UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

POSGRADO EN CIENCIAS BIOLÓGICAS

FACULTAD DE CIENCIAS

BIOLOGÍA EVOLUTIVA

ANÁLISIS FUNCIONAL DE LOS DOMINIOS DE UNIÓN A RNA EN PROTEOMAS VIRALES

TESIS

QUE PARA OPTAR POR EL GRADO DE:

MAESTRO EN CIENCIAS BIOLÓGICAS

PRESENTA:

GERMÁN HERNÁNDEZ ALONSO

TUTOR PRINCIPAL DE TESIS: DR. ARTURO CARLOS II BECERRA BRACHO
FACULTAD DE CIENCIAS, UNAM
COMITÉ TUTOR: DR. ANTONIO EUSEBIO LAZCANO ARAUJO REYES
FACULTAD DE CIENCIAS, UNAM
DR. GABRIEL LÓPEZ VELÁZQUEZ
INSTITUTO NACIONAL DE PEDIATRÍA

CD. MX. NOVIEMBRE 2018

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

POSGRADO EN CIENCIAS BIOLÓGICAS

FACULTAD DE CIENCIAS

BIOLOGÍA EVOLUTIVA

ANÁLISIS FUNCIONAL DE LOS DOMINIOS DE UNIÓN A RNA EN PROTEOMAS VIRALES

TESIS

QUE PARA OPTAR POR EL GRADO DE:

MAESTRO EN CIENCIAS BIOLÓGICAS

PRESENTA:

GERMÁN HERNÁNDEZ ALONSO

TUTOR PRINCIPAL DE TESIS: DR. ARTURO CARLOS II BECERRA BRACHO
FACULTAD DE CIENCIAS, UNAM
COMITÉ TUTOR: DR. ANTONIO EUSEBIO LAZCANO ARAUJO REYES
FACULTAD DE CIENCIAS, UNAM
DR. GABRIEL LÓPEZ VELÁZQUEZ
INSTITUTO NACIONAL DE PEDIATRÍA

MÉXICO, CD. MX. NOVIEMBRE, 2018

UNAM
POSGRADO
Ciencias Biológicas

COORDINACIÓN

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

OFICIO FCIE/DAIP/1097/2018

ASUNTO: Oficio de Jurado

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

Me permito informar a usted que en la reunión ordinaria del Comité Académico del Posgrado en Ciencias Biológicas, celebrada el día **1 de octubre de 2018** se aprobó el siguiente jurado para el examen de grado de **MAESTRO EN CIENCIAS BIOLÓGICAS** en el campo de conocimiento de **Biología Evolutiva** del alumno **HERNÁNDEZ ALONSO GERMÁN** con número de cuenta **411071165** con la tesis titulada **"Análisis funcional de los dominios de unión a RNA en proteomas virales"**, realizada bajo la dirección del **DR. ARTURO CARLOS II BECERRA BRACHO:**

Presidente:	**DRA. CLAUDIA SELENE ZÁRATE GUERRA**
Vocal:	**DRA. ANA LORENA GUTIÉRREZ ESCOLANO**
Secretario:	**DR. ANTONIO EUSEBIO LAZCANO-ARAUJO REYES**
Suplente:	**DR. SANTIAGO ÁVILA RÍOS**
Suplente:	**DR. ALEJANDRO RODRIGO JÁCOME RAMÍREZ**

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

A T E N T A M E N T E
"POR MI RAZA HABLARA EL ESPÍRITU"
Ciudad Universitaria, Cd. Mx., a 13 de noviembre de 2018

DR. ADOLFO GERARDO NAVARRO SIGÜENZA
COORDINADOR DEL PROGRAMA

POSGRADO EN CIENCIAS BIOLOGICAS
COORDINACIÓN

AGNS/VMVA/ASR/ipp

AGRADECIMIENTOS INSTITUCIONALES

Al Posgrado en Ciencias Biológicas de la Universidad Nacional Autónoma de México (UNAM), así como a la Facultad de Ciencias, UNAM.

Al Consejo Nacional de Ciencia y Tecnología (CONACYT), por el apoyo económico recibido, el cual permitió el llevar acabo éste proyecto de maestría (CVU: 576734, No. de Apoyo: 449573, No. Registro Becario: 608502). Nuevamente, agradezco al Posgrado por la asignación del apoyo económico PAEP, por el cual me fue posible el asistir al X Congreso Nacional de Virología (octubre 2017) y al congreso internacional Science in Early Life (junio 2018).

A mi tutor principal, el Dr. Arturo Carlos II Becerra Bracho por darme la oportunidad de realizar éste proyecto, así como por su apoyo incondicional, paciencia, comprensión y continuas enseñanzas. También le doy las gracias al Dr. Antonio E. Lazcano Araujo Reyes, jefe del Laboratorio de Origen de la Vida, por su apoyo y por su invaluable conocimiento compartido. Igualmente, agradezco a todos los miembros del Laboratorio de Origen de la Vida, con especial mención a los técnicos académicos Ricardo Hernández Morales, José Alberto Campillo Balderas y Sara E. Islas Graciano.

Al los miembros del Comité Tutor, el Dr. Gabriel López Velázquez y el Dr. Antonio E. Lazcano Araujo Reyes, por su disposición y comentarios durante el desarrollo de éste trabajo.

Por último, agradezco a los miembros del Jurado por su revisión y comentarios, mismos que fueron de gran importancia para el mejoramiento de la tesis aquí presentada.

AGRADECIMIENTOS A TÍTULO PERSONAL

A la Universidad Nacional Autónoma de México, que me ha dado el espacio y las oportunidades de crecer en aspectos personales, académicos e intelectuales.

A mi tutor, el Dr. Arturo Becerra, le agradezco profundamente la confianza que ha puesto en mí al aceptarme como su alumno sin siquiera conocerme. Igualmente, le agradezco por su cordialidad y sencillez, pero sobretodo por su amistad.

Al Dr. Antonio Lazcano, con quien me siento en deuda por todas sus enseñanzas, le agradezco, igualmente, por su sencillez en el trato a los alumnos, su apoyo y amistad.

Agradezco a todos los miembros del Laboratorio de Origen de la Vida, los Macacos, por su compañerismo y todo el conocimiento compartido, gracias al cual, esta etapa académica llega a buen fin.

Igualmente, agradezco a todos mis amigos y compañeros que han estado presentes, y con quienes he compartido este proceso.

Por último, quiero agradecer a mis padres Blanca Alonso y Cuauhtémoc Hernández, así como a mi hermano David, a quienes, simplemente, debo todo.

A mis padres.

ÍNDICE

RESUMEN

Las proteínas de unión a RNA juegan papeles críticos en todos los procesos que involucran al RNA, desde su síntesis hasta su degradación, incluyendo algunos de gran relevancia como la expresión genética postranscripcional y la traducción. Su relevancia queda demostrada por su alto grado de conservación, su universalidad, así como su gran abundancia dentro de los distintos genomas. Específicamente, los dominios de unión a RNA de estas proteínas se encargan de la interacción con el RNA a través de una serie de mecanismos como la cooperación multidominio o la oligomerización.

Éste tipo dominios también están presentes en proteínas virales, las cuales llevan a cabo procesos fundamentales de su ciclo replicativo y en la manipulación del metabolismo celular de su hospedero. Por ello, su estudio es fundamental para comprender los procesos de interacción con la maquinaria celular y los procesos que marcan su evolución.

En éste trabajo se identificaron y analizaron funcionalmente los dominios de unión a RNA presentes en los proteomas virales. Se logró determinar su diversidad y la manera en que podrían estar participando en ciclo replicativo viral. Para ello se obtuvieron mas de cuatro mil proteomas virales del GenBank, y se emplearon todas las familias de dominios asociadas a la función de unión a RNA de la base de datos ProDom,. La búsqueda por perfiles se realizó a partir de cada una de las familias de dominios. Una vez obtenidos los resultados se creó un catálogo curado de dominios de unión a RNA presentes en proteínas virales.

ABSTRACT

RNA binding proteins play key roles in all RNA involving processes, from synthesis to degradation. Their universal distribution, abundance and high conservation levels demonstrates their relevance. RNA binding domains generate direct interactions with RNA through a variety of mechanisms as the multi-domain cooperation or the oligomerization process.

These proteins are also present in viruses, where participate in fundamental processes along the replicative viral cycle, for example in the hijacking of host metabolism. That is why the study of RNA binding proteins are valuable to the aim of understands how viruses evolve and interact with their hosts.

In this work, RNA binding domains were identified in all available viral proteomes. The viral domains were functionally analyzed allowing us to visualize their diversity and probable way in which they are working inside the viral replicative cycle. We used ProDom database to obtain family domains associated to the function "RNA binding", as well as viral proteomes from GenBank database. To find the homologous viral domains we employ a protein profile search for later create a catalogue.

The final results showed the particular nature of each viral Baltimore group and the closed relationship of these viruses with their hosts. Finally, we created the first viral RNA binding domain compendium that can be used as a revised source to continue more specific researches about the complexity of the interaction between virus and their hosts, the viral genome diversity or the implications of these proteins in the viral replicative cycle.

I. INTRODUCCIÓN

1. Las Proteínas de unión a RNA

Las proteínas de unión a RNA, conocidas como RBP por sus siglas en inglés (*RNA binding protein*), son todas aquellas proteínas que tienen la capacidad de interactuar con moléculas de RNA formando complejos ribonucleoprotéicos (RNP) (Glisovic et al., 2008). Las RBP además son parte integral del metabolismo del RNA, considerado como la parte central y más conservada de la fisiología celular (Anatharaman et al., 2002).

El RNA es la molécula central de la expresión genética y de la biocatálisis, como se observa en el ribosoma, la telomerasa y otras ribozimas. Aunque el RNA puede realizar una gran cantidad de actividades por sí mismo, en las células modernas se encuentra asociado con un gran número de proteínas (Cech, 2012). Los RNA presentan secuencias particulares en su composición nucleotídica, así como distintas estructuras secundarias y terciarias que permiten el reconocimiento e interacción con proteínas específicas (Jeong et al., 2003). Esta interacción RNA-proteína juega un papel crítico en la estabilidad, función, transporte y localización celular del RNA (Glisovic et al., 2008). Por ello, son esenciales en los procesos moleculares de todos los organismos y de los virus (Varadi et al., 2015). Algunos de estos procesos están relacionados con la regulación de la expresión genética postranscripcional (Gerstberger et al., 2014; Matia-Gonzales et al., 2015; Glisovic et al., 2008; Burd & Dreyfuss, 1994), como la traducción, el *splicing* alternativo, modificaciones y edición, poliadenilación, *capping* o el recambio proteínico. (Hennig et al., 2014; Ray et al., 2013; Glisovic et al., 2008). Por otro lado, las RBP no sólo participa en cada uno de estos procesos, sino que también proveen de una conexión entre ellos (Glisovic et al., 2008).

Se ha calculado que del 3 al 11% del genoma celular codifica para proteínas de interacción con el RNA, siendo las bacterias parasitarias con genomas reducidos las que presentan los mayores porcentajes (Anatharaman et al., 2002). Por ejemplo, se ha calculado que en levaduras, éste porcentaje va del 5 al 8%, mientras que para *C. elegans* y *D. melanogaster* sería de aproximadamente un 2% (Glisovic et al., 2008). Por su parte, Gerstberger et al. (2014), han calculado que en humanos estas proteínas representan alrededor de un 7.5% del total de genes codificantes. Sin embargo, es

posible que estos números estén subestimando debido a la existencia de muchas otras RBP no conocidas (Glisovic et al., 2008).

Algunas RBP pobremente descritas son las enzimas metabólicas, muchas de las cuales pueden unirse al RNA de manera específica, regulando la traducción o la estabilidad de los mRNA (Matia-Gonzales et al., 2015; Burd & Dreyfuss, 1994). El trabajo de Matia-Gonzales y colaboradores (2015) incluye un catálogo de las RBP que se unen a mRNA en *S. cerevisiae* y *C. elegans,* donde se encontró que el 73% de todas las proteínas identificadas, en su mayoría enzimas metabólicas, no habían sido relacionadas con la capacidad de unión a RNA. De forma puntual, describen el caso de las proteínas glicolíticas que pueden unirse a sus propios mRNA o al de otras proteínas similares, sugiriendo una regulación mediada por RNA que podría coordinar rutas metabólicas.

En éste mismo sentido, algunas proteínas asociadas a RNA no codificantes podrían estar asociadas con la regulación genética a través de la degradación del RNA, el silenciamiento transcripcional o la activación de loci, así como represión o activación de la transcripción; también se han reportado miRNA (microRNA) como moduladores de rutas metabólicas, y hay lncRNA (del inglés: *long non-coding RNA*) que han sido caracterizados funcionalmente en distintos procesos y enfermedades, incluyendo infecciones, inmunidad innata y adaptativa (Wang et al., 2017). Otras proteínas pueden actuar como componentes estructurales o catalíticos de RNP (Gerstberger et al., 2014). Así, se demuestra la gran complejidad de procesos en los que éstas interacciones RNA-proteína pueden estar implicadas.

En el caso de las RBP, una característica general presente en organismos de los tres dominios celulares es la presencia de regiones intrínsecamente desordenadas. La presencia de éste desorden intrínseco en las RBP ha sido propuesto como necesario para su función y regulación (Wang et al., 2016). Incluso, el desorden se conserva cuando la secuencia de la proteína pierde rastros de identidad, lo que indica su importancia funcional (Varadi et al., 2015). Estas regiones desordenadas establecen una interfase electrostática conservada y extendida con el RNA al que se une por un ajuste o plegamiento inducido. La flexibilidad conformacional que les caracteriza podría ser producto del enriquecimiento de aminoácidos con carga positiva que se emplean para establecer interacciones favorables con las cargas negativas de la cadena del RNA, mismos que promueven desorden al desestabilizar la cadena

proteínica (Varadi et al., 2015). Entre las principales ventajas que se le han atribuído a esta flexibilidad conformacional está el poder identificar distintos RNA, proveyéndoles así de multifuncionalidad, y permitir una evolución rápida con intermediarios funcionales (Varadi et al., 2015). Algunos ejemplos de RBP con altos grados de regiones intrínsecamente desordenadas son las histonas, factores de transcripción, proteínas ribosomales y factores de *splicing*, entre otras (Wang et al., 2016).

Por último, la mayoría de las RBP están conformadas por múltiples dominios de unión a RNA (RBD, por las siglas en inglés de *RNA binding protein*) (Burd & Dreyfuss, 1994), exhibiendo un alto nivel de modularidad (Glisovic et al., 2008). Ello permite generar nuevas interacciones no presentes en los dominios individuales (Burd et al., 1991). Esta característica, aunada a la mezcla de RBD con dominios auxiliares permite un mayor aumento en la diversidad funcional de estas proteínas. Otro modo de expandir éste repertorio es mediante la formación de variantes por *splicing* alternativo, así como por las modificaciones postranscripcionales (fosforilación, metilación, o ubiquitinación), que añaden un nivel más en la complejidad funcional y regulatoria de estas importantes proteínas (Glisovic et al., 2008).

2. Dominios de unión a RNA

Todas las proteínas están constituidas por unidades estructurales o funcionales llamados dominios. Éstos se suelen definir como unidades cuya secuencia está presente en distintas proteínas, las cuales pueden estar formadas por alrededor de 18 a 391 aminoácidos (Traut, 2014) (figura 1). Las proteínas simples pueden presentar uno o pocos dominios, mientras que las proteínas más complejas y grandes pueden tener hasta más de 30. Estos dominios proteínicos son considerados como unidades evolutivas (Kelley & Stenberg, 2015), de plegamiento, de la estructura y como unidades catalíticas o de unión (Traut, 2014).

En términos evolutivos, los dominios juegan un papel de gran importancia, ya que pueden ser incorporados a distintas proteínas, modificando su función estructural, catalítica o de unión (Traut, 2014; Björklund et al., 2005). Se ha propuesto que la evolución de las proteínas se da por fusión, recombinación y diferenciación de dominios existentes. En especial, la fusión de dominios permite la reutilización de

módulos funcionales en lugar de su reinvención (Björklund et al., 2005), lo que conduce a la formación de proteínas multidominio.

Las regiones funcionales específicas que permite la unión de los nucleótidos son los dominios de unión. La forma en la que el dominio reconoce de manera específica las llamadas cadenas nucleotídicas es por medio de los sitios de unión que están formados por secuencias cortas con patrones recurrentes en el DNA o RNA (D'haeseleer, 2006). Estos sitios de unión se encuentran igualmente en los organismos celulares y en los virus (Maris, *et al*. 2005).

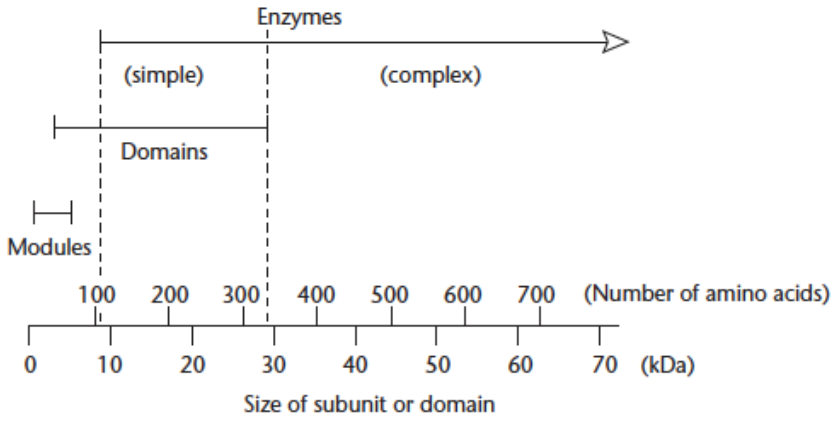


Figura 1. Tamaño de la estructura de proteínas, donde se muestra la variación en el tamaño de los dominios, el cual varía dependiendo de cómo es definido. Tomado de Traut (2014).

En general, RBD poseen una baja capacidad de interacción con moléculas de RNA, uniéndose a secuencias cortas (Holmqvist & Vogel, 2018) de entre dos a seis nucleótidos de cadena sencilla (Glisovic et al., 2008). Aunque también pueden llegar a reconocer algunas estructuras (Ray et al., 2013), un dominio aislado carece del potencial de unirse de manera específica al RNA (Ban et al., 2015). Es por esta razón que la gran mayoría de RBP presentan una arquitectura multidominio de RBD (Holmqvist & Vogel, 2018), lo que genera una superficie de interacción mayor (Ban et al., 2015) que les permite asociaciones específicas y de alta afinidad, además de reconocer secuencias más largas en el RNA (Ban et al., 2015; Gerstberger et al., 2014; Hennig et al., 2014; Glisovic et al., 2008; Burd & Dreyfuss, 1994).

Un caso de proteínas con multidominios RBD sería el de la proteína de unión a poly(A) (PABP), que posee cuatro dominios contiguos, los cuales funcionan

cooperativamente (Burd et al., 1991). Otro ejemplo son las proteínas PUF (Pumilio y FBF) de eucarionte (figura 2), que contienen típicamente ocho repeticiones *PUF RNA-Binding* consecutivas, cada una de aproximadamente 40 aminoácidos. En su conjunto forman una estructura cóncava donde el RNA se une, y en donde cada repetición reconoce un único nucleótido (Ban et al., 2015; Glisovic et al., 2008).

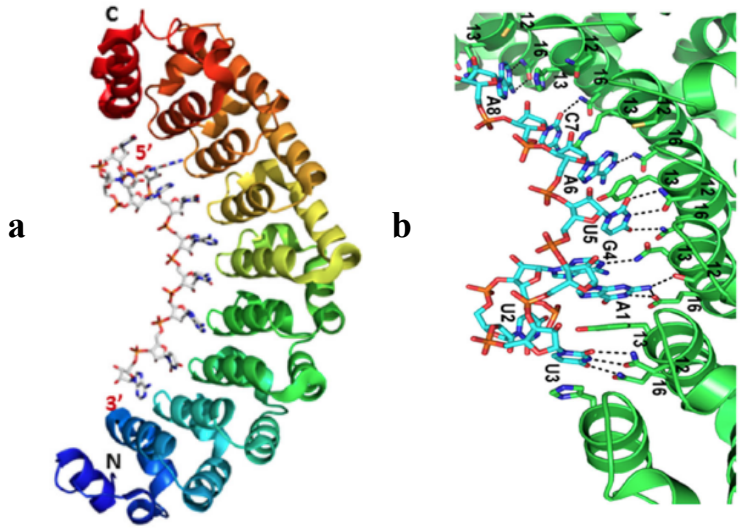


Figura 2. **a**. Vista lateral de la estructura del dominio de humanos Pumilio 1 (PUF) unido a una cadena de RNA. **b**. Acercamiento del reconocimientos de las bases del RNA, donde la proteína PUF aparece en color verde y el RNA como modelo de varillas. Modificado de Ban et al., 2015.

