrados con vectores invertebrados (Tabla 10), el rango de la longitud del genoma, sin casos de segmentación (Tabla 8), va de 13.196kb (*Wongabel virus*) a 15.87kb (*Kotonkan virus*). En el caso de los Secoviridae, los Closteroviridae y los Benyvirus, que infectan a plantas (Tabla 10), el rango de longitud de genomas va desde 9.657kb (*Bean pod mottle virus*) hasta 19.296kb (*Citrus tristeza virus*), presentando dos casos con dos segmentos y uno con cinco (Tabla 8).

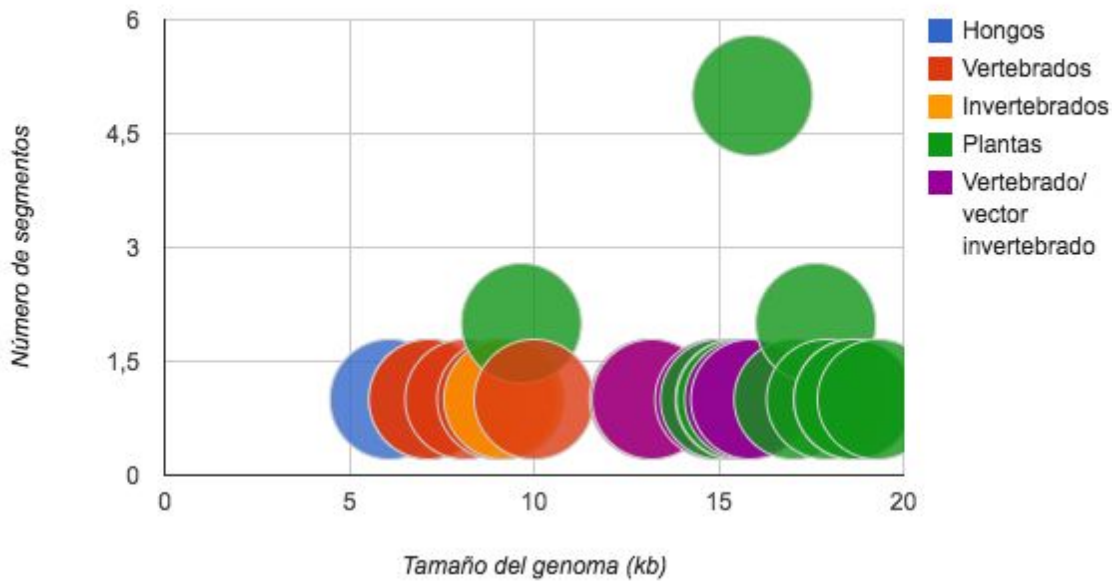


Figura 14. Los virus que infectan a plantas presentan genomas más grandes y con mayor número de segmentos. El color designa a los hospederos que infectan (ver Anexo 4).

En las familias en las que no hay segmentación de genomas y los hospederos son vertebrados, es posible que se seleccionen duplicaciones que promueven la rápida conclusión del ciclo de replicación viral. En estos casos las duplicaciones podrían permitir una producción de viriones más rápida que la tasa de eliminación de estos por el sistema inmune adaptativo. Un ejemplo es el caso de las cápsides de la familia Picornaviridae. Los otros dos casos de duplicación en esta familia también podrían estar relacionados con la rápida producción de viriones. Mientras que la duplicación de Vpg podría aumentar la tasa de replicación del genoma, la duplicación de las proteasas permitiría el rápido reclutamiento del transcrito viral por el ribosoma a través de la inhibición de la traducción celular (Tabla 10).

En los Retroviridae también podrían estar seleccionadas las duplicaciones que permiten un ciclo de replicación viral más rápido, pues en el caso de la duplicación de orfA a orfB puede aumentar tanto la proliferación celular como la tasa de transcripción viral (Tabla 10). Esta duplicación fue encontrada en el género Epsilonretrovirus, que tiene diferentes proteínas accesorias y tropismo celular comparado con los Lentivirus, en los cuales el sistema inmune adaptativo queda deshabilitado. Las duplicaciones de Lentivirus, particularmente de HIV-2,

no parecen estar relacionadas con una mayor tasa de replicación, sino con la evasión de mecanismos celulares de defensa. Por ejemplo, además de mediar la translocación nuclear del complejo de preintegración, vpx está involucrada en la inhibición de SAMHD1 que es un factor de restricción de la retrotranscripción, mientras que vpr inhibe un mecanismo de silenciamiento de DNA foráneo. Como tercera función, vpr lleva a cabo el arresto celular en fase G2 (Swiss Institute of Bioinformatics, 2014). La duplicación subfuncionalizante de vpr a vpx puede permitir que vpr cumpla sus otras funciones sin descuidar la integración del genoma viral. Las funciones de orf1 y orf2 son desconocidas (Simon-Loriere & Holmes, 2013).

Los casos de los Rhabdoviridae, que infectan a vertebrados con vectores invertebrados, son menos claros debido a que se desconoce la función de U1 y U2 (Simon-Loriere & Holmes, 2013), así como la de Gns (Wang & Walker, 1993; Gubala et al. 2010). Gns no forma parte del virión y solo ha sido observada en asociación con membranas celulares gracias a la conservación de dominios transmembranales. Aunque esta podría ocultar una nueva función aún desconocida, su ausencia en el virión parece indicar que más bien se trata de un caso de pseudogenización.

En el casos de los virus que infectan plantas, la presencia de un sistema inmune innato (ausencia de inmunidad adaptativa) no sería suficiente presión como para seleccionar genomas cortos. Como consecuencia los genomas podrían crecer y acumular más mutaciones, beneficiando la selección de genomas segmentados. Interesantemente, los genomas más grandes y con mayor número de segmentos, que presentan casos de duplicación, corresponden a virus que infectan plantas y que pertenecen a clases con un bajo promedio de longitud del genoma y una baja proporción de genomas segmentados (Tablas 8 y 9). Esto podría indicar que la presencia de segmentos es más dependiente de la interacción con el hospedero que de la clase viral. En el caso de los Benyvirus, que infectan plantas, se pueden observar hasta cinco segmentos. Los otros casos de segmentación son los de un virus de la familia Closteroviridae y el de la familia Secoviridae, ambos con dos segmentos (Tabla 8). Interesantemente, en los Closteroviridae y Secoviridae, que infectan plantas, así como en los Dicistroviridae, que infectan invertebrados, se han seleccionado duplicaciones que pueden acelerar el ciclo de replicación viral (Tabla 10). Aunque las plantas no tienen un sistema

inmune adaptativo como el de los vertebrados, es posible que sus defensas antivirales de silenciamiento de RNA y la resistencia sistémica adquirida (ASR) (Zvereva & Pooggin, 2012) generen suficiente presión sobre estos virus. También podría intuirse que la segmentación de genomas reduce la velocidad del ciclo de replicación viral, problema que sería mitigado por las duplicaciones de las cápsides. El caso de los Secoviridae, los Dicistroviridae y los Picornaviridae podría estar mejor explicado por la ancestría común de estos virus. Aun así, se puede apreciar una mayor longitud de genomas en las familias que infectan plantas e invertebrados, respectivamente (Tabla 8 y Figura 14).