Diversidad de los RBD. Los RBD se hallan universalmente distribuidos, y debido a que aproximadamente la mitad de ellos están presentes en especies de los tres dominios celulares, se ha inferido su presencia en el último ancestro en común, LCA (por las siglas en inglés de *Last Common Ancestor*). La diversidad de RBD existentes puede agruparse funcionalmente en enzimáticos o de interacción (Anatharaman et al., 2002).

Considerando esta clasificación, los dominios enzimáticos serían aquellos que participan en la modificación de bases y azucares del RNA, así como la síntesis de nucleótidos no canónicos (tabla 1). Se incluyen además las enzimas dependientes de NTP como RNA helicasas, GTPasas y ATPasas P-loop (*phosphate-binding loop*), y algunas otras como RNA ligasas, nucleotidiltransferasas, RNA polimerasas dependientes de RNA o fosfodiesterasas cíclicas (Anatharaman et al., 2002).

Cabe destacar que las GTPasas P-loop han sido reclutadas muchas veces para funciones específicas en traducción, modificaciones y procesamientos del RNA.

Además, se hallan entre las RBP más antiguas, estando posiblemente nueve de ellas en el LCA formando parte del sistema de traducción (Anatharaman et al., 2002).

En cuanto a los dominios considerados como no catalíticos o de interacción, existen alrededor de 50 súper familias. Algunos de estos dominios son únicos y muy conservados en los genomas, como los que se encuentran en las proteínas ribosomales L30, S6, SRP14, mientras que otro pueden ser muy promiscuos y recurrentes, como el dominio RRM (*RNA recognition motif*) (Anatharaman et al., 2002). Algunas de las súper familias más destacadas, así como algunos de sus dominios y los procesos en los que están involucrados se muestran en la tabla 2.

Tabla 1. Ejemplos de modificaciones del RNA y de formación de bases no canónicas

Modificaciones del RNA	• Metilación • Desmetilación • Desaminación
Formación de bases no canónicas	• Tiouridilación • Pseudouridilación • Tioadenilación • Dihidrouridilación • Síntesis de arqueosina y queuina

Unos de los RBD mejor estudiados es el RRM, que está considerado como canónico debido a que es el más común en eucariontes. Está conformado por 80-90 aminoácidos que se pliegan en cuatro hojas beta anti paralelas y dos hélices alfa (figura 3). Éste motivo reconoce secuencias cortas con poca especificidad, por lo que es muy común encontrarlo en repeticiones o en combinación con otros dominios (Ban et al., 2015). Las uniones las establece principalmente con las hojas beta en residuos conservados como argininas o lisinas que forman puentes salinos con el esqueleto fosfodiestérico del RNA, siendo las regiones variables las que determinan la especificidad (Burd & Dreyfuss, 1994). Éste domino se presenta también con un gran número de variantes como la subclase quasi-RRM o el pseudo-RRM que generan interacciones con distintas partes de su arquitectura, como las hélices alfa o los *loops* de las hojas beta, demostrando flexibilidad y variación en un motivo relativamente simple (Ban et al., 2015). Un característica adicional de éste domino es la presencia de regiones desordenadas en los conectores terminales que los unen a otros dominios, los cuales sufren de un plegamiento inducido en presencia del RNA al que se une;

éste rasgo les permite generar interacciones más fuertes al aumentar la superficie de interacción (Varadi et al., 2015).

Tabla 2. Dominios de unión a RNA con función de interacción agrupados por clases estructurales, así como algunos ejemplos de proteínas o procesos de los que forman parte (Anatharaman et al., 2002).

Estructura de familias de dominios no catalíticos	Dominios	Proteínas o procesos en que están involucrados
OB-fold	• S1 • N-OB • EMAP	• Sistema de traducción • RNasa E • RNasa II • GTPasa • eIF2 *α* • Prp22 • Shock protein
all-β	• KOW • L2 • SM • TUDOR • NusG • PUA • TRAM	• Procesos de *splicing* • Estructurales del ribosoma • Modificación del RNA
α+β y α/β	• RRM • KH • S4 • dsRBD • THUMP • TGS	• S6(ribosomal) • RNA y DNA polimerasas • PSUS tipo II • Sistema inmune • Proceso de *splicing*
all-α	• HhH • PIN • Translin	• S13 y S18 (ribosomales) • PTGR • Degradación de RNA • RNasa II • Localización del RNA
Metal-chelating	• Zn-ribbon • Zn-knuckle • CCCH • C2H2 Zn finger • Little finger • LRP1 finger	• S14 ribosomal • IleRS • S27 • MetRS • eIF5 • L36AE

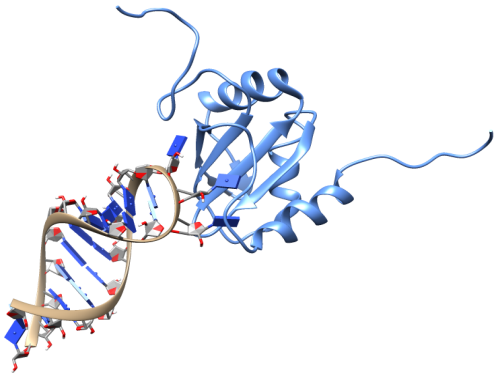


Figura 3. Dominio RRM unido a RNA (modificado de 2N3O, RCSB PDB).

Otros RBD canónicos son *1)* el motivo rico en arginina (ARM), conformado por 10-20 aminoácidos; *2)* RGG box (20-25 aminoácidos), que genera uniones no específicas y que se encuentra de 6 a 18 repeticiones; *3)* el Homólogo K (KH), que está ampliamente distribuído y forma parte de las proteínas S3 ribosomales; *4)* dsRBD (del inglés: *double-stranded RNA binding domain*), que reconoce RNA de doble cadena y está involucrado en diversas funciones celulares, como las postranscripcionales y respuesta inmune (Glisovic et al., 2008; Burd & Dreyfuss, 1994); *5)* los Dedos de Zinc, *6)* Cold-shock; y *7)* el DEAD-DEAH box, entre otros (Gerstberger et al., 2014; Glisovic et al., 2008).

Evolución de los RBD. El origen de las RBP y sus dominios de unión debe remontarse a etapas muy tempranas de la evolución, debido a su amplia distribución entre los organismos celulares. Por esta razón se ha sugerido la presencia de algunas de estas proteínas y RBD en el LCA, donde el sistema de traducción agruparía el mayor número de proteínas y dominios conservados (Fox, 2010; Anatharaman et al., 2002), seguido de algunas enzimas de modificación (metiltransferasas, pseudouridin sinterasas, enzimas de tiouridina y tioadenina) (Anatharaman et al., 2002).

Se han descrito dos posibles fases de radiación y movilización en los RBD. La primera fase pudo ocurrir tempranamente, una vez conformados los dominios más antiguos (como podrían ser EMAP, PUA, PIN, TRAM, THUMP, S4, KOW, N-OB, NusB, entre otros), mismos que rápidamente se fusionaron entre ellos, llevando a una radiación en la diversidad funcional. Posteriormente, ocurrió una segunda gran fase de movilidad con el surgimiento de los eucariontes. Éste mismo aumento de movilidad coincide con el origen del *splicing,* la regulación genética postranscripcional, la edición del mRNA, entre otros mecanismos (Anatharaman et al., 2002).

En los eucariontes ocurrió una diversificación de los RBD debido a la aparición y el reclutamientos de estructuras o motivos antiguos, lo que dio lugar a RBD como el RRM, PWI, SWAP, CCCH, *Little finger* o Zn *Knuckle*. Parte de la explicación de éste aumento de movilidad y de diversidad de RBD está relacionada con el origen quimérico de los eucariontes, los cuales surgen de la combinación entre los linajes Archaea y Bacteria. Por esta misma razón se pueden observar grupos de RBD con distribución exclusiva en los linajes Archaea-Eukarya (como algunos involucrados en maduración del RNA o en traducción) o Bacteria-Eukarya (por ejemplo RNasa 3'-5' o algunos dominios que actúan en mitocondrias o cloroplastos) (Anatharaman et al., 2002).

Además, es bien sabido que en los eucariontes ocurrió un aumento en la complejidad del sistema de traducción y de regulación de la expresión genética, por lo que se requirió de una expansión en el número y función de RBP (Glisovic et al., 2008; Anatharaman et al., 2002). Éste mismo incremento de la complejidad se relaciona con el incremento de intrones, lo que permitió la utilización de estas proteínas en una gran cantidad de arreglos y combinaciones (Glisovic et al., 2008). Una prueba de esto es el hecho de que, sobre todo en eucariontes, las RBP con mayor cantidad de dominios son las relacionadas a mRNA, reflejando una rápida expansión de procesos relacionados a esta molécula en particular, como son el *splicing* o las modificaciones postranscripcionales (Gerstberger et al., 2014).

Interacciones RNA-proteínas. La manera en que las proteínas y sus dominios interactúan con el RNA está mediada por diversos mecanismos que, en su conjunto, muestran una gran complejidad. Estas interacciones pueden describirse y entenderse desde un nivel atómico hasta uno macromolecular en donde interviene la cooperación entre distintas proteínas. De esta manera es posible generar las interacciones específicas, así como regular el funcionamiento de las RBP y los procesos en que están involucradas.

A nivel atómico, los RBD interaccionan con los nucleótidos del RNA mediante distintas fuerzas intermoleculares, como los puentes de hidrógeno o las fuerzas de van der Waals (Jeong et al., 2003). Los contactos se llevan a cabo entre la base, el azúcar o el fosfato del nucleótido, y la cadena lateral o la cadena peptídica del aminoácido (Hoffman et al., 2004). De entre estos contactos se ha observado que el más

recurrente es el que se da con los fosfatos de la cadena de RNA (Hoffman et al., 2004; Treger & Westhof, 2001).

Pese a ser muy comunes, tanto las fuerzas de van der Waals como los puentes de hidrogeno mediados por agua no intervienen en uniones específicas, pero contribuyen a la estabilidad y fuerza de la interacción. Por otra parte, los puentes de hidrogeno (figura 4) son los causantes de la especificidad en las interacciones, principalmente cuando el contacto ocurre con la base de un nucleótido, ya que los puentes formados con el esqueleto solo contribuyen a la estabilidad (Luscombe et al., 2001).

Existen tres tipos de interacción por puente de hidrogeno definidos por el número de puentes formados: *1)* simples, donde se genera un solo puente de hidrogeno entre un aminoácido y una base; *2)* bidentadas, que generan dos o más puentes de hidrogeno entre un aminoácido y una base o bases pareadas (se considera que hay aumento en especificidad) (figura 5a); y *3)* complejas, en donde un solo aminoácido puede interactuar con más de una base a la vez, permitiendo el reconocimiento de secuencias cortas y, por tanto, aumentando el grado de especificidad (figura 5b) (Luscombe et al., 2001).

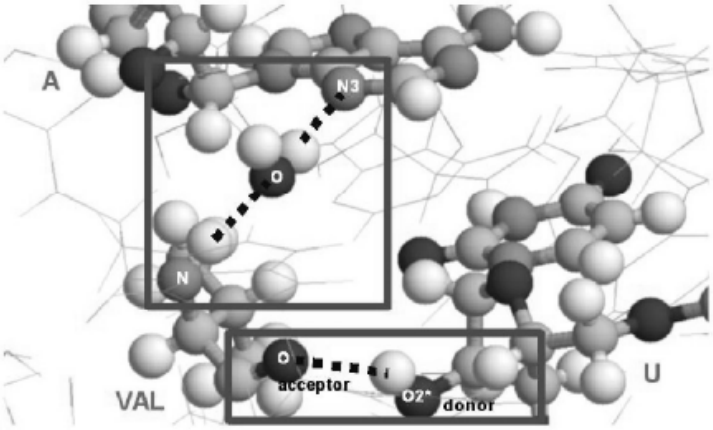


Figura 4. Diagrama tridimensional de los átomos que forman puentes de hidrógeno entre aminoácidos y bases nucleotídicas. Las líneas punteadas indican el puente de hidrógeno. En el rectángulo superior se muestra un puente de hidrogeno mediado por agua entre nitrógenos de una valina y una adenina. En el rectángulo inferior se muestra un puente de hidrógeno directo entre oxígenos de una valina y un uracilo (tomado de Jeong et al., 2003).

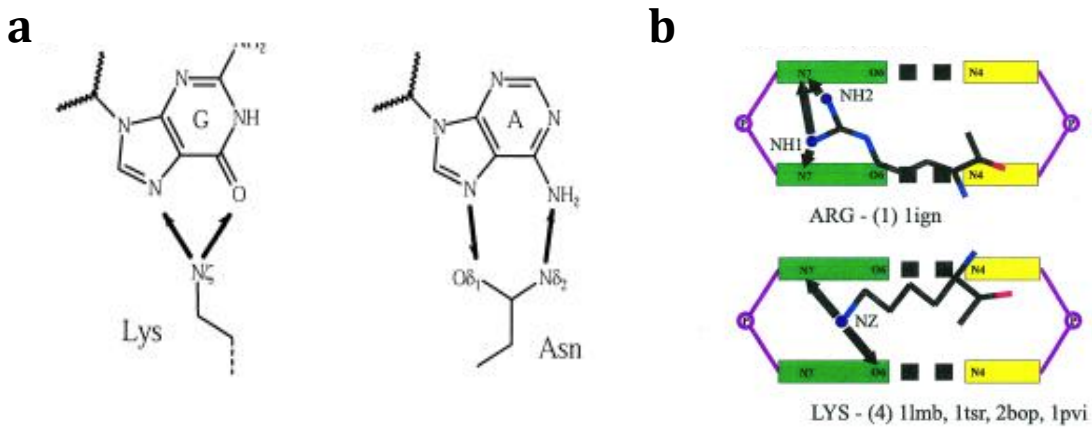


Figura 5. Tipos de interacciones entre aminoácidos y nucleótidos. **a.** Interacciones bidentadas entre lisina – guanina y ácido aspártico – adenina, donde las fechas indican los átomos que interactúan, con dirección de aceptor hacia donador. **b**. Interacciones complejas en bases pareadas y apiladas, donde los aminoácidos se presentan como líneas negras y las bases como rectángulos de colores. Las flechas indican los átomos que generan los puentes de hidrógeno. Modificado de Luscombe et al., 2001.

Generalmente, debido a que los aminoácidos con cargas positivas tienden a interactuar con las cargas negativas de los fosfatos de la cadena de RNA, estos median el mayor número de contactos en los RBD (Hoffman et al., 2004) (figura 6). Algunos de estos aminoácidos son por ejemplo, arginina (R), serina (S), lisina (K) o treonina (T) (Luscombe et al., 2001). La arginina es el aminoácido más común en éste tipo de interacciones (Blanco et al., 2018; Varadi et al., 2015; Hoffman et al., 2004; Jeong et al., 2003; Luscombe et al., 2001; Burd & Dreyfuss, 1994), el cual, por estar positivamente cargado y tener una cadena lateral larga y flexible con capacidad de formar varios puentes de hidrógeno (Luscombe et al., 2001), presenta una alta afinidad con el RNA. La lisina es el aminoácido que sigue a la arginina en frecuencia (Blanco et al., 2018; Varadi et al., 2015; Hoffman et al., 2004; Jeong et al., 2003; Luscombe et al., 2001; Burd & Dreyfuss, 1994), el cual presenta características similares. Como ejemplo, Varadi y colaboradores (2015), analizaron las estructuras disponibles de RBP en *Protein Data Bank* (PDB) para observar la frecuencia de amino ácidos, encontrando que en relación a la misma base de datos, había un incremento de 40% en argininas, y 33% en lisinas. El resultado más evidente al analizar las frecuencias fue en las zonas de interacción RNA-proteínas, donde el porcentaje aumenta a 180% en argininas y 116% en lisinas.

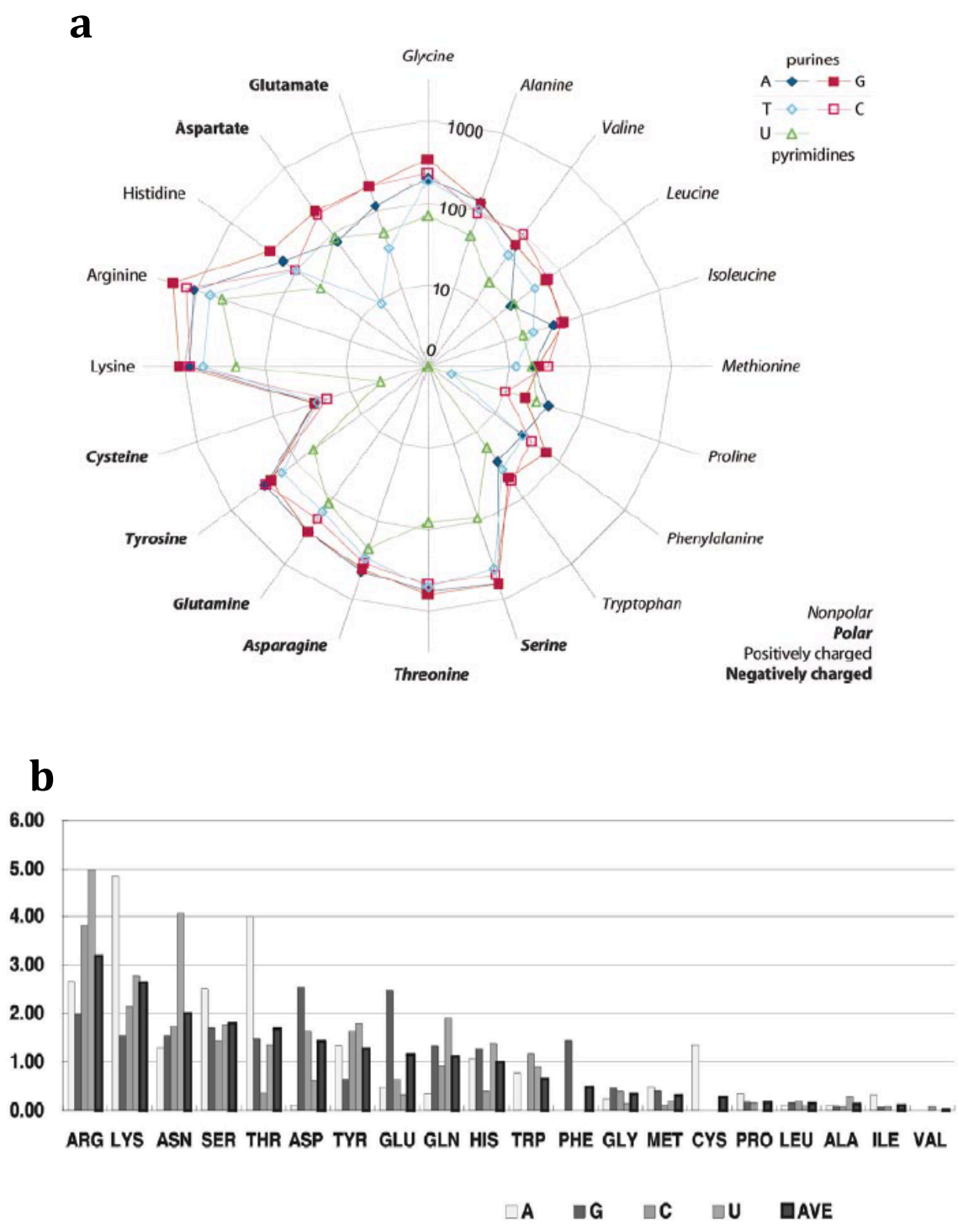


Figura 6. Diagramas de frecuencia de interacciones entre aminoácidos y nucleótidos mediados por puentes de hidrógeno, generados por análisis de estructuras terciarias. El diagrama superior (a) muestra el número de interacciones (tomado de Hoffman et al., 2004). El diagrama inferior (**b**) muestra los porcentajes de propensiones de interacción (tomado de Jeong et al., 2003).

Las interacciones RNA-proteínas no específicas son esenciales en la regulación del metabolismo del RNA por medio de interacciones en donde hay poca relación entre los RBD y las secuencias del RNA (Ban et al., 2015). Un ejemplo son los receptores citoplasmáticos de RNA viral de vertebrados, que interactúan con los fosfatos y azúcares de la parte terminal de las dobles cadenas de RNA viral. Otros ejemplos son la proteína argonaute-2 (AGO2), involucrada en silenciamiento genético, y la proteína IFIT5 (del inglés *Interferon-induced protein with tetratricopeptide repeats 5*), que actúa como sensor de RNA viral y mediador del respuesta inmune innata. Los dominios de unión involucrados en estas interacciones reconocen su blanco al unirse a distintos grupos de marcadores en las partes terminales de la cadena de RNA, como por ejemplo el fosfato 5' terminal, o un trifosfato 5' libre que se puede encontrar en el RNA viral. Otro mecanismo empleado es la formación de canales o surcos positivamente cargados donde se puede acomodar un RNA (Ban et al., 2015).