En síntesis, podemos decir que los hospederos pueden generan condiciones que favorecen estrategias que seleccionan el aumento en el tamaño del genoma sin: 1) incrementar la acumulación de mutaciones o 2) el tiempo de generación del virión. La primer estrategia es la segmentación de genomas, que permite el crecimiento del genoma sin aumentar la tasa de acumulación de mutaciones. (Figura 15A). La segunda estrategia es la duplicación de genes de proteínas que aceleran el ciclo de replicación viral, como las proteínas de cápside (Figura 15B). La duplicación de las proteínas de la cápside además podría proporcionar una mayor variación del epítopo debido a un mayor número de copias y longitud del genoma.

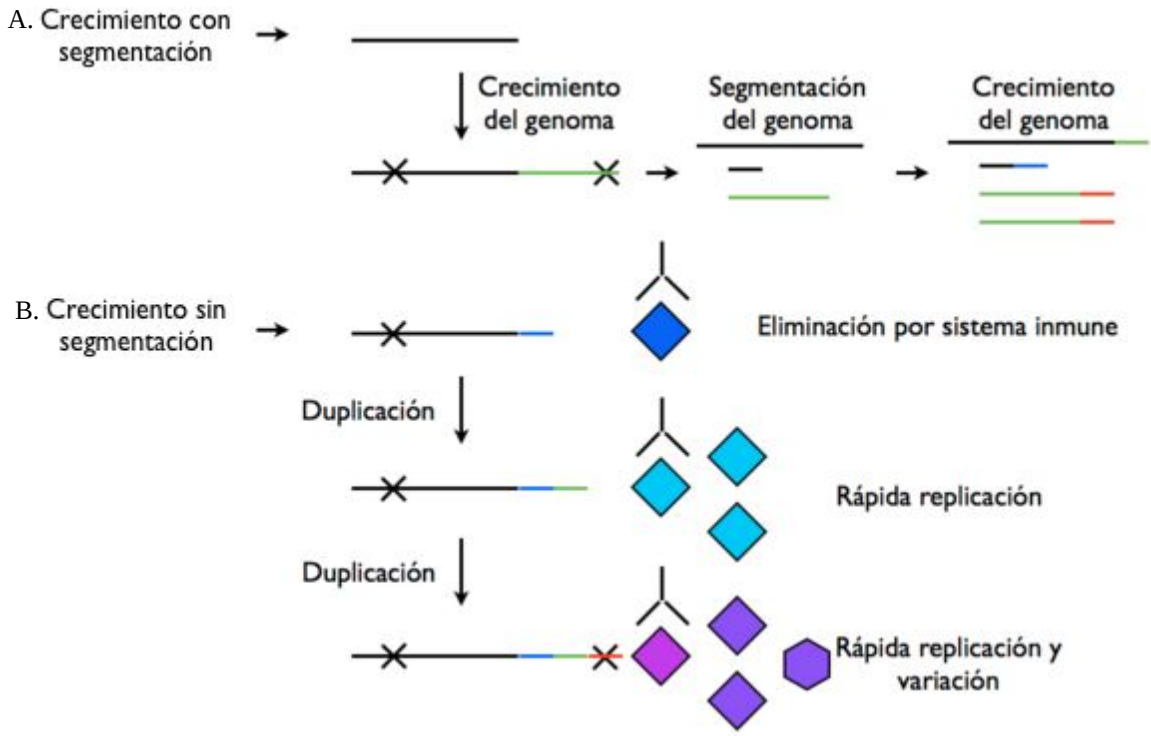


Figura 15. Mecanismos que permiten el crecimiento del genoma sin comprometer la adecuación del virus. (A) Crecimiento con segmentación del genoma. (B) Crecimiento sin segmentación del genoma, con selección de duplicaciones que aceleran el ciclo de replicación viral. La acumulación de mutaciones está señalada como una equis sobre las secuencias.

4.7 Destinos funcionales de los genes duplicados en virus de RNA

Bajo algunos modelos se acepta que uno de los resultados funcionales más comunes luego de la duplicación es la redundancia. Esta redundancia permite una relajada presión de selección y por lo tanto una rápida evolución y divergencia funcional entre las copias. Sin embargo, aunque se ha demostrado una mayor dispensabilidad de los genes redundantes, como una menor frecuencia de la redundancia en comparación con otros destinos funcionales, también se ha demostrado que, en ocasiones, esta redundancia resulta adaptativa y es conservada (Kafri & Pilpel, 2010). En contraste con lo que se ha demostrado con un estudio de evolución experimental sobre una disminuida adecuación en genomas con duplicaciones redundantes (Willemsen *et al.* 2016), en la mayoría de los casos presentados el resultado funcional que se ha mantenido parece ser la redundancia, que en algunos casos podría estar seleccionada para acelerar el ciclo de replicación viral. Sin embargo, la redundancia de p25 y p26 del *Beet necrotic yellow vein virus* (BNYVV) del género Benyvirus, que infecta a plantas del género

Beta y es transmitido por el protista *Polymyxa betae,* podría estar relacionada con un incremento en la severidad de las infecciones (Link *et al.* 2005). Interesantemente, la ubicuidad de las duplicaciones de proteínas de la cápside podría estar explicada por algunas evidencias que parecen indicar que las proteínas que forman homómeros son más duplicables que aquellas que forman heterómeros (Yang *et al.* 2003). Por otro lado, mientras que en el caso de la neofuncionalización de las proteasas también podría acelerarse el ciclo de replicación, la neofuncionalización de los dominios KP6 permiten el establecimiento de la simbiosis con hongos que compiten con otras cepas no resistentes a la toxina. Otro destino fue la subfuncinoalización de la proteina vpx que permite que vpr cumpla sus otras funciones sin descuidar la integración del genoma viral. Interesantemente, se ha sugerido que los genes complejos (longitud, número de dominios, funciones, regulación) se duplican con más frecuencia porque pueden ser retenidos por subfuncionalización (He & Zhang, 2005). Por último, es posible que la pseudogenización de Gns sea un ejemplo de un acontecimiento común en virus de RNA debido a sus limitaciones genómicas (Tabla 10).

Tabla 10. Hospederos y destino funcional de las duplicaciones de las duplicaciones reportadas hasta ahora. Basado en Simon-loriere & Holmes (2013) y ViralZone www.expasy.org/viralzone, Swiss Institute of Bioinformatics. La gama de hospederos está limitada a la de cada tipo viral en el que se han encontrado duplicaciones.

Clasificación	Familia	Hospedero	Gen duplicado	Funcion	Destino funcional
(+)ssRNA	Closteroviridae	Plantas	CP -> CPm	Ensamblaje de la cápside	Redundancia
			CPm1 -> CPm2	Ensamblaje de la cápside	Redundancia
	Picornaviridae	Vertebrados	Vpg -> Vpg	Inicio de la replicación	Redundancia
			VP1 -> VP3 y VP2	Ensamblaje de la cápside	Redundancia
			3C -> 2A	Maduración de la poliproteína e inhibición de la transcripción/inhibición de la traducción	Neofuncionalización
	Dicistroviridae	Invertebrados	VP1 -> VP3 y VP2	Ensamblaje de la cápside	Redundancia
	Secoviridae	Plantas	CPS -> CPL1 y 2	Ensamblaje de la cápside	Redundancia
	Benyvirus	Plantas	p25 -> p26	Factor de patogenicidad	Redundancia
(-)ssRNA	Rhabdoviridae	Vertebrados/ vector invertebrado	G -> Gns	Reconocimiento celular y endocitosis/desconocida	Neofuncionalización /Pseudogenización
			U1 -> U2	Desconocida	Desconocido
(RT)ssRNA	Retroviridae	Vertebrados	orfA -> orfB	Proliferación celular y regulación de la transcripción viral	Redundancia
			orf1 -> orf2	Desconocida	Desconocido
			vpr -> vpx	Supresión del ciclo celular y translocación del complejo de preintgración/translocación del complejo de preintgración	Subfuncionalización
dsRNA	Totiviridae	Hongos	KP6α -> KP6β	Reconocimiento celular/ citotoxicidad	Neofuncionalización

Los resultados obtenidos sugieren que a pesar de las restricciones que limitan el tamaño de los genomas de virus de RNA, estos pueden crecer por fijación de duplicaciones, o bajo ciertas circunstancias, de genes transferidos de manera horizontal. La capacidad de crecer de estos genomas tan estrictamente limitados, permite suponer que el crecimiento de los genomas durante el mundo de RNA-proteínas pudo darse de manera similar. A pesar de que

algunas evidencias indican que los virus de RNA son de origen reciente (Campillo-Balderas *et al*. 2015), estos podrían proporcionar indicios valiosos sobre la dinámica de los genomas previo a la aparición de genomas de DNA (Reyes-Prieto *et al.* 2012). Sin embargo, al hacer la extrapolación hacia el mundo de RNA-proteínas, es importante destacar algunas de las posibles diferencias entre virus de RNA y la células ancestrales. Por ejemplo, el hábito parásito de los virus, que podría ser otro factor que limita el tamaño de su genoma, no habría afectado a las celulas del mundo de RNA-proteínas. Por otro lado, es posible que el crecimiento de los genomas en el mundo de RNA-proteínas haya dependido del origen de enzimas como las exonucleasas y las helicasas. Por ejemplo, dentro de la clasificación de virus de cadena sencilla positiva, cuyo promedio de tamaño de genoma es menor a la media de virus de RNA, se encuentra la familia Coronaviridae, que incluye a algunos de los virus con genoma más largo (hasta 31kb), quizás debido a la presencia de una helicasa capaz de desdoblar las estructuras secundarias del RNA para su replicación (Holmes, 2009). También es posible que en el mundo RNA-proteínas se hubiesen dado organizaciones de genomas como los genomas segmentados y las basadas en cadena sencilla de los virus de RNA (Reyes-Prieto *et al.* 2012). Sin embargo, esta posibilidad también depende del origen de las exonucleasas que reducen la tasa de mutación, y de las helicasas que permiten la apertura de dobles hélices. Al parecer la presencia de las dobles hélices es importante para que pueda crecer el genoma. Por ejemplo, a pesar de que son replicados por las mismas polimerasas, los virus de DNA de doble cadena tienen genomas que alcanzan hasta 2473kb (Pandoravirus) mientras que los virus de DNA de cadena sencilla se comportan básicamente como virus de RNA (Holmes, 2009). Esto se debe a que el mecanismo de corrección es dependiente de doble cadena. De manera consistente, el único reporte encontrado en la literatura sobre un caso de duplicación en un virus de DNA de cadena sencilla es el de una proteína de cápside en un virus bipartito de la familia Geminiviridae. La copia se encuentra en un segmento diferente y no es encapsidado, pero se ha sugerido que está relacionada con la diseminación viral (Boyko, 1992). En contraste, en virus de DNA de doble cadena, se ha observado que la duplicación de genes es un evento recurrente (Holmes, 2009). Por ejemplo, se han encontrado hasta quince genes duplicados dentro de un mismo genoma de la familia Poxviridae (Hughes & Friedman, 2005).

5. CONCLUSIONES

El alineamiento de estructuras terciarias logró identificar casos de duplicación que no se detectan cuando se utilizan otras metodologías como el alineamiento de estructuras primarias. En algunos casos, la falta de identidad no permitió alinear las secuencias, mientras que en otros, se generaron alineamientos no significativos. A pesar de que el alineamiento de estructuras terciarias es adecuado para detectar homologías, en esta tesis se encontraron pocos casos de duplicaciones. Una limitación de este trabajo es la falta de estructuras resueltas para proteínas virales, sobre todo en el caso de dsRNA y (-)ssRNA, cuyos genomas tienen características que permiten suponer la presencia de duplicaciones. Otra razón podría estar relacionada con las restricciones impuestas por las presiones que claramente seleccionan genomas cortos en virus de RNA. Recientemente, también se ha sugerido a partir de trabajos de evolucion experimental, que la baja adecuacion que resulta de la acumulacion de duplicaciones podria tener una relación con el incremento en el costo energético para la producción de más proteínas y con el incorrecto procesamiento de la poliproteína (Willemsen *et al*. 2016). De cualquier forma, la comparación de estructuras terciarias resultó ser efectiva para detectar proteínas relacionadas en genomas con altas tasas de mutación. Este hecho abre nuevas posibilidades en el uso de herramientas para la detección de proteínas homólogas que no pueden ser identificadas por comparación de secuencias.

Los productos de las duplicaciones suelen estar adyacentes en el genoma viral y en pocos casos separados o en segmentos diferentes. Esto podría estar relacionado con los mecanismos propuestos para generar las duplicaciones, como la recombinación y el reclutamiento de segmentos homólogos. Sin embargo, debido a que los mecanismos que generan duplicaciones también pueden generar transferencia horizontal de genes, la ubicación de los genes no es evidencia suficiente para distinguir entre parálogos y xenólogos en genomas virales. Para solucionar este problema se podrían realizar análisis filogenéticos o matrices de similitud basadas en comparaciones de estructuras terciarias.