Para poder generar interacciones específicas RNA-proteínas se requiere de distintos mecanismos como el acomodo de RBD en tándem, que permite reconocer secuencias más largas, y que da variación estructural a la interacción. Las proteínas con repeticiones de pentatricopéptidos (PPR), que contienen hasta 10 repeticiones del dominio de unión, son un ejemplo de cómo los dominios en tándem actúan cooperativamente para la unión de secuencias largas en el RNA (Ban et al., 2015).

La dimerización y la oligomerización son también mecanismos recurrentes en las RBP para aumentar la capacidad de interaccionar de forma específica con el RNA. Éste proceso es inducido por la presencia del RNA, y en ambos casos los RBD de cada proteína actúan cooperativamente fortaleciendo la afinidad de la interacción. Una vez unidas, se forma un surco o ranura positivamente cargado en la zona de interfase donde se sitúan los RBD de ambas proteínas, es decir, el mismo sitio donde el RNA es embebido. De esta forma se asegura la interacción con la molécula de RNA tanto por la estructura como por la secuencia que es reconocida. Una proteína que ejemplifica lo anteriormente mencionado es THA8 de maíz que forma dímeros en presencia de RNA, o las proteínas SF1 y U2AF del complejo del spliceosoma (Ban et al., 2015).

Pese a lo mencionado arriba, aún no hay un entendimiento completo de cómo las RBP logran la unión especifica al RNA con un grupo limitado de RBD (Hennig et al., 2014), y menos aún se conoce acerca de la red de interacciones entre estas proteínas.

(Gerstberger et al., 2014). Incluso algunos de los miembros de los RBD más comunes, no han sido estudiados en detalle, incluyendo al dominio RRM, algunas helicasas, nucleasas y factores de transcripción, entre otros (Gerstberger et al., 2014).

3. Evolución de los virus

Los virus han sido definidos como elementos genéticos independientes que requieren de una célula hospedera para multiplicarse (Domingo y Perales, 2014). Estas entidades biológicas juegan un papel fundamental en la red de interacciones de la vida, siendo parásitos obligados de todos los organismos de los tres dominios celulares (Mihara et al., 2016; Forterre, 2006), y jugando un papel en la regulación de la biogeoquímica del planeta, en la que controlan ciclos fundamentales como el del carbón, nitrógeno y fósforo a través de la regulación y estructuración de las comunidades de microorganismos (O'Malley, 2016). Uno de los aspectos importantes que los define es la presencia de un par de características típicas de los organismos vivientes: replicación y evolución darwiniana (Domingo & Perales, 2014).

Los virus están conformados principalmente por un genoma de RNA o DNA, una cápside de origen proteínico y, en algunos casos, de una envoltura lipoprotéica. La información contenida en su material genético puede dividirse en dos grupos (Krupovič & Bamford, 2008): *1)* propios de los virus o *hallmark genes*, que se consideran como genes exclusivos o casi exclusivos de los virus, y que codifican para componentes estructurales del virión (forma libre e infecciosa del virus), así como para proteínas de la replicación (Domingo & Perales, 2014; Krupovič & Bamford, 2008). Y *2*) no propios del virus, los cuales son aquellos genes involucrados en la interacción con el hospedero, y que pueden ser intercambiados entre virus no relacionados o incluso entre el virus y el hospedero (Krupovič & Bamford, 2008).

Todas las características que conforman a un virus, desde su morfología, contenido genético o ciclo replicativo, dependen de su hospedero y el ambiente, debido a una coevolución constante entre estos (Hunter, 2017; Domingo & Perales, 2014). Por tanto, los genes y estructuras conservadas en virus, en general, se asocian a funciones como el proceso de ensamblaje, empaquetamiento del material genético o replicación; mientras que aquellas relacionadas con interacciones específicas con el hospedero están menos conservadas entre los distintos virus (Bamford, 2002).

Además de la alta tasa de mutación que presentan los virus en relación a los organismos celulares, misma que les permite generar variación y adaptación constante, en estas entidades la transferencia horizontal de genes juega un papel fundamental en su evolución. Éste fenómeno es recurrente entre los virus, así como entre virus y hospederos (Hunter, 2017; Filée, 2014; Rohwer & Thurber, 2009). Es sabido que los organismos celulares, sobre todo los eucariontes, presentan secuencias de origen viral en sus genomas (Bamford, 2002), como en humanos y ratones, donde retrovirus endógenos ocupan del 8-10% de su genoma total (Dupressoir et al., 2012). Pese a esa gran cantidad de transferencias virus-hospedero, algunos autores afirman que la mayoría de estos intercambios genéticos están direccionados hacia los virus (Hunter, 2017).

Existen varios registros de transferencia horizontal de genes de organismos celulares hacia virus, principalmente de sus hospederos. Algunos ejemplos incluyen: cianófagos marinos, que comúnmente presentan genes fotosintéticos en sus genomas; el del virus EhV, que contiene enzimas de su hospedero, el alga *Emiliania huxleyi* (Monier et al., 2009); el de los viromas marinos, en donde se encuentra una gran variedad de genes metabólicos (Rohwer & Thurber, 2009); así como varias chaperonas, nucleasas, helicasas, y oncogenes en retrovirus (Flint et al., 2009). Otro ejemplo destacado, y uno de los más estudiados, es el de los virus gigantes (NCLDV por las siglas en inglés de *nucleocytoplasmic large DNA viruses*), los cuales presentan una gran cantidad de genes y fragmentos genómicos de origen celular, algunos de sus hospederos (Jeudy et al., 2012), pero también de bacterias (Filée et al., 2006) y de diversos eucariontes (Schulz et al., 2017; Filée, 2014). El mantenimiento de los genes celulares en los genomas virales posiblemente sea resultado de una presión de selección positiva que les permitiría una mejor adaptación al ambiente celular de su hospedero.

El origen de los virus permanece como un objeto de debate y especulación (Krupovič & Koonin, 2017), representando uno de los grandes retos para la biología contemporánea. Existen tres hipótesis principales que proponen el origen de los virus: *1)* vestigios del mundo prebiótico, *2)* elementos genéticos que escaparon de las células, y *3)* parásitos con evolución retrograda. La primera de estas ideas establece que al menos los virus de RNA aparecen antes del surgimiento de las primeras células, siendo relictos del periodo conocido como Mundo del RNA (figura 7) (Domingo & Perales, 2014; Fisher, 2010; Holmes, 2009; Forterre, 2006). Esta propuesta es

apoyada principalmente por la presencia de los *hallmark genes* y la gran diversidad de estrategias replicativas que presentan los virus (*cf.* Campillo-Balderas et al., 2015).

La hipótesis de los genes escapados establece que los virus evolucionaron de manera independiente en diferentes dominios celulares a partir de genes auto replicativos que se volvieron infecciosos (Krupovič & Koonin, 2017; Fisher, 2010; Holmes, 2009) (figura 7).

Por último, en la hipótesis de evolución retrograda (figura 7) se supone la pérdida de funciones en un organismo parásito intracelular mediante un proceso evolutivo, ya que estas mismas funciones metabólicas de síntesis de proteínas o producción de energía estarían presentes en el hospedero del cual dependen (Domingo & Perales, 2014; Fisher, 2010; Holmes, 2009; Forterre, 2006).

Pese a los contrastes entre las distintas hipótesis, algunos sugieren que éstas no son mutuamente excluyentes, y que en parte pudieran explicar distintas etapas o procesos de la aparición y evolución de los virus, los cuales pudieron evolucionar en ocasiones múltiples e independientes (Krupovič & Koonin, 2017).

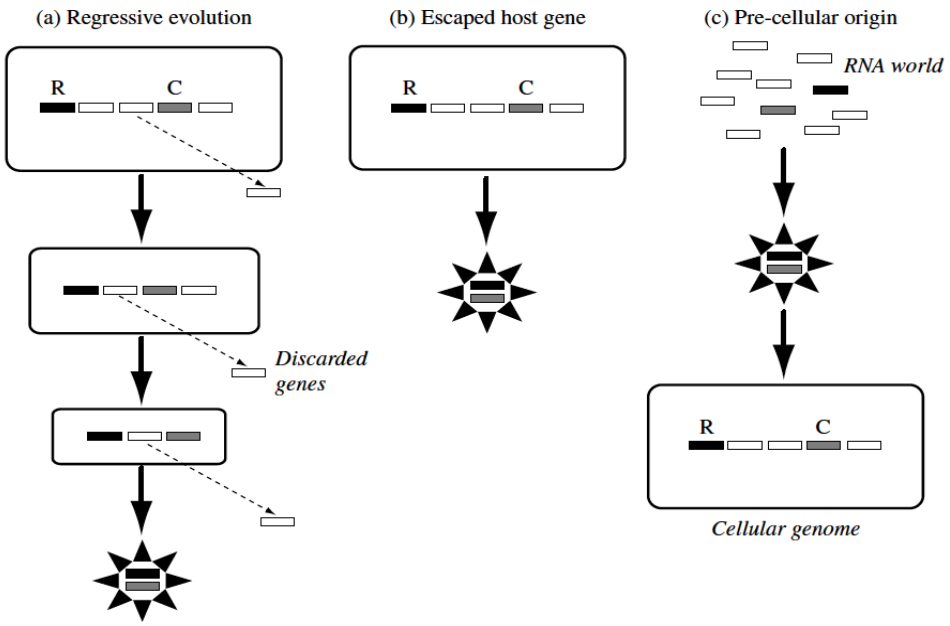


Figura 7. Representación de las principales teorías del origen de los virus. Tomado de Holmes (2009).

Existe una estrecha relación de los virus con el desarrollo y evolución de los organismos celulares. De hecho, se considera que los virus jugaron papeles fundamentales en la evolución de la vida como la conocemos hoy. Algunos de estos eventos clave son la posible invención del DNA y sus mecanismos de replicación, el origen del núcleo eucarionte (Forterre, 2006; Domingo & Perales, 2014), y otras aportaciones como la inducción a duplicaciones del complejo mayor de histocompatibilidad (MHCI), regulación de producción de la amilasa salival, resistencia a retrovirus Fv-1 (Flint et al., 2009), o el origen de la placenta de mamíferos (Dupressoir et al., 2012; Flint et al., 2009). Como ya se ha mencionado, debido a la gran cantidad de material genético que transfieren entre organismos celulares, los virus son considerados como modeladores de genomas, por lo que son de suma importancia como fuente de variación y diversificación genética y, en consecuencia, de los procesos evolutivos de los seres vivos.

4. Las RBP en virus

Las RBP son esenciales a lo largo del ciclo replicativo de los virus. Por un lado, debido a que todos los virus deben sintetizar mRNA para ser traducido por los ribosomas celulares (Baltimore, 1971), requiriendo de proteínas que interaccionen para regular, transcribir, o expresar la información genética. Por otro lado, los virus requieren de un buen número de RBP para interactuar y manipular el metabolismo y expresión genética de sus hospederos (Flint et al., 2009). Además, algunos virus, especialmente los de RNA, han evolucionado para emplear las RBP de sus hospederos (Viktorovskaya et al., 2016; Li & Nagy, 2011; Steitz et al., 2010; Boyle & Holmes, 1986). Un ejemplo bien estudiado son virus de ssRNA (+), que utilizan un gran número de RBP celulares a los largo de su ciclo replicativo, como transportadores núcleo-citoplasma, factores de traducción, chaperonas de RNA, activadores de traducción independientes de *Cap*, estabilizadores del RNA, reguladores de replicación, entre otros (*cf.* Lloyd, 2015; Li & Nagy, 2011; Chavali et al., 2007).

Las RBP virales están asociadas a procesos como la replicación del genoma, transcripción, traducción, transporte y procesamiento del RNA, así como empaquetamiento y encapsidación del material genético (Boyle & Holmes, 1986), pero igualmente intervienen en distintos procesos celulares como el bloqueo del

sistema inmune innato, manipulación del metabolismo o de la respuesta a estrés (Viktorovskaya et al., 2016; Li & Nagy, 2011; Flint et al., 2009).

Al igual que en el caso de las RBP celulares, sus contrapartes virales se caracterizan por un alto grado de desorden, que se ha especulado, les permite enfrentar presiones de selección y generar nuevas interacciones, así como el compactar sus genomas (Varadi et al., 2015).

Se ha sugerido que estas proteínas virales tienen un origen celular, producto de la transferencia horizontal, la lista incluye componentes del aparato de traducción en virus gigantes (Jeudy et al., 2012), o incluso algunas proteínas y dominios propuestos como exclusivos de virus, tales como las proteínas de la nucleocápside, proteasas, o los dominios *jelly-roll* de las cápsides (Krupovič & Koonin, 2017).

Se puede pues considerar que las RBP y sus respectivos RBD juegan un papel fundamental en los procesos virales y celulares, por ello, están sujetos a la selección natural.

Los virus requieren del empleo de estos elementos proteínicos para su éxito replicativo y para una adecuada adaptación al ambiente celular de sus hospederos. Por ello, para el estudio de los virus es fundamental el considerar la información de los hospederos celulares, para así revelar procesos de coevolución y detectar interacciones genéticas por transferencia horizontal de genes (Mihara et al., 2016).

En el presente trabajo se busca generar el primer catálogo de RBD presentes en proteínas virales con el objeto de proveer de una fuente curada que permita avanzar en el entendimiento de la complejidad en las interacciones virus-hospedero, la diversidad genómica viral, así como para guiar futuras investigaciones acerca de las RBP y RBD en el ciclo replicativo de los virus.

II. OBJETIVOS

1. Objetivo general

Catalogar los dominios de unión a RNA encontrados en los proteomas virales disponibles.

2. Objetivos particulares

- Identificar y catalogar los distintos RBD existentes en los proteomas virales disponibles en las bases de datos mediante el uso de herramientas bioinformáticas.
- Realizar un análisis funcional del catálogo obtenido.

III. MÉTODOS

1. Bases de datos

Familias de dominios proteínicos de unión a RNA. Se utilizó la base de datos ProDom (http://prodom.prabi.fr/prodom/current/html/home.php), que incluye familias de dominios de proteínas generadas por un proceso automático de comparación global de todas las proteínas disponibles en SWISS-PROT-41.23 y TrEMBL24.11. Está basada en familias de dominios bien conocidos y caracterizados para poder reclutar dominios homólogos a partir de la similitud de la secuencia, empleando algoritmos de PSI-BLAST (Bru et a., 2005).

Para generar la base de datos para éste proyecto, se llevó a cabo una búsqueda en ProDom seleccionando las familias que tuvieran asignada la función molecular "*RNA binding*" [GO: 0003723], obteniendo un total de 6,469 familias. Las secuencias que conforman estas familias fueron obtenidas manualmente al igual que la información sobre su origen celular o viral, el número de secuencias que forma la familia, la longitud máxima de las mismas, y proteína más frecuente, si contaban con esta anotación.

Proteomas. Un total de 4,035 proteomas virales fueron obtenidos de la base de datos GenBank, los cuales se utilizaron como base de datos para la búsqueda de los homólogos de las familias proteínicas de unión a RNA. Esta base de datos tiene un total de 201,601 secuencias, y 50,407,260 residuos.

2. Análisis bioinformáticos

Alineamientos múltiples. Se llevaron a cabo alineamientos múltiples de las secuencias que conforman a cada una de las familias de dominios obtenidas de ProDom. Para ello se empleó MAFTT v.7.205, que utiliza un método basado en similitud para realizar de manera precisa y veloz alineamientos múltiples de un gran número de secuencias (Katoh & Standley, 2013). Se empleó un alineamiento local con mil iteraciones, usando la matriz BLOSUM62, que está establecida por *default* para secuencias de aminoácidos.

```
$ mafft --localpair --maxiterate 1000 --reorder fasta_file
```

Formación de perfiles. Para realizar una búsqueda de homólogos distantes con una mayor sensibilidad y que permita la detección de dominios poco conservados se crearon perfiles de cada una de las familias de dominios proteínicos a partir de los alineamientos múltiples anteriormente realizados, empleando la herramienta *hmmbuild* de HMMER 3.1b1 con los parámetros de *default* recomendados. Estos perfiles son modelos estadísticos de alineamientos múltiples, los cuales capturan la información de las posiciones específicas sobre qué tan conservada es cada columna del alineamiento **(**Eddy, 2015), proveyendo de un peso numérico para cada *match* o *mismatch* posible entre el residuo de una secuencia y la posición de un perfil (Sigrist et al., 2002).

Búsqueda de perfiles en proteomas virales. Para llevar a cabo la búsqueda de los 6,469 perfiles en los proteomas virales, se empleó la herramienta *hmmsearch* de HMMER 3.1b1 utilizando los parámetros de *default*, y únicamente cambiando el valor de expectación (*E-value)* de <=10 a <= 0.1*,* a partir del cual se reportan las secuencias encontradas.

```
$ hmmsearch –E 0.1 profile.hmm proteomes_database
```

Para considerar un *match* como significativo se empleó el valor de expectación. Éste es un valor estadístico que refleja la distancia evolutiva entre dos secuencias alineadas, y está considerado de mayor utilidad para inferir homología que el porcentaje de similitud comúnmente empleado (Pearson, 2013). El valor de expectación reporta el número de veces que un *score* ocurriría por azar, por lo que entre más cerca esté el valor de cero éste será mejor. El valor depende del tamaño de

la base de datos, por lo que se considera que sería más sencillo detectar homólogos distantes en una base de datos pequeña (<100,000 – 500,000) (Pearson, 2013).

El valor de *bit-score* obtenido también está incluido en los resultados. Éste valor representa el tamaño requerido de una base de datos en la cual un *match* en particular podría ser encontrado por suerte, por lo que entre más grande sea el valor, éste será más significativo (Pearson, 2013).

Debido a que se usaron búsquedas con perfiles proteína:proteína (las cuales presentan valores estadísticos más precisos y sensibles que las de DNA:DNA), que aumentan el grado de resolución en la búsqueda de homólogos distantes, se puede considerar que un *E-value* < 0.001 es significativo, mientras que para un *E-value* de 0. 001- 0.01 se puede sospechar de una relación evolutiva. En el caso de los valores de *bit-score* alrededor o mayor de 50 puede considerarse significativo (Pearson, 2013).

3. Curación de datos y catálogo de dominios de unión a RNA virales

Base de datos. Se agregó información descriptiva a las familias de dominios que al menos obtuvieron un *match* dentro de los parámetros establecidos en los proteomas virales. Para éste fin se utilizó la función "*Retrieve/ID mapping*" disponible en UniProt (https://www.uniprot.org/), donde se introdujeron los identificadores de secuencia que conforman a cada familia de dominios para recuperar anotaciones de las principales categorías de *The Gene Ontology Consortium* (GO): *1) Biological process*, *2) Molecular function,* y *3) Cellular component*. Además, se obtuvieron los grupos taxonómicos superiores, Phylum y Clase, a los que pertenecen las secuencias, así como el identificador de estructuras terciaria de la base de datos PDB si éste existía. Las familias de dominios que no presentaron anotación alguna de GO y las que presentaron anotaciones muy vagas o específicas de unión a DNA fueron excluidas.

Después se eliminaron las redundancias en las anotaciones de GO, para así obtener un solo representante de cada anotación asociada a las familias. Lo mismo se hizo en el caso de las columnas que contienen los datos taxonómicos.

Por tanto, la base de datos final se conformó por el identificador de ProDom para cada familia, el número de secuencias que conforman la familia, la longitud máxima de las secuencias, la proteína más frecuente, anotaciones asociadas a GO, el dominio

celular, Phyla y Clase al que pertenecen las secuencias y, por último, el identificador de estructuras terciarias de PDB. En el caso de secuencias virales, se reportó el tipo de genoma, así como la clasificación taxonómica a nivel de Familia u Orden viral dependiendo de la información disponible.

Catálogo de dominios de unión a RNA virales. El catálogo de dominios de unión a RNA está conformado por todos los *match* encontrados en los proteomas virales. Se reportaron los valores estadísticos de *E-value* y *bit-score* obtenidos de los archivos de salida de *hmmsearch*, así como los identificadores de cada secuencia viral encontrada, su descripción, el nombre de la proteína y el nombre del virus al que pertenecen. El nombre de los virus se utilizó para realizar una búsqueda en la base de datos Virus-Host DB (https://www.genome.jp/virushostdb/) (Mihara et al., 2016) con la intención de obtener información sobre el grupo viral de acuerdo a la clasificación de Baltimore (Baltimore, 1971), al que pertenece cada virus, su clasificación taxonómica a nivel de Familia u Orden en el caso de que la familia no esté determinada, así como el desglose taxonómico de los hospederos que infecta. Por último, se asociaron las secuencias encontradas con alguna función de *Clusters of Orthologous Groups* (COG) (tabla 3) a partir de la información disponible en UniProt, en donde se consideraron descripciones, anotaciones GO, dominios proteínicos y estructuras asociadas.