A pesar de que la relación observada entre el tamaño del genoma y el número de duplicaciones no coincide con lo esperado, posiblemente por sesgos de muestreo, es posible observar que la ocurrencia de duplicaciones se ve reflejada en el tamaño de los genomas

largos. Por otro lado, se observaron excepciones que, sin embargo, parecen forman parte de una tendencia sobre los patrones de retención de genes duplicados. Por ejemplo, que las duplicaciones son mejor retenidas en genomas segmentados, porque permiten el crecimiento del genoma sin incrementar la tasa de fijación de mutaciones; o cuando estas permiten la rápida conclusión del ciclo de replicación viral, bajo presión del sistema inmune del hospedero. Otras duplicaciones también son retenidas cuando una enzima tiene múltiples funciones, en este caso relacionadas con inhibición de mecanismos de defensa intracelulares, cuando estas permiten incrementar la virulencia o incluso cuando permiten establecer relaciones simbióticas. Cabe destacar la relación que hay entre algunas duplicaciones con el incremento en la patogenicidad de ciertos virus. La evolución experimental ha demostrado que cuando un virus recupera su adecuación, en este caso por la deleción de genes duplicados, se recupera la patogenicidad del mismo (Willemsen *et al*. 2016). Esto parece indicar que hay una relación entre la adecuación y la patogenicidad de los virus, lo cual señalaría la relevancia de las duplicaciones retenidas que por sí solas incrementan la severidad de las infecciones. Debido a que la patogenicidad y virulencia pueden estar asociados con la transmisibilidad de algunos virus (Weiss, 2002), valdría la pena enfocar esfuerzos sobre el estudio de las duplicaciones en virus de RNA causantes de enfermedades emergentes.

Los resultados de esta tesis en conjunto con otros trabajos demuestran que los genomas de RNA conservan más genes duplicados de lo que se tenía pensado, lo cual indica que a pesar de las restricciones que limitan su tamaño, bajo ciertas circunstancias, estos genomas son susceptibles de crecer. Incluso es factible pensar que en un futuro, con el crecimiento de las bases de datos y el refinamiento de las técnicas de comparación de estructuras terciarias, se puedan encontrar más casos de duplicación. Sin embargo, el crecimiento de estos genomas podría depender del reclutamiento de enzimas con actividad de edición, mismas que dependen de la presencia de dobles hélices y helicasas capaces de abrirlas, pues la presencia de estos mecanismos podrían explicar la diferencia en el tamaño de los genomas entre los virus de dsDNA y los ssDNA.

6. AGRADECIMIENTOS

Se agradece el apoyo financiero brindado por el PAPIIT-UNAM (IN223916), así como la colaboración académica de Sara Islas Graciano, Ricardo Hernandez Morales, Alejandro Rodrigo Jacome Ramirez & Emeline Oueda Cruz.

7. REFERENCIAS

1. Agol, V. 1997. Recombination and Other Genomic Rearrangements in Picornaviruses. *Seminars in Virology,* 8. 77-84
2. Allen A., Chatt E. & Smith T. J. 2013. The Atomic Structure of the Virally Encoded Antifungal Protein, KP6. *J. Mol. Biol.* 425: 609-621.
3. Allison, A., B., Mead, D., G., Palacios, G., F., Tesh, R., B. & Holmes, E., C. 2014. Gene duplication and phylogeography of North American members of the Hart Park serogroup of avian rhabdoviruses. *Virology* 448:284–292
4. Andersson, D., I. & Hughes, D. 2009. Gene Amplification and Adaptive Evolution in Bacteria. *Annu. Rev. Genet*. 43:167–95.
5. Andersson, D., I., Jerlstrom-Hultqvist, J. & Nasvall, J. 2015. Evolution of New Functions De Novo and from Preexisting Genes. *Cold Spring Harb Perspect Biol*.
6. Artymuk, P., J., Grindley, H., M., Kumar, K., Rice, D., W. & Willett, P. 1993. Three-dimensional structural resemblance between the ribonuclease H and connection domains of HIV reverse transcriptase and the ATPase fold revealed using graph theoretical techniques. *FEBS.* 324: 1. 15-21
7. Barker, W., C., Ketcham, L., K. & Dayhoff, M., O. 1978. A Comprehensive Examination of Protein Sequences for Evidence of Internal Gene Duplication. *J. Mol. Evol.* 10, 265-281.
8. Barrett, A., J. & Rawlings, N., D. 2001. Evolutionary Lines of Cystein Peptidases. *Biol. Chem.* 382. 727-733.
9. Bazan, J., F. & Fletterick, R., J. 1988. Viral cysteine proteases are homologous to the trypsin-like family of serine proteases: Structural and functional implications. *Proc. Natl. Acad. Sci*. 85. 7872-7876.

10. Blasdell, K., R., Voysey, R., Bulach, D., Joubert, D., A., Tesh, R., B., Boyle, D., B. & Walker, P., J. 2012. Kotonkan and Obodhiang viruses: African ephemeroviruses with large and complex genomes. *Virology* 425:143–153.
11. Boyko, V., P, Karasev, A., V, Agranovsky, A., A, Koonin, E., V. & Dolja, V., V. 1992. Coat protein gene duplication in a filamentous RNA virus of plants. *Proc. Natl. Acad. Sci.* 89:9156–9160.
12. 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*:143.
13. Chen Ch, Li W & Sung H. 2007. Patterns of internal gene duplication in the course of metazoan evolution. *Gene,*396:59-65.
14. Conant, G. & Wolfe, K. 2008. Turning a hobby into a job: How duplicated genes find new functions. *Nature*,9:938-950.
15. Crick, F., H., C. 1958. The Biological Replication of Macromolecules. *Symp. Soc. Exp. Biol.* XII, 138.
16. Crick, F., H., C. 1970. Central Dogma of Molecular Biology. *Nature*, 22. 561-563.
17. 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.
18. Fazeli, C., F. & Rezaian, M., A. 2000. Nucleotide sequence and organization of ten open reading frames in the genome of Grapevine leafroll-associated virus 1 and identification of three subgenomic RNAs. *Journal of General Virology*, 81, 605–615.
19. Fletcher, T., M., Brichacek, B., Sharova, N., Newman, M., A., Stivahtis, G., Sharp, P., M., Emerman, M., Hahn, B., H. & Stevenson, M. 1996. Nuclear import and cell cycle arrest functions of the HIV-1 Vpr protein are encoded by two separate genes in HIV-2/SIV(SM). *EMBO. J.* 15 (22), 6155–6165.
20. Forss, S. & Schaller, H. 1982. A tandem repeat gene in a picornavirus. *Nucleic Acids Res*. 10:6441–6450.
21. Gerstein M, Bruce C, Rozowsky J, Zheng D, Du J, Korbel J, Emanuelsson O, Zhang Z, Weissman S & Snyder M. 2007. What is a gene, post-ENCODE? History and updated definition. Cold Spring Harbor Laboratory Press. 17:669-681.
22. Gorbalenya, A., E., Donchenko, A., P., Blinov, V., M. and Koonin, E., V. 1989. Cysteine proteases of positive strand RNA viruses and chymotrypsin-like serine

proteases: A distinct protein superfamily with a common structural fold. *FEBS Lett.* 243. 103-114.

23. Gregory R. 2005. The Evolution of the Genome. Elsevier Academic Press. USA. 768pp.
24. Gubala, A., Davis, S., Weir, R., Melville, L., Cowled, C., Walker, P. & Boyle, D. 2010. Ngaingan virus, a macropod-associated rhabdovirus, contains a second glycoprotein gene and seven novel open reading frames. *Virology* 399:98–108.
25. He, X. & Zhang, J. 2005. Gene Complexity and Gene Duplicability. *Current Biology,* Vol. 15, 1016–1021.
26. Holm, L. & Rosenström, P. 2010. Dali server: conservation mapping in 3D. *Nucl. Acids Res.* 38. 545-549.
27. Holmes E. 2009. *The Evolution and Emergence of RNA Viruses.* Oxford University Press. UK. 267pp.
28. Hughes, A., L., Ekollu, V., Friedman, R. & Rose, J,. R. 2005. Gene Family Content-Based Phylogeny of Prokaryotes: The Effect of Criteria for Inferring Homology. *Syst. Biol.* 54(2):268–276.
29. Hughes, A., L. & Friedman, R. 2005. Poxvirus genome evolution by gene gain and loss. Molecular Phylogenetics and Evolution. 35. 186-195.
30. Hughes, A., L. & Friedman, R. 2010. Myths and REalities of Gene Duplication. In: Dittmar, K. & Liberles, D. *Evolution after gene duplication*. Wiley-Blackwell. 329pp.
31. Johnson, C., M. & Grossman, A., D. 2015. Integrative and Conjugative Elements (ICEs): What They Do and How They Work *Annu. Rev. Genet.* 49:13.1–13.25.
32. Kaessmann H. 2010. Origins, evolution and phenotypic impact of new genes. Cold Spring Harbor Laboratory Press. 20. 1313-1326.
33. Kafri, R. & Pilpel, T. 2010. Evolutionary and Funcional Aspects of Genetic Redundancy. In: Dittmar, K. & Liberles, D. *Evolution after gene duplication*. Wiley-Blackwell. 329pp.
34. Kambol R, Kabat P, Tristem M. 2003. Complete nucleotide sequence of an endogenous retrovirus from the amphibian, Xenopus laevis. *Virology* 311:1–6.
35. Karasev, A., V., Boyko, V., P., Gowda, S., Nikolaeva, O., V., Hilf, M., E., Koonin, E., V., Niblett, C., L., Cline, K., Gumpf, F., J., Lee, R., F., Garnseg, S., M.,

Lewandowski, D., J. & Dawson, W., O. 1995. Complete sequence of the citrus tristeza virus RNA genome. *Virology* 208:511–520.

36. Kreuze, J., F., Savenkov, E., I. & Valkonen, J., P., T. 2002. Complete genome sequence and analyses of the subgenomic RNAs of Sweet potato chlorotic stunt virus reveal several new features for the genus Crinivirus. *J Virol*. 76:9260–9270.
37. Lai, M., M., C. 1992. RNA Recombination in Animal and Plant Viruses. *Microbiological Reviews.* 56(1): 61-79.
38. LaPierre, L., A., Holzschu, D., L., Bowser, P., R. & Casey, J., W. 1999. Sequence and transcriptional analyses of the fish retroviruses walleye epidermal hyperplasia virus types 1 and 2: evidence for a gene duplication. *J Virol*. 73:9393–9403.
39. Le Guillou-Guillemettea, H., Ducancellea, A., Bertrais, S., Lemairec, C., Pivert, A., Veillona, P., Bouthrya, E., Alaind, S., Thibault, V.,Abravanelf, F., Rosenbergg, A., R., Henquell, C., André-Garnier, E., O. Petsaris, O., Valletj, S., Bour, J., B., Baazial, Y., Trimoulet, P., André, P., Gaudy-Graffino, C., Bettinger, D., Larrat, S., Signori-Schmuck, A., Saoudinr, H., Pozzettor, B., Lagathus, G., Minjolle-Chas, S., Stoll-Keller, F., Pawlotsky, J., M., Izopetf, J., Payanj, C. & Lunel-Fabiani, F. Identification of a duplicated V3 domain in NS5A associated with cirrhosis and hepatocellular carcinoma in HCV-1b patients. *Journal of Clinical Virology* 69:203–209.
40. Lesk, A., M. & Fordham, W., D. 1996. Conservation and Variability in the Structures of Serine Proteinases of the Chymotrypsin Family. *J. Mol. Biol*. 258, 501–537.
41. Levinson G & Gutman G. 1987. Slipped-Strand Mispairing: A Major Mechanism for DNA Sequence Evolution. Mol. Biol. Evol. 4(3):203-221.
42. Liljas L, Tate J, Christian P & Johnson J. 2002. Evolutionary and taxonomic implications of conserved structural motifs between picornaviruses and insecte picorna-like viruses. Arch Virol 147: 59–84.
43. Lin, T., Cavarelli, J. & Johnson, J., E. 2003. Evidence for assembly-dependent folding of protein and RNA in an icosahedral virus. *Virology* 314. 26–33.
44. Link, D., Schmidlin, L., Schirmer, A., Klein, E., Erhardt, M., Geldreich, A., Lemaire, O. & Gilmer, D. 2005. Functional characterization of the Beet necrotic yellow vein virus RNA-5-encoded p26 protein: evidence for structural pathogenicity determinants. *Journal of General Virology,* 86, 2115–2125