Las secuencias encontradas en los proteomas virales tras la búsqueda con los perfiles fueron extraídas para emplearlas en otros análisis, siempre y cuando éstas tuvieran valores estadísticos significativos. De éste conjunto de secuencias se obtuvo la longitud máxima, la cual también fue agregada en la información del catálogo.

A manera de resumen, el catálogo contiene en primer lugar el identificador de ProDom para cada familia, para relacionar la base de datos final y los resultados obtenidos agrupando los resultados encontrados en relación con cada familia de dominios. Además, el catálogo está conformado por la longitud máxima del grupo de secuencias encontradas, los valores estadísticos *E-value*, *bitscore* y las desviaciones de los *bitscore* (*bias*) para cada *match*, el identificador de cada secuencia, su descripción, nombre de la proteína, nombre del virus al que pertenece, función COG asignada, e información asignada de UniProt, principalmente la relacionada a dominios o función. Por último, también contiene la información de la clasificación

de cada virus tanto a nivel de la clasificación de Baltimore como a nivel taxonómico de Familia, y la información taxonómica de sus hospederos celulares.

Tabla 3 . Clasificaciones funcionales de *Clusters of Orthologous Groups* (COG) agrupados por categorías generales. * Categorías adicionadas para éste trabajo con el fin de clasificar secuencias con funciones virales o funciones difíciles de definir.

Information Storage and Processing	Cellular Processes and Signaling	Metabolism	Poorly Characterized	*Viral Processes
[A] RNA processing and modification	[D] Cell cycle control, cell division, chromosome partitioning	[C] Energy production and conversion	[R] General function prediction only	*[Vc] Capside/membrane structure
[B] Chromatin structure and dynamics	[M] Cell Wall/membrane/envelope biogenesis	[E] Amino acid transport and metabolism	[S] Function unknown	*[Vp] Viral pathogenesis
[J] Translation, ribosomal structure and biogenesis	[N] Cell motility	[F] Nucleotide transport and metabolism		*[VY] Viral polyprotein
[K] Transcription	[O] Posttranscriptional modification, protein turnover, chaperones	[G] Carbohydrate transport and metabolism		*[VX] Viral Processes
[L] Replication, recombination and repair	[T] Signal transduction mechanisms	[H] Coenzyme transport and metabolism		
*[X] Information processing	[U] Intracellular trafficking, secretion, and vesicular transport	[I] Lipid transport and metabolism		
	[V] Defense mechanisms	[P] Inorganic ion transport and metabolism		
	[W] Extracellular structures	[Q] Secondary metabolites biosynthesis, transport and catabolism		
	[Y] Nuclear structure			
	[Z] Cytoskeleton			

4. Análisis de secuencias virales

Alineamientos múltiples con familias celulares. Las secuencias virales obtenidas fueron agregadas a los alineamientos múltiples preexistentes de las familias de dominios por cuyo perfil fueron encontradas. Se empleó MAFTT v.7.205 con las mismas condiciones utilizadas anteriormente en las secuencias de las familias de ProDom.

```
$ mafft --localpair --maxiterate 1000 --add secuencias_virales.fasta alineamientos
```

Predicción de sitios de unión a RNA. Se empleó el predictor de sitios de unión a RNA en proteínas RNABindRplus, el cual realiza la predicción a partir de la combinación de dos metodologías distintas, una basada en homología de secuencias (HomPRID) haciendo búsquedas con BLAST en una base de datos de secuencias de proteínas de unión a RNA descritas experimentalmente (PRIDB, *Protein-RNA Interface Database*), y la otra metodología basada en aprendizaje automatizado, el cual utiliza un clasificador de *Support Vector Machine* (SUM) entrenado en una base de datos de 198 proteínas de unión a RNA conocidas (Walia et al., 2014). Éste software está disponible en la página web del *Artificial Intelligence Research Laboratory* de la Universidad de Pensilvania (http://ailab1.ist.psu.edu/RNABindRPlus/). Para el empleo de RNABindRplus fue necesario eliminar los gaps de las secuencias a analizar. Por otro lado, el software sólamente permite el análisis de cierto número de secuencias por vez, por lo que se determinó eliminar la redundancia de los grupos de secuencias virales obtenidas utilizando el software CD-HIT v. 4.6, herramienta que permite agrupar secuencias a distintos niveles de identidad (Li et al., 2001) a partir de un filtro de palabras cortas (Li & Godzik, 2006). Se empleó un valor límite de identidad de secuencia de 50% con un filtro de palabra de 3 como es recomendado para el valor de identidad seleccionado.

```
$ cdhit -i secuencias.fasta -o output -c 0.5 -n 3
```

Predicción de estructura secundaria. Se llevaron a cabo análisis de predicción de estructura secundaria empleando el servidor JPred4 (http://www.compbio.dundee.ac.uk/jpred/), el cual utiliza el algoritmo JNet, uno de los métodos más precisos para éste fin. Además, éste servidor hace predicciones de accesibilidad del solvente y de regiones coiled-coil (Drozdetskiy et al, 2015). Para el análisis se utilizaron los alineamientos múltiples anteriormente realizados de secuencias virales con las secuencias de las familias de dominios de la base de datos, ya que esto permite una mejora en la precisión de la predicción (Cuff & Barton, 1999). Cuando fue necesario se colocó la secuencia más larga en la parte superior del alineamiento con el fin de que el análisis cubriera el total de posiciones en el alineamiento, o al menos la mayor parte, debido a que el servidor toma la primera secuencia de los alineamientos como referencia para generar la predicción de estructura secundaria.

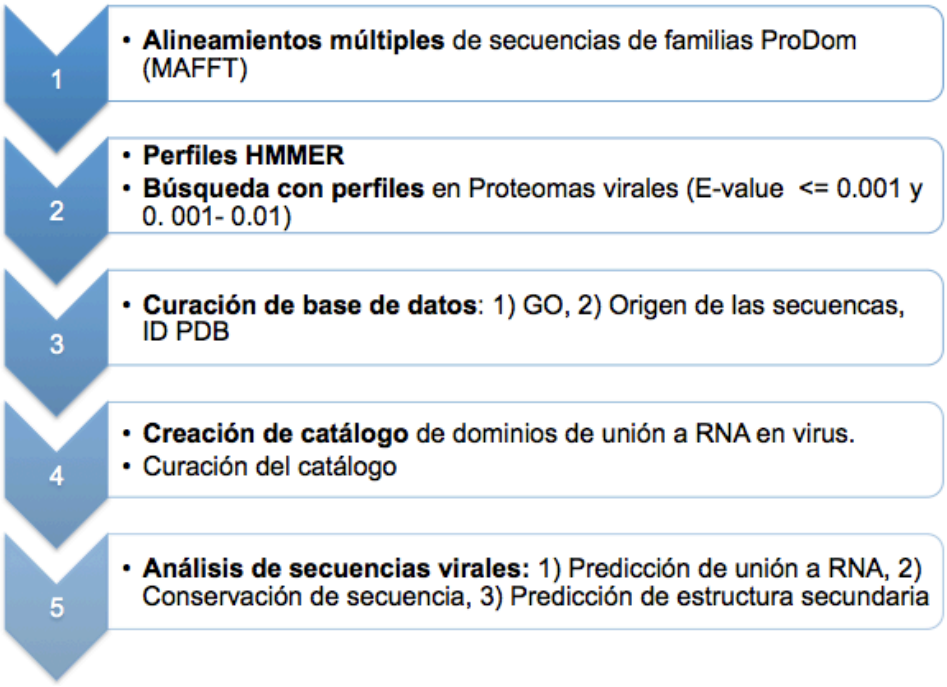


Figura 8. Análisis y procesamiento de datos. Se muestran de forma general, el orden seguido en los principales puntos realizados en la metodología para el análisis de los datos.

IV. RESULTADOS

Las familias de dominios de unión a RNA obtenidas de ProDom pertenecen a proteínas de los tres dominios celulares y de virus. En su gran mayoría las familias son de organismos del dominio Eukarya (~87.5%), mientras que los demás dominios representan menos del 10% (Bacteria ~5.5%; Archaea ~0.7%). Las familias de RBD pertenecientes a virus son 72 (~1.1%). Las 360 familias restantes pertenecen al menos a dos de los cuatro grupos mencionados anteriormente (figura 9).

Como resultado de la búsqueda de los 6,469 perfiles de familias de dominios proteínicos de unión a RNA en los proteomas virales se obtuvieron 1,360 perfiles que presentaron al menos un *match*. Una vez revisados y curados los resultados, el número final de familias de dominios proteínicos de unión a RNA que conformaron la base de datos fue de 438.

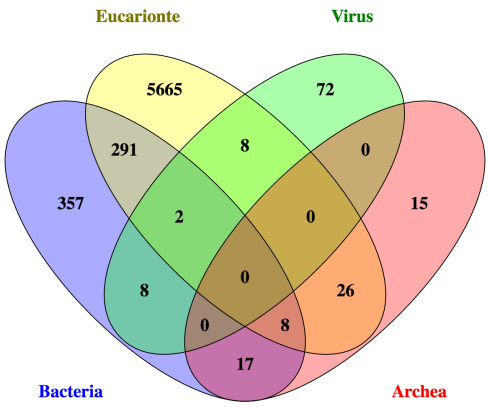


Figura 9. Número de familias de dominios proteínicos con función de unión a RNA distribuidas según su pertenencia a los Dominios celulares o virus. [1]

1. Curación de base de datos

Los resultados de estos 1,360 perfiles se revisaron manualmente para confirmar que presentaran valores estadísticos significativos o cercanamente significativos (*E-value* < 0.001 ó 0.001-0.01). Una vez realizada la revisión se redujo el número de perfiles con *hits* significativos a 500, mismos que conformaron una primer etapa de la base de datos final. Se realizó luego una curación en la que fueron eliminadas 35 familias de dominios: dos de ellas debido a que las secuencias que las conformaban han sido eliminadas en las bases de datos; 32 más, debido a que no presentan ninguna anotación GO, por lo que no fueron consideradas como datos fiables. Estas últimas familias, en su mayoría, pertenecen a organismos eucariontes y unas pocas a bacterias (cuatro familias) y virus (tres familias). También se descartó una familia adicional que obtuvo como resultado un solo *hit* en una secuencia de origen bacteriano, probablemente debido a un error de anotación en la base de datos de proteomas virales. Por último, se removieron 27 familias más que no presentaron anotaciones relacionadas con unión a RNA o nucleótidos, incluyendo familias relacionadas específicamente a unión a DNA o con anotaciones escasas que no permitían relacionarlas con una actividad de interés para éste trabajo. Así, la base de datos final tuvo un número total de 438 familias de dominios de proteínas de unión a RNA (figura 10). [2]

[1] Esta figura se realizó con la herramienta *Calculate and Draw custom Venn diagrams* del servidor web Bioinformatics & Evolutionary Genomics (http://bioinformatics.psb.ugent.be/webtools/Venn/).

[2] Por el momento la base de datos curada se encuentra disponible en el Laboratorio de Origen de la Vida de la Facultad de Ciencias, UNAM. En un futuro cercano estará disponible en un sitio web para su consulta.

De éste total de 438 familias, tres pertenecen exclusivamente al dominio Archaea, 30 a bacterias, 290 a eucariontes, y 69 a virus; las demás familias contienen secuencias de distintos dominios y/o virus, teniendo casi todas las combinaciones posibles: Archaea-Bacteria (5 familias), Archaea-Eukaryota (3), Bacteria-Eukaryota (26), Bacteria-Virus (7), Eukaryota-Virus (8), Bacteria-Eukaryota-Virus (1) (figura 11). En particular, las secuencias de estas familias pertenecen a 27 Phyla distintos de bacterias y 51 Clases taxonómicas; cinco Phyla y nueve Clases de archeas; así como 28 Phyla y 58 Clases de eucariontes (apéndice 1.1). Debido al tipo de información arrojada por UniProt para el caso de virus, estos fueron agrupados por su tipo de genoma en lugar de Phyla, y en familias virales en lugar de la Clase taxonómica (apéndice 1.2).

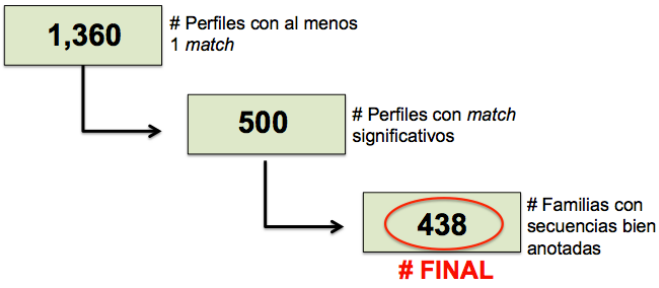


Figura 10. Proceso de curación de base de datos, en donde los números indican el número de perfiles de proteínas de las familias de dominios de unión a RNA. El primer valor incluye todos aquellos perfiles con al menos un *match* en la búsqueda (*E-value* = 0.1). El valor significativo de *E-value* fue de < 0.01, aunque también se incluyeron valores cercanos a éste umbral (<0.01).

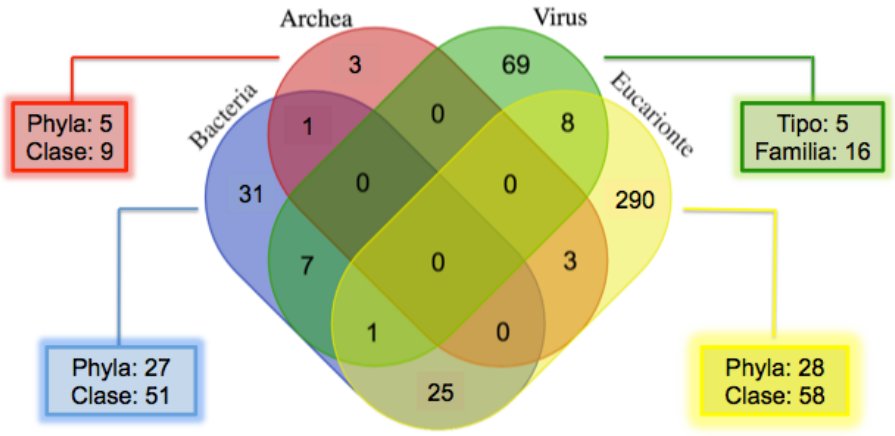


Figura 11. Distribución de las familias de proteínas de unión a RNA incluidas en la base de datos final, según la pertenencia de sus secuencias a los Dominios celulares o virus (Diagrama de Venn[3]), así como el número de Phyla y Clases para cada grupo en el diagrama.

[3] El diagrama de Venn en esta figura fue realizado, al igual que el presentado en la figura 4, con el servidor web Bioinformatics & Evolutionary Genomics.

Anotaciones GO. Respecto a las anotaciones obtenidas de *The Gene Ontology Consortium*, éstas fueron revisadas para clasificarlas en categorías generales a las que pertenecen, lo que permitió observar, a grandes rasgos, cómo están agrupadas. Para el caso de las anotación de *Cellular component* se encontraron 219 anotaciones distintas que se agrupan en dos grandes campos: componentes intracelulares y extracelulares; para éste último sólo se tienen unas cuantas anotaciones relacionadas a proyecciones celulares, uniones celulares, y procesos extracelulares de membrana como la sinapsis. Para el caso de los componentes intracelulares, estos pueden reagruparse en organelos celulares, complejos catalíticos, virión, y célula del hospedero, misma que se subdivide en organelos. El mayor número de anotaciones pertenece a componentes del núcleo, complejos proteínicos y ribonucleoproteínas del citoplasma, y la mitocondria. Todos estos generalmente relacionados a procesos con RNA (figura 12a).

Para el caso de las anotaciones de *Molecular function* se obtuvieron 236 anotaciones distintas, las cuales fueron agrupadas principalmente en: *1)* Funciones de unión a proteínas, a iones, y mayoritariamente a ácidos nucleicos; *2)* función de actividad catalítica, en general actividades enzimáticas de modificación de nucleótidos; *3)* transporte; y *4)* transducción de señales (figura 12b).

Por último, las anotaciones para *Biological process* fueron las más abundantes con 655 anotaciones distintas, siendo los principales grupos: *1*) Procesos celulares, que incluye crecimiento, muerte y proliferación celular, entre otros; *2)* señalización, básicamente procesos de transducción de señales; *3)* localización, en donde se hallan procesos de transporte, localización de organelos y de macromoléculas, principalmente proteínas y RNA ; *4)* procesos virales; *5)* procesos de desarrollo, como desarrollo embrionario, de tejidos y sistemas, así como ciclo celular y diferenciación celular; y *6)* procesos metabólicos, principalmente expresión genética, así como metabolismo del RNA y DNA, proteínas y compuestos orgánicos, entre otros (figura 12c).

2. Catálogo de dominios de unión a RNA virales[4]

El número de *hits* que obtuvimos en los proteomas virales fue de 3,956 en total, de los cuales 2,995 (75.7%) presentan un *E-value* significativo (< 0.001), mientras que

[4] El catálogo completo de dominios de unión a RNA virales se encuentra disponible de manera temporal en el Laboratorio de Origen de la Vida de la Facultad de Ciencias, UNAM. En un futuro cercano estará disponible en un sitio web para su consulta.

961 (24.3%) están en el rango de < 0. 001- 0.01 (figura 13), por lo que se pueden considerar con una posible relación evolutiva (Pearson, 2013).

Las familias de dominios de unión a RNA con origen eucarionte fueron las que mayor número de *hits* obtuvieron (2,009 *hits*), seguidas por las de origen viral (813) y bacteriano (462). Las otras familias obtuvieron un número de *hits* muy inferior, siendo las origen Archaea, Archaea-Bacteria y Archaea-Eukaryota las que presentaron los números más bajos (tabla 4, figura 14).

Diversidad viral. Los *hits* obtenidos fueron asociados a grupos virales según la clasificación de Baltimore. Un total de 3,951 *hits* pudieron ser asociados, teniendo una distribución que cubre el total de categorías (siete), siendo el grupo predominante el de los virus de DNA de doble cadena (dsDNA) (grupo I de Baltimore) y como grupo menos representado el de los virus de DNA de cadena sencilla (ssDNA) (grupo II) (tabla 5). Las familias virales a las que pertenecen estos mismos *hits* fueron 22 para el caso de virus de dsDNA, siendo *Myoviridae* la más recurrente; dos y tres familias para los virus de ssDNA y de RNA de doble cadena (dsRNA) (grupo III); siete para los virus de RNA de cadena sencilla (ssRNA), tanto de polaridad positiva como negativa (grupos IV y V), destacándose *Coronaviridae* y *Arenaviridae*. En el caso de los retrovirus, dos familias para los de dsDNA (grupo VII de Baltimore) y solo una para retrovirus ssRNA (grupo VI) (apéndice 2). Asimismo, el total de variedades virales distintas por las que está conformado éste catálogo es de 931. La forma en que están asociados estos grupos encontrados con las familias de dominios de la base de datos se puede observar en la figura 15, en donde, tanto la mayor diversidad viral como la cantidad de resultados están asociados a familias de dominios de origen eucariontes, virales o bacterianas principalmente. En esta misma figura se destacan las familias de virus de dsDNA *Poxviridae*, *Mimiviridae* y *Myoviridae*, así como las de retrovirus *Reovirida*e (ssRNA) y *Caulimoviridae* (dsDNA).

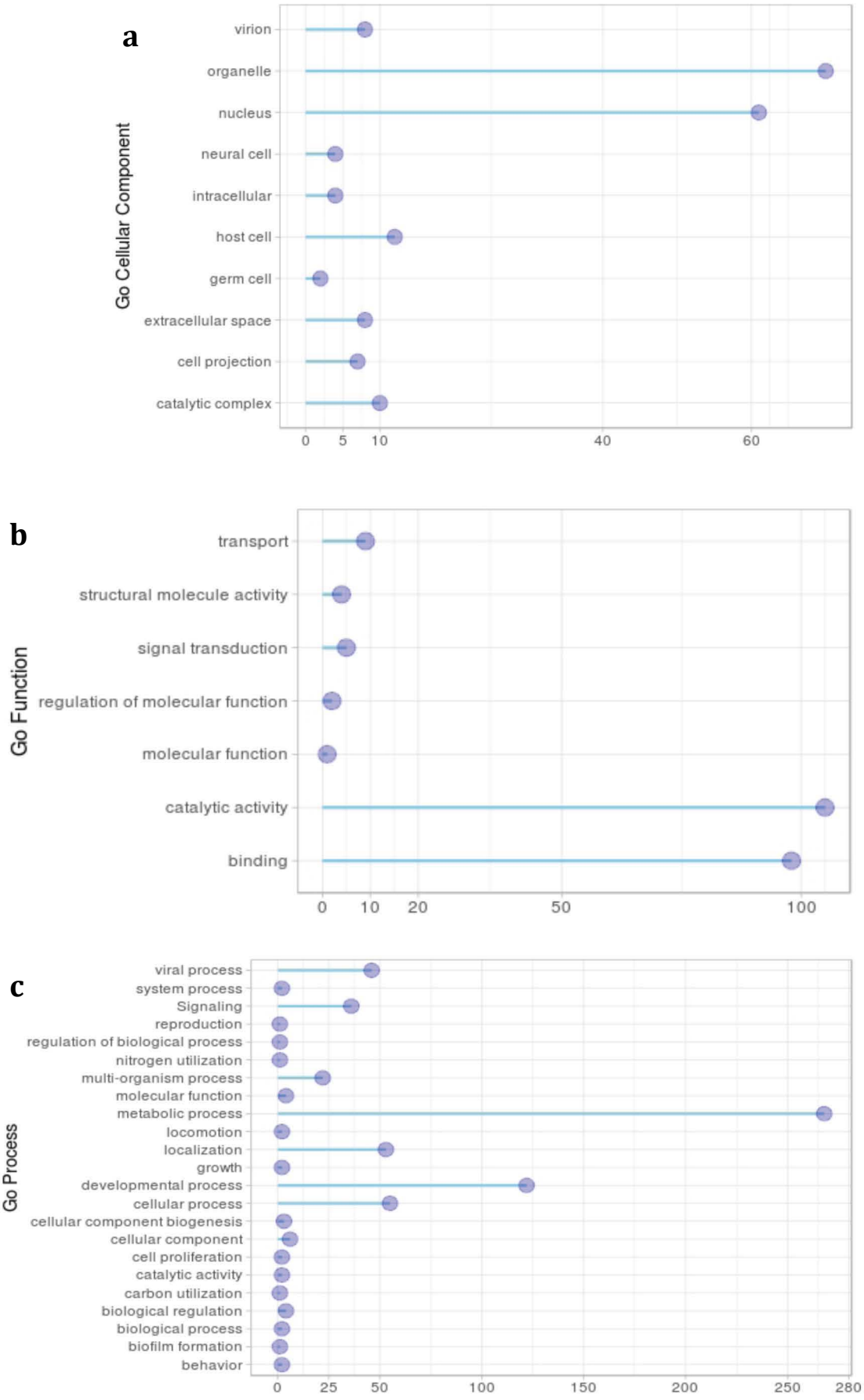


Figura 12. Número de ocurrencias por categorías generales de *The Gene Ontology Consortium* asociadas a las familias de dominios proteínicos de unión a RNA de la base de datos final. **a.** *Cellular component*. **b** *Molecular function.* **c.** *Biological process*

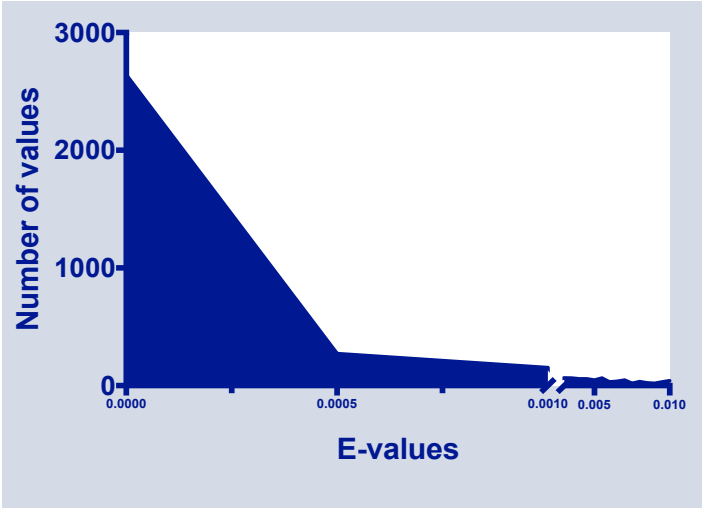


Figura 13. Distribución de frecuencias de valores de *E-value* categorizados en grupos de 0.0005.

Tabla 4. Número total y porcentaje de *hits* obtenidos en la búsqueda de perfiles en proteomas virales. Los *hits* se agrupan según el origen de las familias cuyos perfiles los obtuvieron (A: Archaea; B: Bacteria; E: Eukaryota; V: Virus).

Número y porcentaje de *hits* hallados por perfiles de familias de RDB según su origen		
A	5	0.12 %
AB	2	0.05 %
AE	8	0.2 %
B	462	11.67 %
BE	79	1.99 %
BV	184	4.65 %
E	2009	50.78 %
EV	219	5.53 %
V	813	20.55 %
BEV	175	4.42 %
TOTAL	3956	

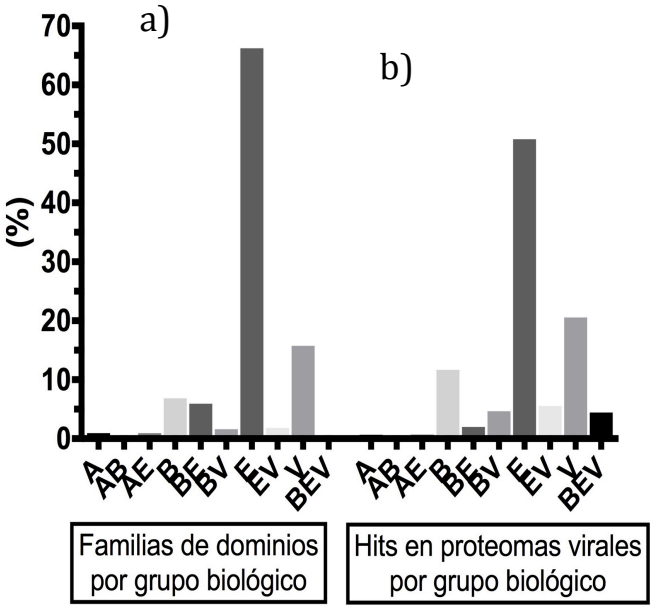


Figura 14. **a)** Porcentajes de las familias de dominios de unión a RNA (**438)** de acuerdo a la pertenencia de sus secuencias a grupos biológicos (A = Archaea; B: Bacteria; E: Eukaryota; V: Virus). **b)** Porcentaje de los hits encontrados en los proteomas virales de acuerdo a los grupos biológicos al que pertenecen los perfiles que les encontraron (total = **3,956** hits).

Tabla 5. Número de hits asociados a alguno de los siete grupos virales de Baltimore

Número de hits asociados a grupos virales de Baltimore	
dsDNA	2564
ssDNA	3
dsRNA	15
ssRNA (+)	105
ssRNA (-)	550
Retrovirus dsDNA	399
Retrovirus ssRNA	315
TOTAL	3951

El número de perfiles de familias de dominios a partir de las cuales se logró obtener secuencias virales fue de 429, los restantes, pese a obtener valores significativos en la búsqueda, no presentaron secuencias particulares con valores significativos, por lo que no fueron recopiladas. El total de secuencias obtenidas de los proteomas virales fue de 3,749.

Longitudes máximas de secuencia. En relación con las longitudes máximas de las secuencias de las familias de dominios y las secuencias virales encontradas se puede observar una reducción generalizada, con excepción de las secuencias pertenecientes a familias virales (tabla 6 y figura 16). El promedio de la longitud máxima de las secuencias virales se mantiene dentro del rango de longitud para lo establecido en la definición de dominio en proteínas (Traut, 2014), aunque en el caso de las secuencias halladas por perfiles de familias de origen Archaea-Bacteria y Archaea-Eukaryota, éste promedio se encuentra cerca del límite inferior (tabla 6).

En el caso de las longitudes máximas de las secuencias de las familias de la base de datos, los valores mínimos son de 20 aminoácidos, menos en el caso particular de la familia con origen Bacteria-Eukaryota-Virus. Los valores máximos igualmente presentan valores no muy variables entre sí (71-440 aminoácidos), con excepción de la secuencia viral de 769 aminoácidos que se aleja de la distribución de las demás secuencias tanto virales como celulares (tabla 6 y figura 16).

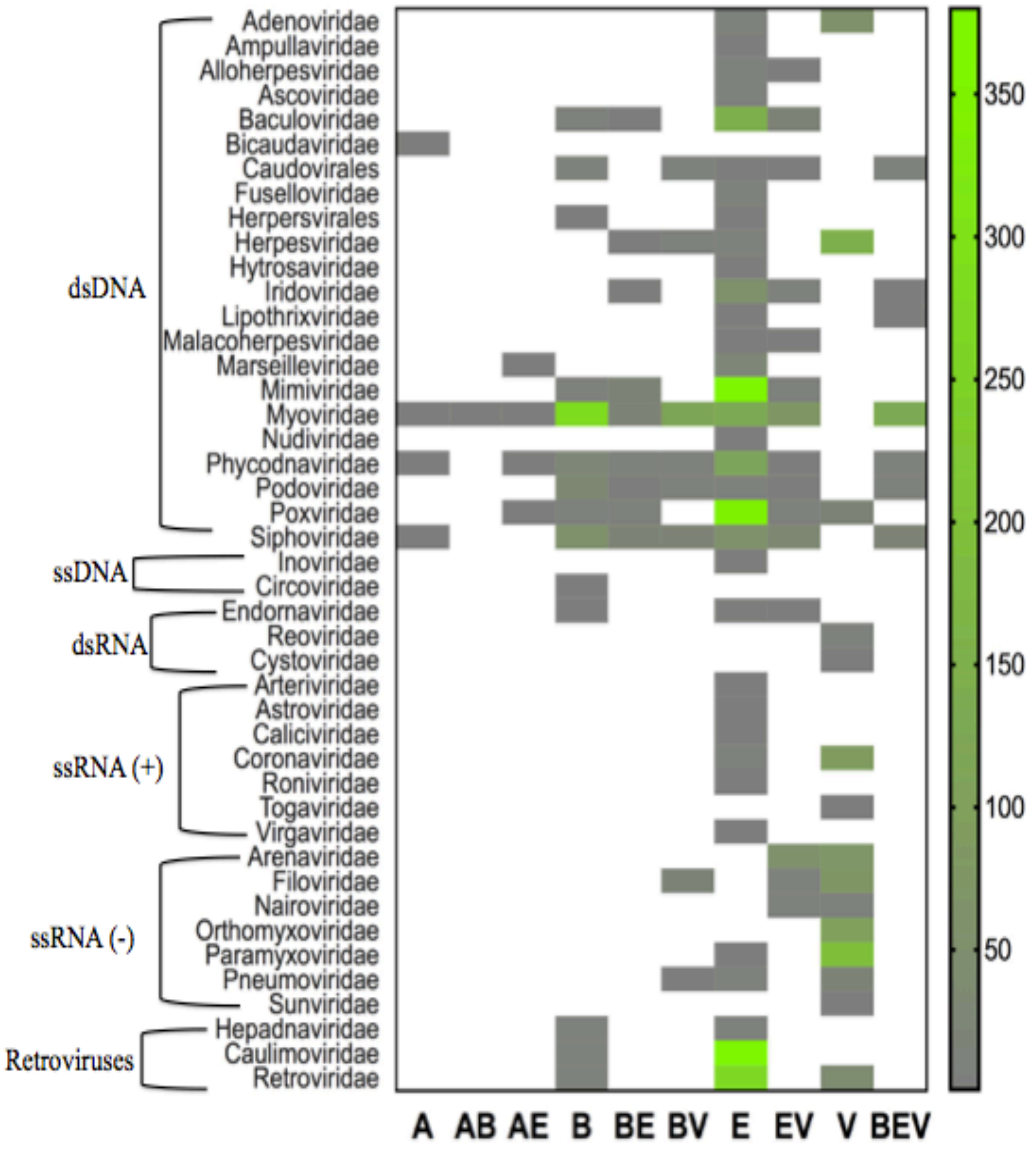


Figura 15. Heatmap que muestra la distribución de los hits encontrados en los proteomas virales, organizados en familias virales (líneas), y relacionado a familias de dominios agrupadas por grupos biológicos (columnas) a los que sus secuencias pertenecen.

En el caso de las secuencias virales obtenidas, de manera general, estas presentan una reducción en el número de aminoácidos, con varios casos de secuencias con valores mínimos debajo de 20 aminoácidos. Pese a ello, los promedios, valores máximos y medias, permiten ver que en su mayoría las longitudes de las secuencias se mantiene dentro o cerca de los rangos normales de la definición de un dominio proteínico (tabla 6).

Tabla 6. Valores relativos a las longitudes máximas de las secuencia de las familias proteínicas agrupados por Dominio celular y virus a los cuales sus secuencias pertenecen (A: Archaea; B: Bacteria; E: Eukaryota; V: Virus), así como los hits encontrados en los proteomas virales por estas mismas familias de dominios.

	A	AB	AE	B	BE	BV	E	EV	V	BEV
Long. máxima de secuencias de Familias en catálogo										
Min.	49	71	25	20	22	27	20	33	20	132
Media	129	71	48	69.5	65	72	67	64.5	141	132
Máx.	167	71	99	205	440	116	410	323	769	132
Promedio	115	71	57.33	83.63	87.5	72	86.57	100.9	175	132
Long. máxima de secuencias virales obtenidas										
Min.	22	38	12	15	17	26	9	33	20	88
Media	42	38	33	48	33.5	72	37	61	131	88
Máx.	124	38	34	183	182	109	265	290	790	88
Promedio	62.6	38	26.3	58.24	46.2	70.8	43.75	92.25	175.3	88

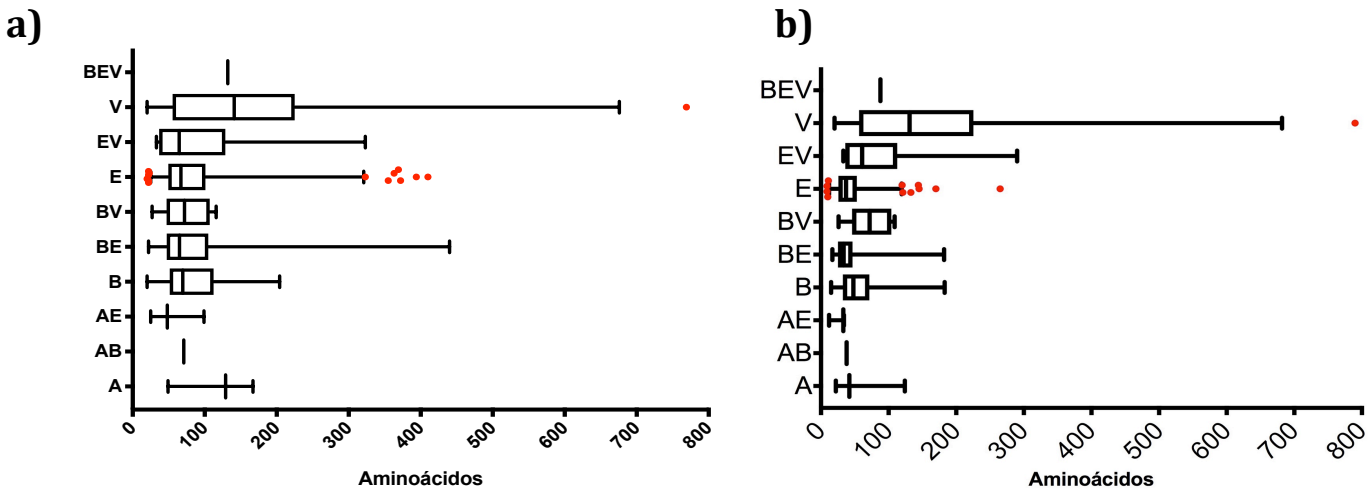


Figura 16. Diagramas de caja con máximos y mínimos de 2.5% y 97.5%. **a)** Se muestra las diferencias entre las longitudes máximas de las secuencias de las familias de dominios que conforman la base de datos, agrupadas según el origen de sus secuencias (A: Archaea; B: Bacteria; E: Eukaryota; V: Virus). **b)** Se pueden observar la variación y diferencia de las longitudes máximas de las secuencias de los hits encontrados en los proteomas virales en relación a la familia de dominios proteínicos a partir de cuyo perfil fueron encontradas.

Distribución de los hospederos celulares. En éste estudio se recopiló la información taxonómica de los hospederos empleando la base de datos Virus-Host DB. De manera general, se tuvieron 28 hospederos pertenecientes al dominio Archaea, 1055 a Bacteria y 2873 de Eukaryota. La distribución de estos hospederos en relación con el grupo y familia viral puede verse en la figura 17, en la cual se muestra que los virus de dsDNA presentan la mayor diversidad de hospederos, mientras que los virus de los demás grupos virales de la clasificación de Baltimore parasitan en su gran mayoría a eucariontes, salvo los virus de las familias *Inoviridae* (grupo II) y *Cystoviridae* (grupo III) que tienen como hospederos células bacterianas.

Acerca de la diversidad encontrada en los hospederos celulares, los pertenecientes al dominio Archaea están clasificados en los taxa Crenarcheota y Euryarcheota; los taxa que conforman los hospederos del Dominio Bacteria son Actinobacteria, Bacterioidetes, Cyanobacteria, Deinococci, Firmicutes, Proteobacterias y Tenericutes, siendo el más recurrente Proteobacterias (574) seguido de Cyanobacterias (204) y Firmicutes (188) (apéndice 3.1). Los hospederos Eukaryota pertenecen a diversos grupos: Bicosoecida, Ascomycota, Isochrysidales, Longamoebia, Metazoa, Pelagophyceae, Phaeophyceae y Viridiplantae, siendo Metazoa el más común con 1839 ocurrencias, de las cuales 1073 pertenecen a mamíferos (apéndice 3.2 y 3.3).