45. Lowry, K., Woodman, A., Cook, J. & Evans, D., J. 2014. Recombination in Enteroviruses Is a Biphasic Replicative Process Involving the Generation of Greater-than Genome Length 'Imprecise' Intermediates. *PLOSpathogens*. 42(6):1-17.
46. Lynch, M. & Conery, J., S. 2003. The Origins of Genome Complexity. *Science.* 302 (5649), 1401-1404.
47. Majorek, K., a., Dunin-Horkawicz, S., Steczkewicz, K., Muszewska, A., Nowotny, M., Ginalski, K. & Bujnicki, J., M. 2014. The RNase H-like superfamily: new members, comparative structural analysis and evolutionary classification. *Nucleic Acid Research,* Vol 42, No. 7. 4160-4179.
48. Malik H., S., & Eickbush T., H. 2001. Phylogenetic Analysis of Ribonuclease H Domains Suggests a Late, Chimeric Origin of LTR Retrotransposable Elements and Retroviruses. *Cold Spring Harbor Laboratory Press.* 11: 1187-1197.
49. Matlin A, Clark F & Smith C. 2005. Understanding Alternative Splicing: Towards a Cellular Code. Nature. 6. 386-398.
50. Mills, D. R., Peterson, R., L. & Spiegelman, S. 1967. An extracellular Darwinian experiment with a self-duplicating nucleic acid molecule. *Proceedings of the National Academy of Sciences* 58 (1): 217–24.
51. Nacher J, Hayashida M & Akutsu T. 2010. The role of internal duplications in the evolution of multi-domain proteins.BioSystems,101.127-135.
52. Neme R & Tautz, D. 2014. Evolution: Dynamics of De Novo Gene Emergence. Current Bioogy. 24(6). 238-240.
53. Ohno, S. 1970. *Evolution by Gene Duplication*. Springer-Verlag.160pp.
54. Olendzenski, L., Zhaxybayeva, O. & Gogarten, P. 2005. Orthologs, Paralogs and Xenologs in Human and Other Genomes. *eLS, John Wiley and Sons, Ltd.*
55. Petersen, J., F., W., Cherney M., M., Liebig, H-D,. Kuechler, E. & James, M., N., G. 1999. The structure of the 2A proteinase from a common cold virus: a proteinase responsible for the shut-off of host-cell protein synthesis. EMBO. Vol 18. No 20. 5463-5475.
56. Peng, C., W., Peremyslov, V., V., Mushegian, A., R., Dawson, W., O. & Dolja, V., V. 2001. Functional specialization and evolution of leader proteinases in the family Closteroviridae. *J Virol*. 75:12153–12160.

57. Porter, A., G. 1993. Picornavirus Nonstructural Proteins: Emerging Roles in Virus Replication and Inhibition of Host Cell Functions. *Journal of Virology*. 6917-6921
58. Reanney, D. 1982. The evolution of RNA viruses. *Ann. Rev. Microbiol.* 36: 47-73.
59. Reyes-Prieto, F., Hernandez-Morales, R., Jacome, R., Becerra, A. & Lazcano, A. 2012. Coenzymes, viruses and the RNA world. *Biochimie*. 94. 1467-1473.
60. Romero, D. & Palacios, R. 1997. Gene Amplification and Genomic Plasticity in Prokaryotes. *Annu. Rev. Genet*. 31:91–111.
61. Rose, S. 2016. How to Get Another Thorax: Epigenetics. *London Review of Books*. 38(17).
62. Rost, B. 1999. Twilight zone of protein sequence alignments. *Protein Engineering* 12(2):85–94.
63. 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.
64. Shirogane Y, Watanabe S & Yanagu Y. (2013) Cooperation: another mechanism of viral evolution. *Trends in Microbiology*.21(7). 320-324.
65. Sikic, K., Tomic, S. & Carugo, O. 2010. Systematic Comparison of Crystal and NMR Protein Structures Deposited in the Protein Data Bank. *The Open Biochemistry Journal*, 4, 83-95.
66. 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.
67. Simon-Loriere, E. & Holmes, E., C. 2012. Why do RNA virus recombine? *Nat. Rev. microbiol.* 9(8): 617–626.
68. Soucy S, Huang J & Gogarten P. 2015. Horizontal gene transfer: building the web of life. *Nature*. 16. 472-482.
69. Taylor, W., R. & Sadowski, M., I. 2010. Protein Products of Tandem Gene Duplication: A Structural View. In: Dittmar, K. & Liberles, D. *Evolution after gene duplication*. Wiley-Blackwell. 329pp.
70. 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. *Nature Structural Biology.* 6(8): 765-774.

71. Treangen, T., J. & Rocha, E., P., C. 2011. Horizontal Transfer, Not Duplication, Drives the Expansion of Protein Families in Prokaryotes. *PLoS Genet* 7(1).
72. Tristem, M., Marshall, C., Karpas, A., Petrik, J. & Hill, F. 1990. Origin of vpx in lentiviruses. *Nature* 347:341–342.
73. Tzanetakis, I., E. & Martin, R., R. 2007. Strawberry chlorotic fleck: identification and characterization of a novel Closterovirus associated with the disease. *Virus Res*. 124:88–94.
74. Tzanetakis, I., E., Postman, J., D. & Martin., R., R. 2005. Characterization of a novel member of the family Closteroviridae from Mentha spp. *Phytopathology* 95:1043–1048.
75. Valli, A., López-Moya, J., J. & García, J., A. 2007. Recombination and gene duplication in the evolutionary diversification of P1 proteins in the family Potyviridae. *J Gen Virol*. 88:1016–1028.
76. Wagner, A. 2010. On the Energy and Material Cost of Gene Duplication. In: Dittmar, K. & Liberles, D. *Evolution after gene duplication*. Wiley-Blackwell. 329pp.
77. Walker PJ, Byrne KA, Riding GA, Cowley JA, Wang Y, McWilliam S. 1992. The genome of bovine ephemeral fever rhabdovirus contains two related glycoprotein genes. *Virology* 191:49–61.
78. Weiss, R., A. 2002. Virulence and Pathogenesis. *TRENDS in Microbiology* 10(7). 314-317.
79. 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 Biology and Evolution.*
80. Xue, C., Huang, R., Maxwell, T., J. & Fu, Y., X. 2010. Genome Changes After Gene Duplication: Haploidy vs. Diploidy. *Genetics Society of America*. 186: 287–294.
81. Yang, J., Lusk, R. & Li, W., H. 2003. Organismal complexity, protein complexity, and gene duplicability. *PNAS,* 100(26). 15661–15665.
82. Zhang, J. 2003. Evolution by gene duplication: an update. *TRENDS in Ecology and Evolution* Vol.18 No.6.
83. Zvereva, A., S. & Pooggin, M., M. 2012. Silencing and Innate Immunity in Plant Defense Against Viral and Non-Viral Pathogens. *Viruses.* 4. 2578-2597.

ANEXOS

Anexo 1. Tabla completa con la información de los datos analizados.