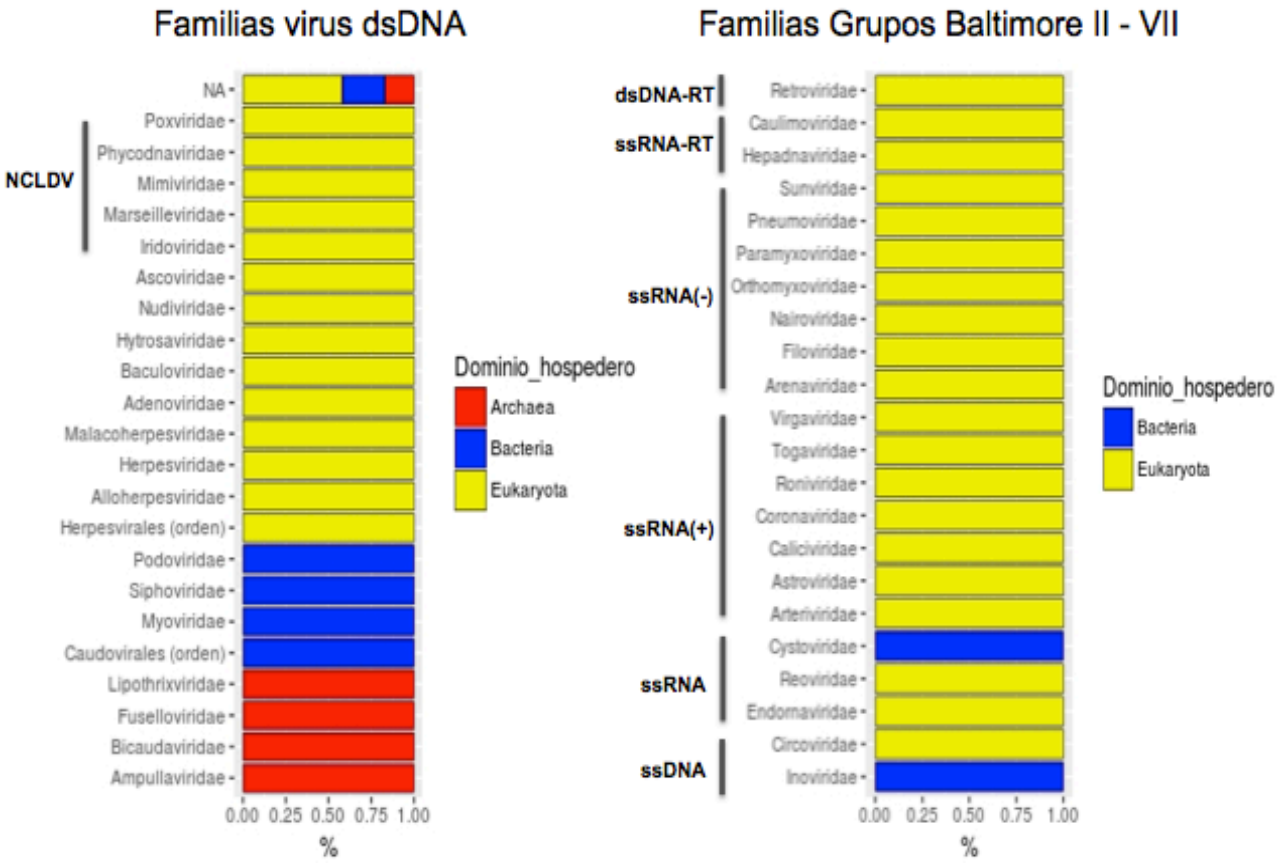
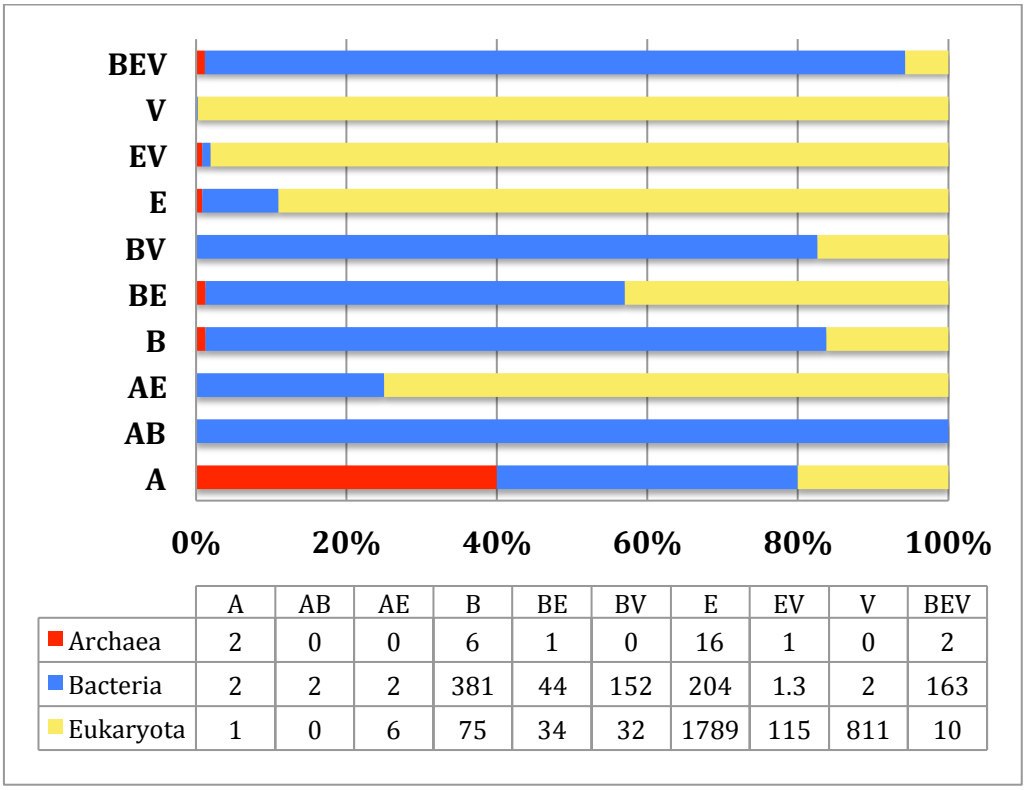


Figura 17. Distribución en porcentaje de los hospederos según su Dominio celular agrupados por el grupo de Baltimore y la familia viral que los parasita

Otra manera de ver la distribución de estos hospederos es relacionando el origen taxonómico de las familias de dominios de unión a RNA de la base de datos y el hospedero celular del virus cuya secuencia fue encontrada por estas mismas familias de dominios, para así observar si existe una relación entre estos datos. De esta manera,

se puede ver una coincidencia general en donde familias de dominios de origen bacteriano están presentes en proteínas de virus que predominantemente parasitan bacterias, lo mismo puede decirse para el caso de eucariontes, y también parcialmente para el de arqueas, en donde las pocas familias de dominios de origen arquea se encuentran en proteínas de virus que parasitan principalmente arqueas, pero también bacterias y eucariontes (figura 18).



	A	AB	AE	B	BE	BV	E	EV	V	BEV
Archaea	2	0	0	6	1	0	16	1	0	2
Bacteria	2	2	2	381	44	152	204	1.3	2	163
Eukaryota	1	0	6	75	34	32	1789	115	811	10

Figura 18. Distribución en porcentaje de hospederos según su Dominios celulares, cuyo parasito viral presentó en sus proteínas secuencias homólogas a las familias de dominios de unión a RNA, las cuales están agrupadas según su origen taxonómico en el eje vertical.

Funciones COG. La asignación de funciones COG a las proteínas del catálogo se realizó a partir de la información disponible en la base de datos UniProt, tomando en cuenta información GO cuando estaba disponible, estructuras asociadas, pero principalmente los dominios proteínicos detectados en las secuencias. Durante el proceso de asignación de funciones COG se decidió adicionar nuevas funciones relacionadas con los procesos exclusivamente virales como patogenicidad, estructuras de cápside, poliproteínas que presentan principalmente funciones de replicación y estructura, así como proteínas multifuncionales que están implicadas en diversos procesos, como replicación, transcripción o inmunoevasión, entre otros. También se

agregó una función general "[X]" para proteínas que están implicadas en más de una función del procesamiento de la información (tabla 3). Con esto se buscó evitar sesgos de las anotaciones y descripciones originales de las secuencias, las cuales en muchas ocasiones carecían de información o no coincidían con la información actualizada arrojada por UniProt.

Se reportó al menos una función de cada una de las cinco categorías generales de los COG que se emplearon. Del total de funciones existentes (incluyendo las adicionadas para éste trabajo), 18 fueron asignadas, siendo las más recurrentes las de la categoría "Pobremente caracterizadas", y particularmente las de función general; otras funciones recurrentes fueron las de la categoría "Almacenamiento y procesamiento de la información" y "Procesos virales", así como la función [P] *inorganic ion transport and metabolism*, que es la única función empleada de la categoría "Metabolismo". De igual manera, varias funciones COG de la categoría "Procesos y señalización celular" fueron asignadas, pero con una menor frecuencia (figura 19A, 19B).

Las proteínas asignadas con la función [R] *General function prediction only*, que ocupan más del 20% de las asignaciones, fueron aquellas cuya información no era suficiente para relacionarla con un proceso específico; sin embargo, contienen información para asociarlas a ciertos procesos o estructuras, mismos que pueden agruparse de manera general en: *1)* unión a nucleótidos, *2)* dedos de zinc, *3)* unión a ATP, *4)* modificación de nucleótidos e *5)* interacción entre proteínas (figura 19C). De éstas, la gran mayoría (684) están asociadas con interacción de proteínas, siendo los dominios más comunes los de las repeticiones de ankyrinas, repeticiones WD40, repeticiones ricas en leucina, repeticiones FNIP y repeticiones F-box, que usualmente están asociadas a las repeticiones FNIP.

En el caso de las proteínas con asignación [P] *inorganic ion transport and metabolism*, éstas son las únicas dentro de la clasificación general de "Metabolismo", y en todas las ocasiones se trata de la misma proteína: *Phosphate starvation-inducible protein PhoH*, la cual es parte del regulón de fosfato, una vía metabólica típica de bacterias (Santos-Beneit, 2015).

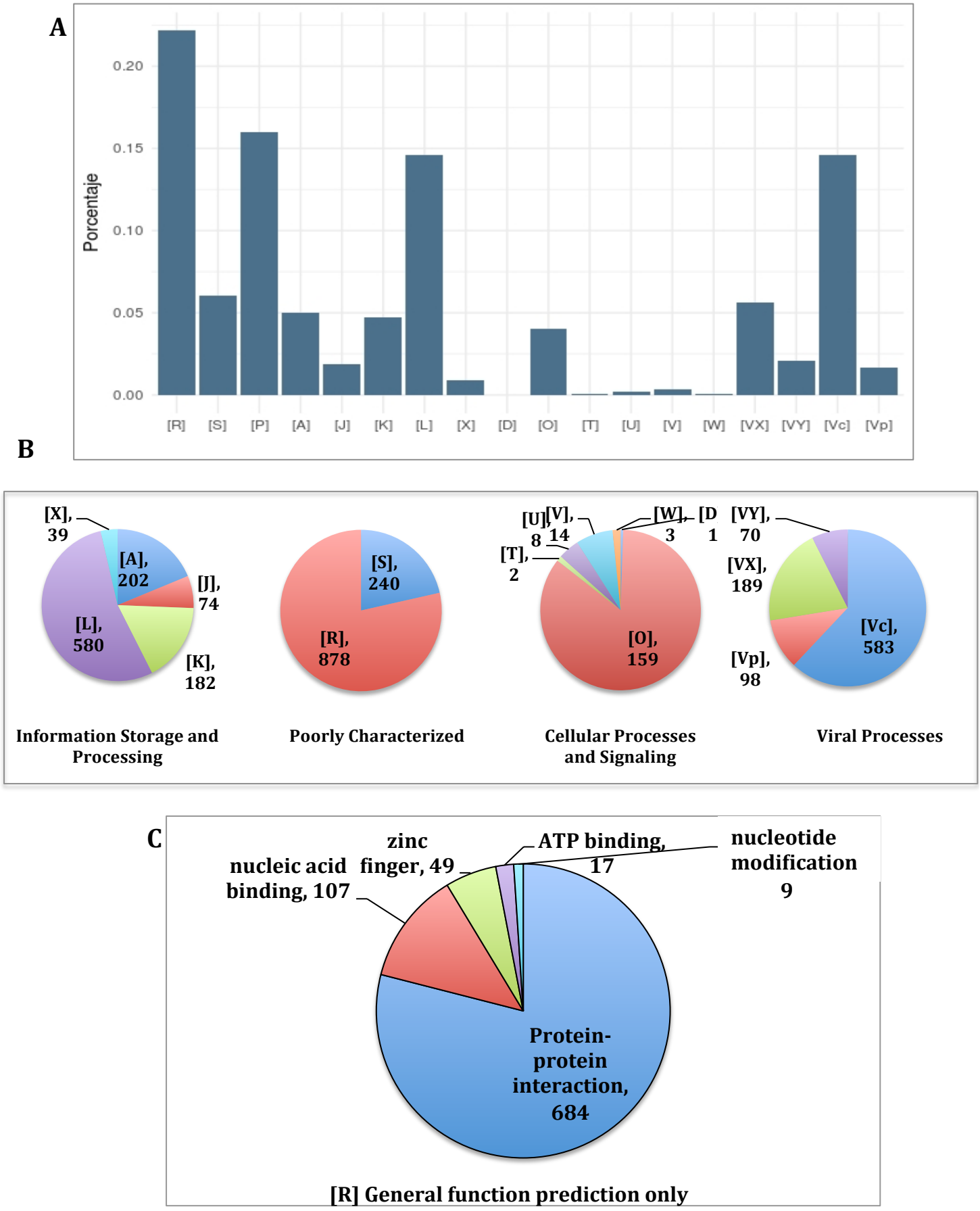


Figura 19. Frecuencia de funciones COG (tabla 3) asignadas a las proteínas virales que conforman el catálogo. A. Frecuencia en porcentaje del total las funciones asignadas. B. Conteo de funciones asignadas según su clasificación general de funciones COG. C. Conteo de las estructuras y funciones que conforman la función COG [R] *General function predicton only*.

Distribución de funciones COG en grupos virales. La manera en que las funciones COG se distribuyen en relación con los grupos virales según su tipo de genoma es dispar. El grupo de virus con genoma dsDNA (grupo I de Baltimore) es el que contiene la mayor diversidad de funciones, teniendo representantes de todas las categorías generales; en su gran mayoría son proteínas pobremente caracterizadas, seguido de proteínas relacionadas con metabolismo (proteínas phoH), exclusivamente de transporte y metabolismo de fosfato, y con números similares aparecen las funciones relacionadas al procesamiento de la información. Para el caso de los virus con genoma ssDNA y dsRNA (grupos II y III de Baltimore), estos aparecen en cantidades muy reducidas; sin embargo, éstas pocas proteínas incluidas se asociaron a funciones de procesamiento de la información, así como a funciones poco caracterizadas y exclusivas de procesos virales. Algo similar se muestra para los virus de ssRNA, tanto positivos como negativos, en donde predominan las funciones de procesos virales. En cuanto a los retrovirus, tanto los del grupo VI y VII, presentan mayoritariamente funciones de procesamiento de la información y de procesos virales, con la peculiaridad de que los retrovirus de ssRNA, en su mayoría, presenta funciones de procesos virales, mientras que en los retrovirus de dsDNA se da lo contrario, teniendo como funciones predominantes las del procesamiento de la información (figura 20). Únicamente cinco secuencias no pudieron ser asociadas a un grupo viral de Baltimore debido a la falta de información en distintas bases de datos.

Estos mismos datos muestran interesantes patrones una vez analizados al nivel de las familias virales. Las familias que conforman a los virus de dsDNA pueden ser agrupadas en virus que parasitan arqueas, bacteriófagos, virus de animales, y por último el de los virus gigantes. El grupo de virus de arqueas conformado por las familias *Ampullaviridae*, *Bicaudaviridae*, *Fuselloviridae* y *Lipothrixviridae*, es el que presenta menor número de datos, mismos que se agrupan entre los poco caracterizados y los de almacenamiento y procesamiento de la información. Por su parte, el grupo de bacteriófagos (Orden Caudovirales y famlias *Myoviridae*, *Siphoviridae* y *Podoviridae*) es el que presenta mayores frecuencias y una amplia diversidad de funciones, destacando la metabólica [P], la de procesamiento del RNA [A], las relacionadas a replicación, recombinación y reparación [L], así como las poco caracterizadas. El siguiente grupo, el que conforman los virus que parasitan animales, puede ser subdividido en virus que afectan vertebrados (*Alloherpesviridae*,

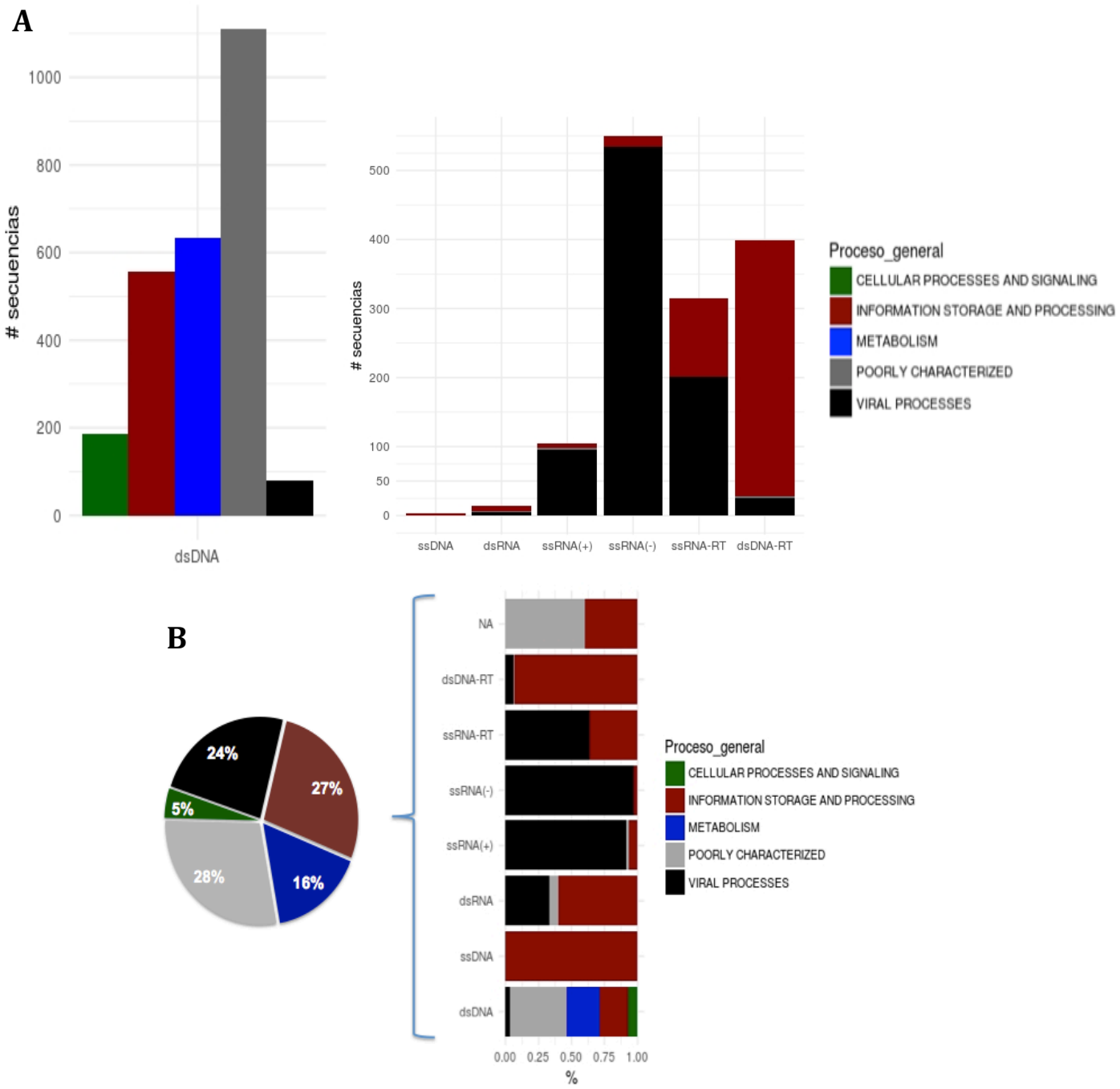


Figura 20. Distribución de funciones COG en relación a los grupos virales de Baltimore que conforman el catálogo de dominios de unión a RNA en virus. A. Conteo de secuencias con función COG asociada y organizadas por grupos virales de Baltimore. B. Porcentaje de funciones COG que conforman el total para cada uno de los grupos de Baltimore.

Herpesviridae, y *Adenoviridae*) y virus que infectan invertebrados (*Malacoherpesviridae*, *Baculoviridae*, *Hystrosaviridae*, y *Nudiviridae*). Éste grupo de virus de animales concentra sus funciones en los procesos virales, funciones poco caracterizadas y algunos del almacenamiento y procesamiento de la información; un caso que destaca es la alta frecuencia en la función de modificaciones postranscripcionales y chaperonas [O] en la familia *Baculoviridae*. Por último, los virus gigantes (*Ascoviridae*, *Iridoviridae*, *Marseilleviridae*, *Mimiviridae*,

Phycodnaviridae y *Poxviridae*) presentan una gran cantidad de proteínas poco caracterizadas, pero también cuentan con una amplia diversidad de funciones dentro de las distintas clasificaciones generales COG (figura 21).

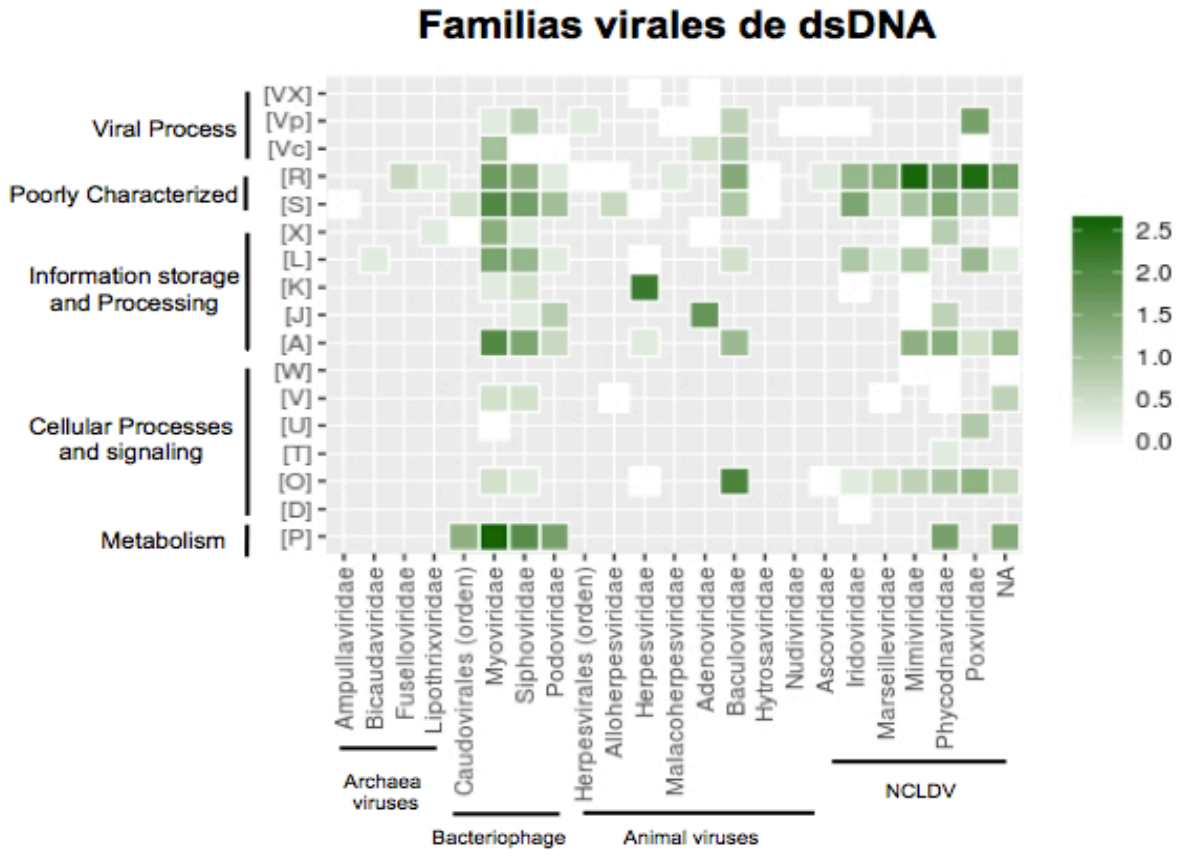


Figura 21. Heatmap que muestra la frecuencia y distribución de las funciones COG asignadas en relación con las familias virales del grupo I de Baltimore que conforman el catálogo. La escala ha sido normalizada con log10 para mejorar la visualización y comparación de los datos.

En las Familias que conforman los grupos de Baltimore del II al VII la gran mayoría de las funciones están asociadas a los distintos procesos virales, principalmente a estructuras de cápside y membrana [Vc]. Como excepción, los virus del grupo II con genoma de DNA de cadena simple (*Inoviridae* y *Circoviridae*) sólo presentan funciones relacionadas a replicación, recombinación y reparación [L]. De forma general, estas familias sólo contienen funciones que pertenecen a las clasificaciones generales de "Proceso viral", "Almacenamiento y procesamiento de la información" y "Pobremente caracterizadas". Sobresale el hecho de que los virus de cadena simple de RNA (grupo IV y V) presentan casi exclusivamente funciones asociadas a los procesos virales. Por su parte, los retrovirus (grupos VI y VII), además de funciones asociadas a procesos virales, presentan una alta frecuencia en la función

de replicación, recombinación y reparación [L], que principalmente incluye retrotranscriptasas (figura 22).

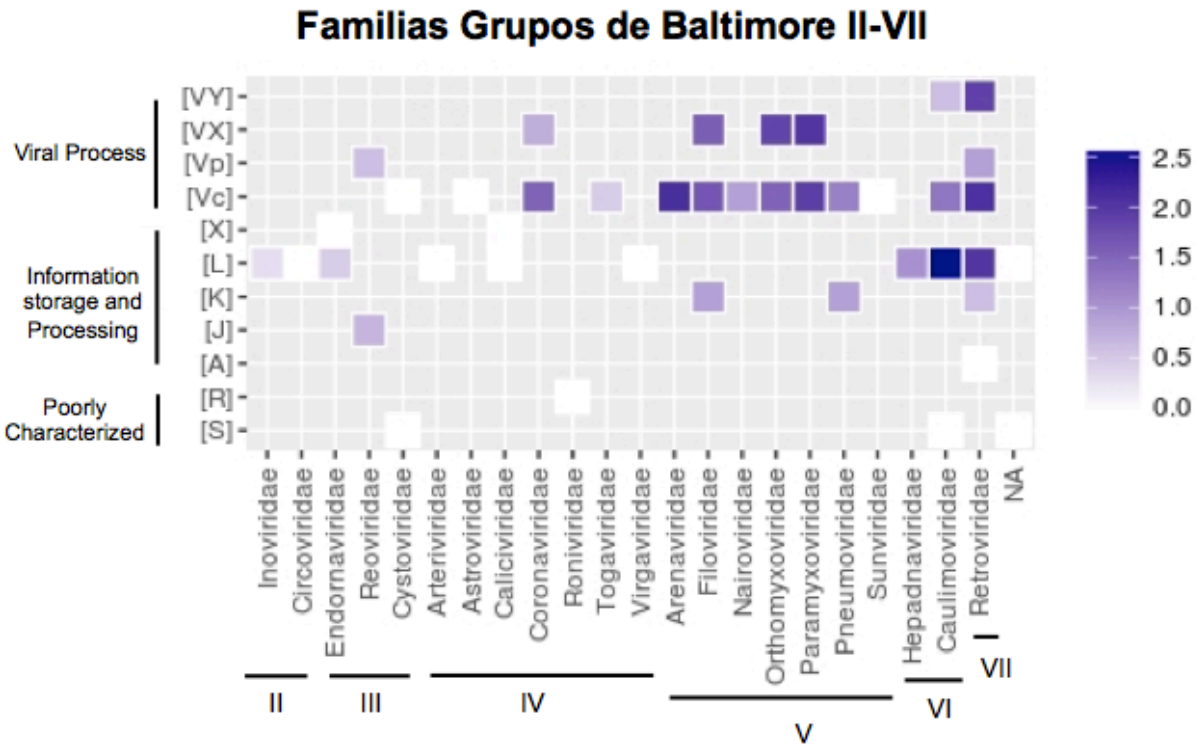


Figura 22. Heatmap donde se observa la frecuencia y distribución de las funciones COG asignadas en relación con las familias virales de los grupos II-VII de Baltimore que conforman el catálogo. La escala ha sido normalizada con log10 para mejorar la visualización y comparación de los datos.

Para visualizar estos mismos datos a una escala más fina, se graficó la distribución de las funciones COG, pero únicamente para los virus que infectan a los grupos más abundantes de hospederos registrados en el catálogo: Viridiplantae (Streptophyta y Chlorophyta) y Metazoa (vertebrados, artrópodos y moluscos) (apéndice 3.2 y 3.3). En el caso de los virus que infectan a organismos pertenecientes a Viridiplantae, la mayor cantidad de ocurrencias y diversidad de funciones recaen en un miembro de la familia *Phycodnaviridae*, una familia de virus gigantes que infecta algas; estas funciones se distribuyen en todas las categorías generales COG, menos en la de procesos virales, siendo las más destacadas las de [A] procesamiento de RNA, [P] transporte y metabolismo de iones inorgánicos, y [R] funciones generales. También hay ocurrencias en la familia *Retroviridae*, *Endornaviridae* y *Virgaviridae*, que tienen como hospedero a plantas (figura 23).

En cuanto al caso de virus que infectan Metazoa, éstos presentaron una mayor diversidad, tanto en familias virales, como en funciones COG (figura 23). De forma

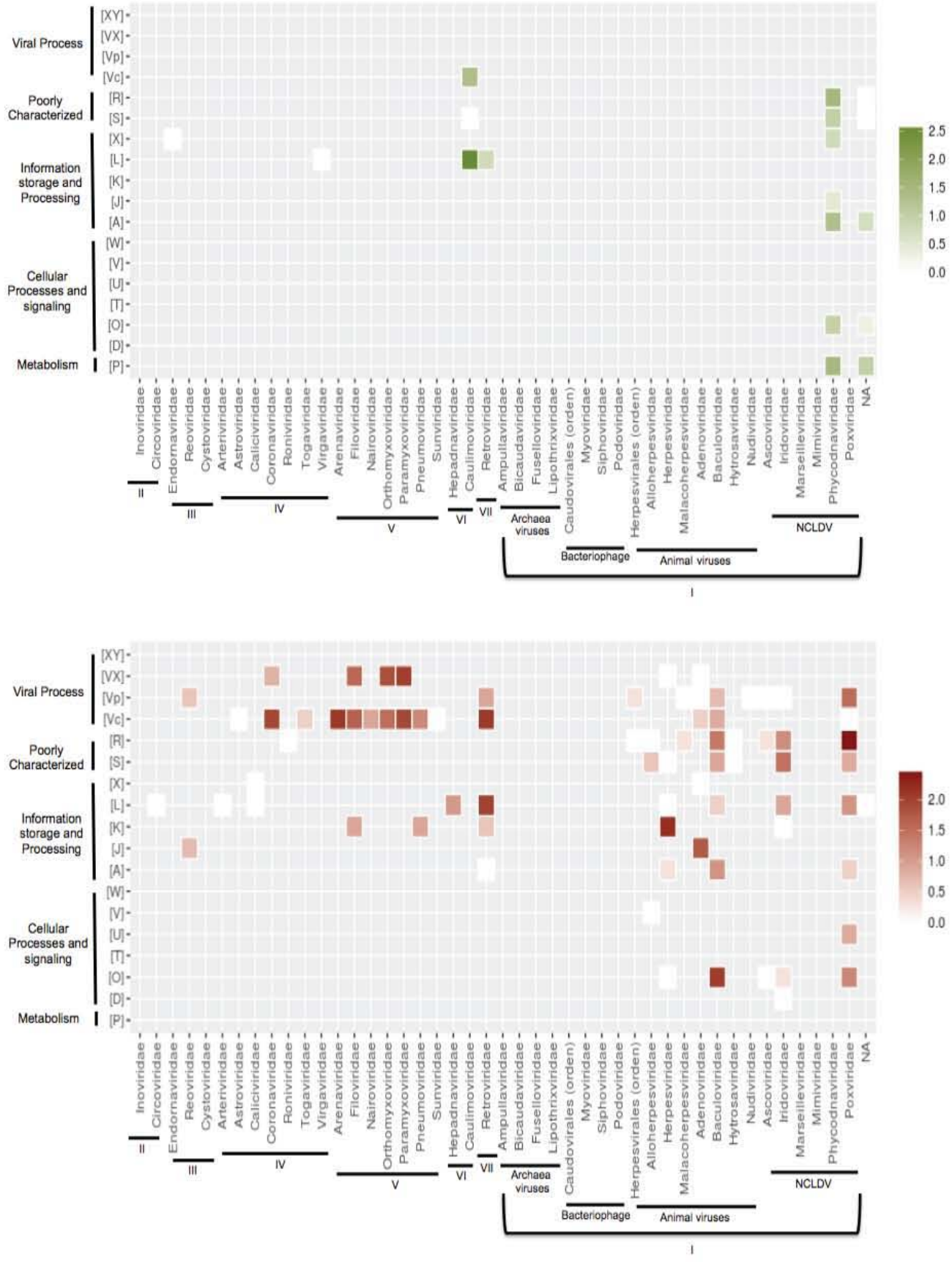


Figura 23. Heatmaps con distribución de funciones COG en familias virales que infectan eucariontes de los taxa Viridiplantae (superior) y Metazoa (inferior). De lado izquierdo se muestran las funciones COG agrupadas por sus categorías generales, y en la parte inferior los números indican el grupo viral de Baltimore al que pertenecen las familias virales. La escala ha sido normalizada con log10.

general, las funciones se distribuyen principalmente en las categorías funcionales de procesos virales, pobremente caracterizadas y en la de almacenaje y procesamiento de la información. En cuanto a la distribución de virus, se destacan los de ssRNA, principalmente de cadena negativa, los retrovirus, *Baculoviridae*, *Poxviridae* y algunos virus más de dsDNA.

Se destaca que en el caso de los virus ssRNA(-) hay un gran número de ocurrencias en la función [Vc] estructura de cápside/membrana, al igual que en las familias *Retroviridae* (dsDNA-RT) y *Coronaviridae* (ssRNA (+)). También llama la atención el caso de la familia *Baculoviridae* y *Poxviridae* que presentan una alta diversidad en funciones, incluyendo casi todas las categorías generales COG, menos la de metabolismo (función [P]) (figura 23).

Arquitectura de las Proteínas de unión a RNA virales. A partir del supuesto de que cada secuencia representa una proteína, se estimó cuántos dominios de unión a RNA hay en cada proteína contenida en el catálogo, para así entender cómo se estructuran y distribuyen. Una vez realizada esta descripción, se clasificó a las proteínas como monodominio y multidominio (dos o más dominios). Se observó que en los virus de dsDNA predominan las proteínas monodominio, a diferencia de lo que ocurre con los virus de ssRNA y retrovirus, en donde las proteínas multidominio son más abundantes. Para el caso de los virus de ssDNA, de las pocas ocurrencias de estos virus, aparentemente todas las proteínas son monodominio, mientras que para los virus de dsRNA, pese a presentar ambas categorías de proteínas, las monodominio son más frecuentes (figura 24).

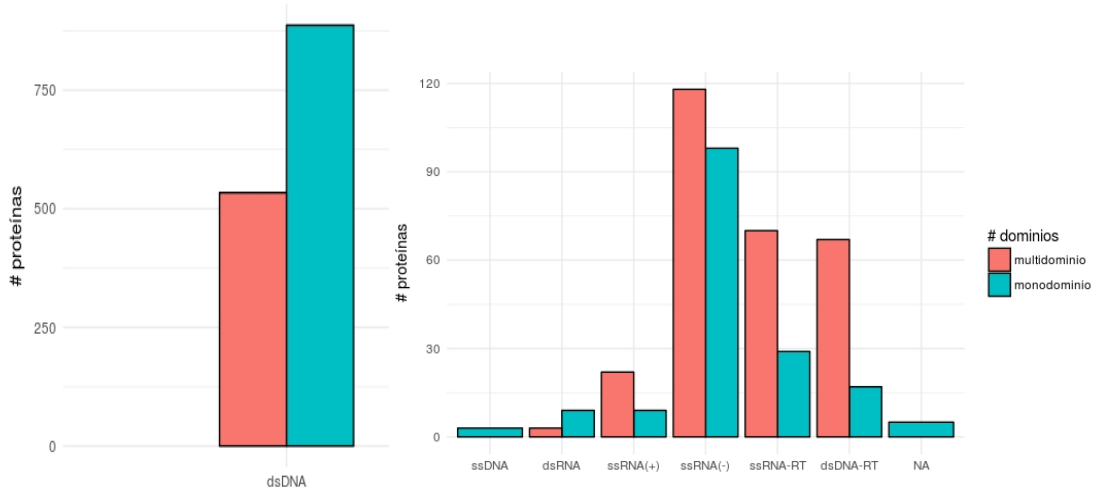


Figura 24. Número de proteínas monodominio y multidominio según el grupo viral de Baltimore.

Con la intención de explicar estas arquitecturas, se observó la distribución de estas mismas proteínas mono y multidominio en relación con las funciones COG. Lo que proporciona un mayor número de proteínas monodominio en los virus de dsDNA son las proteínas poco caracterizadas y las de procesamiento del RNA [A]. Respecto a las proteínas multidominio, estás se agrupan en las funciones [R] *General function prediction only* y [P] *inorganic ion transport and metabolism.* En las demás categorías, pese a las pocas ocurrencias, en su mayoría son proteínas monodominio (figura 25).

Por otro lado, la predominancia de proteínas multidominio en los virus de ssRNA y retrovirus se asoció a las funciones de replicación, recombinación y reparación [L], poliproteínas virales [XY] y cápside/membrana [Vc] (figura 25). Pero la distribución de éstas no es homogénea; para las funciones [L], la gran mayoría pertenece a retrovirus de dsDNA, siendo ésta la única categoría funcional en donde se destacan. Además todas estas proteínas aparecen como multidominio. Para la función [XV], la cantidad de proteínas mono y multidominio es similar, destacando un poco las monodominio. Cabe destacar en éste punto que éstas proteínas pertenecen casi en su totalidad a virus de ssRNA (-), mismo grupo que se destaca en la función [Vc] con proteínas multidominio. Por último los retrovirus de ssRNA son prácticamente los únicos con presencia en las funciones [Vp] y [VY] con proteínas multidominio.

3. Análisis de secuencias virales

Predicción de sitios de unión a RNA. Para realizar la predicción en las proteínas virales recuperadas se eliminó la redundancia de secuencias empleando CD-HIT v. 4.6 con los parámetros antes indicados. Se logró reducir el total de secuencias virales de 3,749 a 1,283. Estas secuencias se analizaron mediante RNABindRplus sin obtener resultados positivos.

Independientemente de la longitud de las secuencias analizadas se obtuvieron resultados dispares y poco congruentes entre los métodos empleados por RNABindRplus. Por un lado, se observaron algunos análisis sin predicción alguna, otros con predicciones de unión completamente negativas por ambos métodos o con predicciones positivas muy escasas. Por otro lado, se encontraron casos de análisis con predicciones muy dispares entre los dos métodos, en donde uno hallaba un número importante de residuos con posibilidad de unión a RNA, mientras que el otro

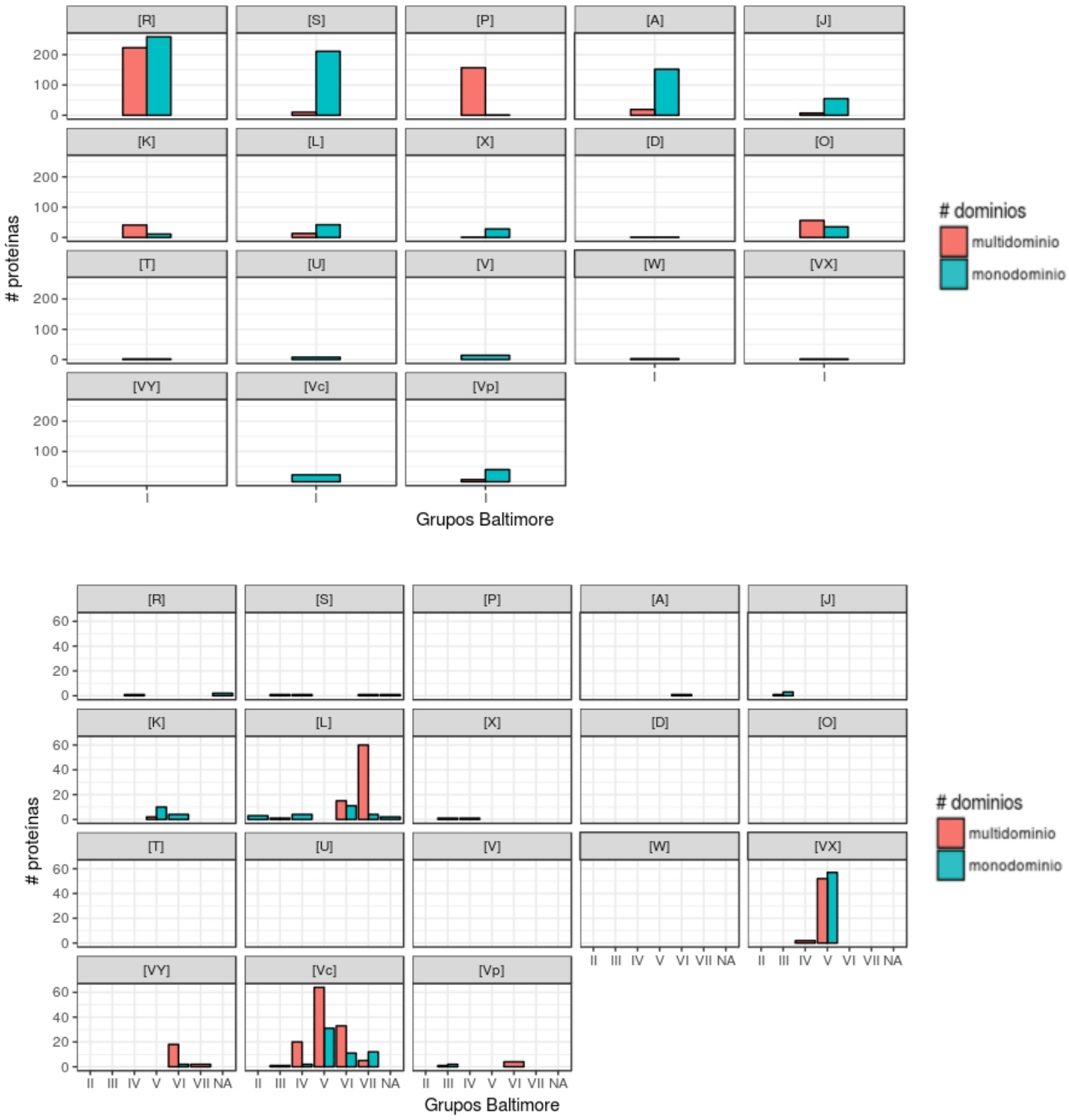


Figura 25. Número de proteínas mono y multidominio asociadas a funciones COG (tabla 3) y agrupadas por grupos virales de Baltimore: I (dsDNA), II (ssDNA), III (dsRNA), IV (ssRNA (+)), V (ssRNA (-)), VI (RT-ssRNA) y VII (RT-dsDNA). El primer conjunto de gráficas, situado en la parte superior, corresponde exclusivamente a los virus del grupo I de Baltimore, mientras que el conjunto inferior incluye a los demás grupos virales.

no hacía ninguna predicción positiva o muy pocas. En todos estos casos el resultado del consenso entre ambas metodologías se veía afectado por estas disparidades.