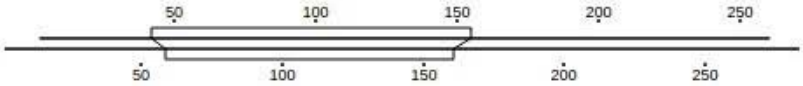
Clasificación	Familia	Número de tipos virales	Número de entradas PDB	Alineamientos significativos
dsRNA	Birnaviridae	3	6	N/S
	Cystoviridae	4	8	N/S
	Totiviridae	2	2	1
	Reoviridae	21	33	N/S
(RT)ssRNA	Retroviridae	19	54	N/S
(-)ssRNA	Arenaviridae	7	11	N/S
	Bornaviridae	1	2	N/S
	Bunyaviridae	14	17	N/S
	Filoviridae	2	10	N/S
	Orthomyxoviridae	3	17	N/S
	Paramyxoviridae	9	18	N/S
	Rhabdoviridae	5	9	N/S
(+)ssRNA	Alphatetraviridae	2	2	N/S
	Arteriviridae	2	8	N/S
	Astroviridae	3	3	N/S
	Bromoviridae	3	3	N/S
	Caliciviridae	5	9	N/S
	Coronaviridae	11	44	N/S
	Dicistroviridae	2	2	2
	Flaviviridae	11	36	N/S
	Leviviridae	4	6	N/S
	Nodaviridae	6	6	N/S
	Picornaviridae	9	23	3
	Potyviridae	3	3	N/S
	Secoviridae	3	3	2
	Sobemovirus	5	6	N/S
	Togaviridae	6	13	N/S
	Tombusviridae	2	2	N/S
	Tymoviridae	2	3	N/S
	Virgaviridae	1	2	N/S

Anexo 2. Regiones alineadas por comparación de secuencias con PRSS.

KP6α y KP6β (Totiviridae)

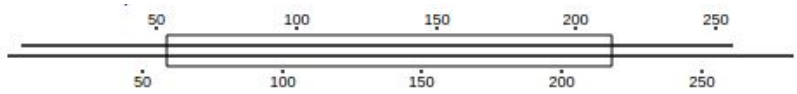

```
               10          20        30        40        50           60        70
4GVB:A NNAFCAGFGLSCKWECWC--TAHGTGNELRYATAAGCG-DHLSKSYYDARAGHCL--FSDDL--RNQFYSHCSSLNNNMS
                         : :  :..: : .:  .: :.:. : .  .::  . : :   .: ::   .:::. :
4GVB:B        GKRPRPVMCQCVDTTNG-GVRLDAVTRAACSIDSFIDGYYTEKDGFCRAKYSWDLFTSGQFYQACLRYSHAGT
                        10         20        30        40        50        60        70

            80
4GVB:A CRSLSKR

4GVB:B NCQPDPQYE
            80
```

VP1-VP3 (Dicistroviridae)



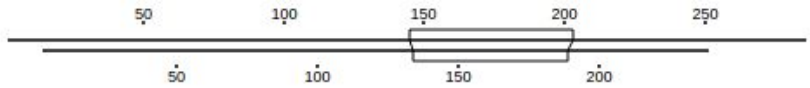

```
               10        20        30        40        50          60        70
1B35:A VMGEDQQIPRNEAQHGVHPISIDTHRISNNWSPQAMCIGEKVVSIRQLIKRFGI-FGDA-NTLQADGSSFVVAPFTVTSP
                                                .:.:: .   ::    .:. ::.  :.    :.:: :
1B35:C GRMANFDGMDMSHKMALSSTNEIETNEGLAGTSLDVMDLSRVLSIPNYWDRFTWKTSDVINTVLWDN---YVSPFKVKPY
        20        30        40        50        60        70        80           90

       80        90       100       110       120       130       140       150
1B35:A TKTLTSTRNYTQFDYYYYLYAFWRGSM--RIKMVAETQDGTGTPRKKTNFTWFVRMFNSLQDSFNSLISTSSSAVTTTVL
       . :.:.    :..      ...:::::   .:.: .::  .:  : . .:  .          .:. :::..  :. :
1B35:C SATITDRFRCTHMGKVANAFTYWRGSMVYTFKFV-KTQYHSG--RLRISFIPYY---------YNTTISTGTPDVSRTQK
          100       110       120       130       140               150       160

        160       170       180       190       200       210       220       230
1B35:A PSGTINMGPSTQVIDPIVEGLIEVEVPYYNISHIIPAVIIDDGIPSMEDYLKGHSPPCLIFSPRDSISAINHIIIASFM

1B35:C IVVDLRTSTAVSFTVPYIGSRPWLYCIRPESSWLSKDNTDGALMYNCVSGIVRVEVLNQLVAAQNVFSEIDVICEVNGGP
            170       180       190       200       210       220       230       240
```

VP2-VP3 (Dicistroviridae)



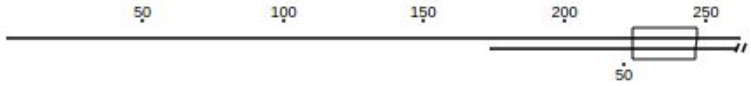

```
            20        30        40        50        60            70         80
1B35:B TSEQKEIVHFVSEGVTPSTTALPDIVNLSTNYLDKNTREDRIHSIKDFLSRPI-----IIATNLW-SVSDPVEKQLYTAN
                                                :. :: .. .:       .: : :: .  .: . . :.:.
1B35:C GRMANFDGMDMSHKMALSSTNEIETNEGLAGTSLDVMDLSRVLSIPNYWDRFTWKTSDVINTVLWDNYVSPFKVKPYSAT
        20        30        40        50        60        70        80        90

          90       100       110       120       130       140       150       160
1B35:B FPEVLISNAMYQDKLKGFVGLRATLVVKVQVNSQPFQQGRLMLQYIPYAQYMPNRVTLINETLQGRSGCPRTDLELSVGT
       . . .  . : .   ..:.  :...:   .  .  ...::: ...:::  :.    : :.      :   .  ..: ..:
1B35:C ITDRFRCTHMGK-VANAFTYWRGSMVYTFKFVKTQYHSGRLRISFIPY--YYN---TTISTGTPDVSRTQKIVVDLRTST
       100       110       120       130       140            150       160       170

        170       180       190       200       210       220       230       240
1B35:B EVEMRIPYVSPHLYYNLITGQGSFGSIYVVVYSQLHDQVSGTGSIEYTVWAHLEDVDVQYPTGANIFTGNEAYIKGTSRY
        : . .::.. . .   :  ..:. :   .  . ... :::   .:
1B35:C AVSFTVPYIGSRPWLYCIRPESSWLSKDNTDGALMYNCVSGIVRVEVLNQLVAAQNVFSEIDVICEVNGGPDLEFAGPTC
             180       190       200       210       220       230       240       250

        250
1B35:B DAAQKAHAA

1B35:C PRYVPYAGDFTLADTRKIEAERTQEYSNNED
             260       270       280
```

VP1-VP3 (Picornaviridae)




```
           110       120       130       140       150       160       170       180
1AYM:1 AQIRRKFEMFTYARFDSEITMVPSVAAKDGHIGHIVMQYMYVPPGAPIPTTRDDYAWQSGTNASVFWQHG-QPFPRFSLP
                                                :.:::   : .: .     ::.  : :. : :     . .:
1AYM:3 PLATTLIGEIASYFTHWTGSLRFSFMFCGTANTTLKVLLAYTPPGIGKPRSRKEAML--GTH--VVWDVGLQSTVSLVVP
             100       110       120       130       140         150         160

             190       200       210       220       230       240       250       260
1AYM:1 FLSIASAYYMFYDGYDGDTYKSRYGTVVTNDMGTLCSRIVTSEQLHKVKVVTRIYHKAKHTKAWCPRPPRAVQYSHTHTT
       ..: ..  .   : :..  :
1AYM:3 WISASQYRFTTPDTYSSAGYITCWYQTNFVVPPNTPNTAEMLCFVSGCKDFCLRMARDTDLHKQTGPITQ
      170       180       190       200       210       220       230
```

VP2-VP3 (Picornaviridae)




```
           190       200       210       220       230       240       250       260
1AYM:2 LIFPHQFINLRSNNSATLIVPYVNAVPMDSMVRHNNWSLVIIPVCQLQSNNISNIVPITVSISPMCAEFSGARAKTVVQ
                                                .::. . ::: :.:.   ::..::
1AYM:3 QFMTTDDMQSPCALPWYHPTKEIFIPGEVKNLIEMCQVDTLIPINSTQSN-IGNVSMYTVTLSPQTKLAEEIFAIKVDIA
             20        30        40        50        60         70        80        90

1AYM:3 SHPLATTLIGEIASYFTHWTGSLRFSFMFCGTANTTLKVLLAYTPPGIGKPRSRKEAMLGTHVVWDVGLQSTVSLVVPWI
             100       110       120       130       140       150       160       170
```

Proteasa 3C y 2A (Picornaviridae)


```
            110       120       130       140       150       160       170
1CQQ:A NLALLANQPEPTIINVGDVVSYGNILLSGNQTARMLKYSYPTKSGYCGGVLYKIGQVLGI-HVGGNGRDGFSAMLLRSYF
                                                : . : ::: :     :.::  .::... .:
2HRV:A YCKHKNRYFPITVTSHDWYEIQESEYYPKHIQYNLLIGEGPCEPGDCGGKLLCKHGVIGIVTAGGDNHVAFIDLRHFHCA
      60        70        80        90       100       110       120       130

     180
1CQQ:A T

2HRV:A EEQ
     140
```

Anexo 3. Tablas de datos para elaboración de las figuras 12 y 13.

Tabla de información para la elaboración de la figura 12.

Clase	Promedio de longitud de genoma por clase	Número de duplicaciones
(RT)ssRNA	8,383	3
(+)ssRNA	9,594	7
dsRNA	11,771	1
(-)ssRNA	13,264	2

Tabla de información para la elaboración de la figura 13.

Clase	Promedio de longitud de genoma por clase	Promedio de longitud de genoma de virus con duplicaciones, por clase
(RT)ssRNA	8,383	10,70833333
(+)ssRNA	9,594	14,15806667
dsRNA	11,771	6,099
(-)ssRNA	13,264	14,9165

Anexo 4. Información general de los virus y sus hospederos. Información tomada de Campillo-Balderas *et al.* (2015), Simon-Loriere & Holmes (2013) y ViralZone www.expasy.org/viralzone, Swiss Institute of Bioinformatics.

Clasificación	Familia	Especie	Tamaño del genoma	Número de segmentos	Hospedero	Sistema inmune	Duplicaciones
dsRNA	Totiviridae	UmV	6	1	Hongos	innato	KP6a->KP6b
(+)ssRNA	Picornaviridae	Rhinovirus A	7,152	1	Vertebrados	adaptativo	3C-2A
(+)ssRNA	Picornaviridae	HRV-A	7,152	1	Vertebrados	adaptativo	VP3-VP2-VP1
(+)ssRNA	Picornaviridae	FMDV	8,134	1	Vertebrados	adaptativo	Vpg-Vpg
(RT)ssRNA	Retroviridae	HIV-2	9	1	Vertebrados	innato	vpr-vpx
(+)ssRNA	Dicistroviridae	CrPV	9,185	1	Invertebrados	innato	VP3-VP2-VP1
(+)ssRNA	Secoviridae	BPMV	9,657	2	Plantas	innato	VP3-VP2-VP1
(RT)ssRNA	Retroviridae	Xen1	10	1	Vertebrados	adaptativo	orf1-orf2
(RT)ssRNA	Retroviridae	WEHV-2	13,125	1	Vertebrados	adaptativo	orfA-orfB
(-)ssRNA	Rhabdoviridae	Wongabel virus	13,196	1	Vertebrado/vector invertebrado	adaptativo/innato	U1-U2
(-)ssRNA	Rhabdoviridae	BEFV	14,9	1	Vertebrado/vector invertebrado	adaptativo/innato	G-GNS
(+)ssRNA	Closteroviridae	LCV 2	15,045	1	Plantas	innato	Cp-Cpm
(+)ssRNA	Closteroviridae	MV 1	15,45	1	Plantas	innato	Cp-Cpm
(+)ssRNA	Closteroviridae	BYV	15,48	1	Plantas	innato	Cp-Cpm
(-)ssRNA	Rhabdoviridae	NGAV	15,7	1	Vertebrado/vector invertebrado	adaptativo/innato	G-GNS
(-)ssRNA	Rhabdoviridae	KOTV	15,87	1	Vertebrado/vector invertebrado	adaptativo/innato	G-GNS
(+)ssRNA	Benyviridae	BNYVV	15,914	5	Plantas	innato	p25-p26
(+)ssRNA	Closteroviridae	SCF-AV	17,039	1	Plantas	innato	Cp-Cpm
(+)ssRNA	Closteroviridae	SPCSV	17,63	2	Plantas	innato	Cp-Cpm
(+)ssRNA	Closteroviridae	GLRaV 3	17,919	1	Plantas	innato	Cp-Cpm
(+)ssRNA	Closteroviridae	GLRaV 1	18,659	1	Plantas	innato	Cp-Cpm
(+)ssRNA	Closteroviridae	GLRaV 1	18,659	1	Plantas	innato	Cpm1-Cpm2
(+)ssRNA	Closteroviridae	CTV	19,296	1	Plantas	innato	Cp-Cpm