Solamente unos pocos análisis generaron predicciones positivas y congruentes entre los métodos. Por todo lo mencionado anteriormente acerca de la predicción de sitios de unión a RNA en las proteínas virales, se decidió no incluir estos resultados.

Predicción de estructura secundaria. Éste análisis fue generado a partir de los alineamientos múltiples de las secuencias que conforman cada familia de RBD de ProDom y sus respectivas contrapartes virales. De esta manera, se obtuvo una predicción general de la estructura secundaria de cada familia de dominios.

Los resultados arrojados por JPred4 fueron revisados manualmente para poder clasificarlos según su topología general en alguna de las grandes clases de dominios estructurales: *all-α, all-β, α+β y α/β.*[5]

Como se observa en la tabla 7, en su gran mayoría, las familias de dominios presentan una estructura *α+β y α/β*, seguido de dominios *all-α*. En estos mismos grupos se presentaron algunos pocos casos de predicciones para estructuras *coiled coil*, también conocidas como hélices superenrolladas. También ocurrieron predicciones en donde no se detectó ninguna estructura secundaria (tabla 7).

Algo notable fue el gran número de resultados en donde un porcentaje importante de la predicción generada estaba ocupada por regiones desestructuradas. Un poco más de la mitad de los análisis (55.47% de 429 análisis realizados) presentaron 50% o más de sus sitios sin estructura.[6]

Tabla 7. Frecuencia de estructuras secundarias obtenidas con JPred4 clasificadas en clases de dominios estructurales. Se muestra también el número de análisis donde se predijo la presencia de regiones *coiled coil.*

Motivo estructural	Frecuencia	Regiones *coiled coil*
all-β	46	-
all-α	116	17
α+β y α/β	247	8
Sin estructura	20	-
total	429	24

[5] La clasificación y descripción de las predicciones de estructura secundaria obtenidas de JPred4 pueden ser consultadas en el Laboratorio de Origen de la Vida de la Facultad de Ciencias, UNAM. En un futuro cercano estarán disponibles en un sitio web para su consulta.

[6] *Idem.*

V. DISCUSIÓN

Como se ha señalado anteriormente, las RBP juegan papeles críticos en un gran número de procesos celulares. Su relevancia queda demostrada por su alto grado de conservación y su universalidad, así como su gran abundancia en los distintos genomas. Específicamente, la interacción de los RBD con el RNA se da gracias a una serie de mecanismos, principalmente la cooperación multidominio, ya sea con dominios de la misma proteína o de otras cuando se involucran procesos de oligomerización.

Éste tipo de proteínas y sus dominios también están presentes en los virus, en donde participan en procesos fundamentales de su ciclo replicativo. De particular importancia es su función en la regulación del metabolismo celular de su hospedero. Por esto mismo, presentan una alta presión de selección, lo que permite emplearlas como marcadores moleculares para el estudio de la evolución de los virus y las interacciones con sus hospederos.

Tomando en cuenta que los genomas virales son reducidos y sus proteínas emplean arquitecturas multidominio para llevar a cabo distintos procesos, en el presente trabajo se emplearon únicamente las secuencias de RBD al considerar que serían más adecuados para ser rastreados en los proteomas virales.

La base de datos ProDom, a diferencia de otras bases de datos de RDB, tiene la característica de incluir a familias de dominios de una gran diversidad de organismos pertenecientes a los tres dominios de la vida, así como de distintos virus. Esto nos permitió realizar una búsqueda muy completa, considerando tanto información de las proteínas virales como de las de los organismos celulares que fungen como sus hospederos.

Es sabido que los RBD son más divergentes que los dominios catalíticos, y por tanto son menos reconocibles en búsquedas estándares de similitud de secuencia (Anantharaman et al., 2002). Por tal razón se utilizó una búsqueda por perfiles, que es un método más sensible para el reconocimiento de secuencias con un mayor grado de divergencia. Pese a ello, debido a la naturaleza altamente cambiante de los virus, es posible que una gran cantidad de datos no hayan sido recuperados a lo largo del análisis realizado, por lo que la diversidad y abundancia de RBD virales presentada en

nuestros resultados probablemente estén subestimadas. Por ello, se tomó la decisión de mantener dos rangos de *E-value*, buscando perder una menor cantidad de datos.

Otro factor importante que indudablemente afectó en los resultados de éste trabajo son los distintos sesgos presentes en las bases de datos. De forma general, la mayoría de datos pertenecen a organismos modelo o a aquellos que son de interés médico, económico o biotecnológico. Esta misma situación ocurre con la información disponible de virus, donde se destacan todos aquellos que infectan a organismos con las características mencionadas anteriormente. Por tanto, existen grandes huecos de información que limitan una descripción más completa de la diversidad biológica y sus relaciones.

Lo comentado en el párrafo anterior es especialmente visible en los datos que fueron obtenidos de ProDom para familias de dominios relacionadas a la función "*RNA binding*", los cuales pertenecen casi en su totalidad a eucariontes (~87.5%). Aunque en éste dominio hayan ocurrido importantes procesos de radiación de RBD, y su empleo es más extendido que en otros grupos, las grandes diferencias en los porcentajes observados con más seguridad demuestran una ausencia de datos en los otros dominios celulares (Archaea y Bacteria) y en virus (figura 9).

Aún con estas problemáticas, en el presente trabajo se ha logrado obtener una descripción cuantitativa y cualitativa de los dominios proteínicos relacionados con la función de unión a RNA presentes en los proteomas virales disponibles. Se obtuvieron datos pertenecientes a virus de los siete grupos de Baltimore, mismos que nos permitieron sugerir funciones e interacciones en las que los RBD están implicados, así como vislumbrar la estrecha relación evolutiva de los virus y sus hospederos.

Así, éste catálogo provee de un valioso recurso que puede servir de guía para futuras investigaciones sobre las interacciones virus-hospedero, para ayudar a entender algunos de los procesos de ciclo replicativo viral, o para simplemente estudiar las RBP y sus dominios con la finalidad de entender sus interacciones y los procesos en los que están involucrados.

1. Base de datos de familias de RBD

Del total de familias obtenidas inicialmente, únicamente los perfiles de 1,360 lograron generar resultados en la búsqueda en los proteomas virales. Un número importante de los 5,106 perfiles restantes que no obtuvieron resultados posiblemente estén relacionados a distintas rutas metabólicas o estén implicados en procesos exclusivos de los organismos celulares, como el sistema de traducción. Por tal razón, pudiera ser que no hayan encontrado secuencias análogas en los proteomas virales.

Una vez curada la base de datos, el número total de familias de RBD fue de 438, en donde las secuencias puramente eucariontes continuaron siendo las mayoritarias (290 familias) (figura 11). Respecto a las proporciones de las familias según su origen biológico, estas fueron similares en relación a las proporciones originales obtenidas de ProDom, siendo solo las familias de origen viral las que aumentaron su proporción en la base de datos de 1.11% al 15.75% (figura 11), debido a que todas las familias con secuencias virales se hallaron a sí mismas en los proteomas virales. Así, fue necesario descartar únicamente tres familias de origen viral debido a la falta de anotaciones GO que las relacionaran con la función de unión a RNA.

La diversidad de las secuencias que conforman a las familias de nuestra base de datos cubrió de forma general a los principales grupos biológicos y algunos representantes de casi todos los grupos virales de Baltimore. En el caso particular de las secuencias con origen Archaea, pese al poco número de ocurrencias y de familias en las que están presentes, se tienen representantes de todos los phyla reconocidos para éste dominio (Parte, 2018) (apéndice 1.1). Las secuencias bacterianas, por su parte, cubren una buena porción de los phyla reconocidos, teniendo 27 de 33 propuestos (Parte, 2018), así como 51 clases taxonómicas (figura 11). En éste grupo destaca por su abundancia y diversidad las de Proteobacteria, Firmicutes, Bacteroidetes y Actinobacteria (apéndice 1.1).

Aunque las secuencias de origen eucarionte fueron las más abundantes con relación a la diversidad de taxones superiores, éstas sólo cubren una porción del total para eucariontes. Existen grandes faltantes, sobre todo de los reinos Fungi, Protoctista, y Plantae. Aun así, hay representantes de los principales grupos que conforman éste dominio (28 phyla y 28 clases), en donde destacan por su diversidad y abundancia los phyla Arthropoda, Ascomycota, Chordata y Streptohyta, y de forma más específica

las clases Insecta, Saccharomycetes, Sordariomycetes, Eurotiomycetes, Mammalia, Actinopteri y Liliopsida (apéndice 1.1).

Por último, la fracción de secuencias virales que conforman la base de datos contiene cinco de los siete grupos de Baltimore, faltando únicamente los virus con genoma de ssDNA (grupo II) y los retrovirus de dsDNA (grupo VII). El grupo viral más representado fue el de los virus ssRNA (-) (grupo V), seguido de los virus de dsDNA (grupo I).

Por ello, consideramos que la cobertura de la diversidad biológica obtenida justifica el empleo de la base de datos ProDom sobre otras bases de datos disponibles como ATtRACT (Giudice et al., 2016), PROSITE (Sigrist et al., 2002) o RBPDB (Cook et al., 2010), las cuales, aunque puedan estar mejor curadas, por su restricción a organismos modelo o a la organización de sus datos, resultaron menos indicadas para los objetivos de éste trabajo.

Anotaciones GO. El incorporar las anotaciones GO a las familias de RBD de la base de datos tuvo dos propósitos. Por un lado, asegurarnos que las secuencias de las familias efectivamente presentaban una relación con procesos que involucran al RNA, y por el otro, servir como guía al interpretar y comparar con las secuencias virales encontradas. El gran número de anotaciones distintas que fueron agregadas obligó a reagruparlas en las categorías jerárquicas superiores a las que pertenecen, para así poder visualizarlas y describirlas de manera más eficiente. Esta tarea se realizó manualmente revisando en el sitio web de GO cada anotación y su jerarquía, eligiendo las categorías superiores que nos fueran más informativas. Éste mismo proceso se realizó empleando el servidor web *Categorizer v.1.0* (http://ssbio.cau.ac.kr/software/categorizer/), el cual permite una clasificación de las anotaciones GO en grupos definidos calculando la distancia en la similitud de la semántica entre las anotaciones (Na et al., 2014). Pese a sus ventajas metodológicas, se decidió no emplear esta categorización de las anotaciones GO debido a que excluyó todas aquellas anotaciones asociadas a virus o a los hospederos de los virus, mismas que eran de interés para nuestro estudio. Además el resultado final de la clasificación de *Categorizer v1.0* fue muy similar al realizado manualmente.

2. Implicaciones evolutivas en la distribución de datos

La manera en que los datos obtenidos por los perfiles se agruparon en relación a los grupos virales fue no aleatoria. De forma general, se muestra una relación entre el origen biológico de las secuencias que conforman el perfil y el grupo viral en donde se encuentran secuencias homólogas (figura 15 y 18), lo que denota la estrecha relación virus-hospedero a través de posibles eventos de transferencia horizontal, fenómeno que muchas veces ha sido reportado (Schulz et al., 2017; Filée, 2014; Jeudy et al., 2012; Flint et al., 2009; Monier et al., 2009; Rohwer & Thurber, 2009; Filée et al., 2006; Lindell et al; 2004). En particular, los genes relacionados al metabolismo del RNA en virus han sido considerados como no propios de los virus, lo que permite interactuar con su hospedero y son propensos a ser transferidos (Krupovič & Bamford, 2008).

Por ejemplo, la gran mayoría de secuencias virales homólogas a familias de origen procarionte se distribuyeron en virus que infectan procariontes, como *Myoviridae, Siphoviridae, Podoviridae, Caudovirida*e o *Bicaudaviridae*. Mientras que la ausencia de homologías de virus de RNA y familias de dominios procariontes se explicaría por el hecho de que, salvo unas cuantas excepciones, no existen virus de RNA que infecten bacterias o arqueas (Campillo-Balderdas et al., 2015).

Algunas familias bacterianas hallaron similitudes en secuencias pertenecientes a retrovirus, lo cual no se explicaría por la interacción virus-hospedero. Las secuencias involucradas corresponden a retrotranscriptasas, por un lado virales y por el otro bacterianas. La relación entre estas proteínas ha sido mencionada en distintas ocasiones, y se ha explicado debido a que todos los retroelementos conocidos comparten un mismo ancestro en común (Rice & Lampson, 1995; Xiong & Eickbush, 1990; Inouye et al., 1989). La distribución actual de los retroelementos puede ser entendida por procesos de transferencia horizontal entre grupos taxonómicos superiores; los virus pudieron evolucionar por la captura de estos retroelementos, o tomarlos de otros virus existentes (Xiong & Eickbush, 1990).

Otras relaciones de homología encontradas entre familias de dominios bacterianos y grupos virales que infectan eucariontes se encontraron principalmente en virus gigantes como las familias *Poxviridae*, *Mimiviridae* y *Phycodnaviridae*. Estos resultados coinciden con los reportes realizados sobre la conformación del genoma de

los virus gigantes, los cuales contienen grupos de genes con diversos orígenes celulares, entre ellos bacterianos (Schulz et al., 2017; Filée et al., 2006). Por último, los *hits* en secuencias de herpesvirus, endornavirus, circovirus y baculovirus fueron casos aislados y con valores de expectación bajos, por lo que algunos de ellos podrían considerarse como posibles falsos positivos.

Respecto a las familias de dominios con secuencias eucariontes, estas tienen el mayor número de homologías en los proteomas virales. Estas se distribuyeron en una gran diversidad de virus que incluyen a todos los grupos de Baltimore. Como ocurrió con las familias de dominios bacterianos, los resultados obtenidos se concentraron mayoritariamente en virus que emplean a los eucariontes como hospederos, por ejemplo, retrovirus, poxvirus, mimivirus, phycodnavirus, baculovirus o iridovirus (figura 15). Sin embargo, pese a que los virus con genoma ssRNA presentan algunos de los grupos más diversos de todos los virus, y casi en su totalidad infectan eucariontes, fueron pocas las homologías encontradas en estos virus (Campillo-Balderas et al., 2015). Una manera de explicar esta escasez en los datos obtenidos son las restricciones presentes en los virus con genomas de RNA, como lo son sus genomas pequeños, inestabilidad del RNA, restricciones en el tamaño y conformación de la cápside, así como una alta tasa de mutación.

La presencia de homologías entre secuencias eucariontes y secuencias de bacteriófagos y virus de arqueas evidencia el alto grado de movilidad por transferencia horizontal de éste tipo de dominios y su posible importancia en la adaptación de los virus al entorno celular. Anteriormente, se han reportado distintos casos de transferencias y homologías entre estos dos grupos, demostrando que su existencia es un fenómeno recurrente (McGeoch & Bell, 2005; di Fagagna et al., 2003; Pedulla et al., 2003; Cermakian et al., 1996).

Por último, las 69 familias de dominios exclusivamente virales incluidas en la base de datos se reencontraron en la búsqueda realizada, donde en su mayoría las secuencias halladas son de virus con genoma ssRNA, así como retrovirus, adenovirus y herpesvirus.

En resumen, a partir de la distribución de los datos obtenidos de los proteomas virales se pueden extraer tres conclusiones importantes. En primer lugar, que la estrecha relación virus-hospedero, sin duda, modela la composición genómica de los virus. En segundo lugar, la presencia de dominios homólogos a los de sus hospederos,

así como a otros grupos biológicos distantes en los proteomas virales demuestra la composición en mosaico de los genomas virales, los cuales se modelan y evolucionan en gran parte por la transferencia horizontal de genes. Por último, los virus con genomas de RNA parecen ser los más divergentes y con menos número de homologías con los dominios de organismos celulares. Esto podría ser consecuencia de su alta tasa de mutación, lo que produciría una señal evolutiva débil en los análisis con estructura primaria, y una capacidad reducida de adicionar genes nuevos debido a las características de sus genomas.

3. Asignación de funciones

Durante el proceso de asignación de funciones a los dominios que constituyen el catálogo se presentaron una serie de dificultades. En primera instancia, las anotaciones originales asociadas a los proteomas presentan irregularidades e inconsistencias en relación con otras bases de datos, por lo que consideramos que no deben ser empleadas como referencia principal. De ser así, hubiésemos lidiado con un gran número de ausencias y errores de notación, lo que habría sesgado el resultado final de éste trabajo. Por otra parte, nos enfrentamos a la escasez en la información descriptiva de las secuencias virales en las bases de datos, las cuales son mayoritariamente asignadas de forma automatizada con base en la similitud de secuencias. Por tal razón, las asignaciones funcionales propuestas en éste trabajo se basaron, principalmente, en el consenso de la información disponible en UniProt, empleando en forma preferencial las descripciones y funciones de los dominios proteínicos allí descritos. Esta decisión se basó en las observaciones que sugieren que proteínas con RBD similares pueden tener especificidades similares o casi idénticas (Ray et al, 2003), y que generalmente las secuencias homólogas celulares capturadas por los virus interactúan de la misma manera que su contraparte celular (Nicholls & Gray, 2004).

Por último, la falta de una clasificación funcional en la que se considere a las proteínas virales con sus características y particularidades, como lo son la multifuncionalidad o la presencia recurrente de poliproteínas y péptidos precursores, nos condujo a sugerir y agregar nuevas funciones a las existentes en la clasificación COG (tabla 3) para así lograr una mejor descripción del catálogo. Pese a que las categorías funcionales aquí propuestas son perfectibles, consideramos que su empleo cumplió con el objetivo de generar una descripción general de cómo los virus podrían

estar empleando los RBD y hacer la diferenciación entre las funciones típicas de organismos celulares y las exclusivas de virus.

4. Recurrencias en las categorías funcionales

Una importante porción de las funciones asignadas se asociaron a unas cuantas categorías funcionales, mostrando a grandes rasgos como podrían estar siendo empleados los dominios en cuestión. Los distintos grupos de categorías serán discutidos a continuación.

Funciones pobremente caracterizadas. En esta categoría se agrupan las funciones generales [R] y desconocidas [S], las cuales suman una fracción considerable del total de funciones asignadas. En el caso de las funciones desconocidas [S], estás incluyen en su mayoría secuencias hipotéticas y con anotaciones nulas o muy escasas. Esta podría ser la primera ocasión en que estas secuencias son asociadas a una función en particular, por lo que les consideramos relevantes como objeto para futuras investigaciones que puedan confirmar o descartar los resultados aquí expuestos.

A su vez, las secuencias asociadas a funciones generales están relacionadas principalmente con procesos de unión a diversos sustratos, como a ácidos nucleicos, ATP o proteínas. Casi en su totalidad se trata de dominios proteínicos promiscuos y relacionados con una gran variedad de procesos, razón por la cual no fue posible clasificarlas en alguna otra función más específica. La presencia de estos dominios en virus ha sido reportada en distintas ocasiones (Shukla et al., 2018; Sonnberg et al., 2008; Levin et al., 2017; Fischer et al., 2010; Afonso et al., 2002), pero las funciones en las que pudieran estar implicados aún son discutidas (Shukla et al., 2018).

Las ankirinas, por ejemplo, han sido asociadas a regulación del ciclo celular, unión entre proteínas, tráfico de proteínas (Shukla et al., 2018), regulación de la transcripción o señalización celular (Miles et al., 2005); incluso algunos han sugerido que tiene capacidad de unirse a DNA (Michaely & Bennett, 1992), y se conoce también un grupo de estas proteínas asociadas al dominio KH de unión a RNA (Miles et al., 2005). Sobre sus funciones en virus, estas podrían estar involucradas en el secuestro celular (Shukla et al., 2018; Al-Khodor et al., 2010), prevención de la apoptosis inducida por infección (Afonso et al., 2002) o manipulación del sistema de ubiquitinación (Levin et al., 2017; Al-Khodor et al., 2010; Sonnberg et al., 2008). Otros dominios reportados en nuestros resultados como FNIP y F-box son conocidos

por su interacción con el sistema de ubiquitinación y por tanto con el proceso de degradación de proteínas (Levin et al., 2017; Fischer et al., 2010). Por su parte, se ha sugerido que el dominio WD40 participa en la regulación genética, modelado de la cromatina y modificación de mRNA, entre otras funciones (Shukla et al., 20018). Los dominios de unión a ATP también han sido asociados con unión a RNA (Anantharaman et al., 2002), como en el caso del factor de transcripción Ro de procariontes (Wei & Richardson, 2001).

Estos dominios junto a otros descritos con función de unión a nucleótidos o proteínas, como dedos de zinc o los LLR, son el conjunto que conforma el grupo de secuencias virales asociadas a funciones generales. Si tomamos en cuenta que *1)* se trata de dominios altamente promiscuos, *2)* que suelen presentarse juntos en una misma proteína y además que *3*) participan en diversas vías celulares debido a su flexibilidad funcional, entonces, éstas proteínas proporcionan un ejemplo acerca de la multifuncionalidad proteínica viral, permitiéndonos comprender la compleja red de interacciones con las que los virus podrían manipular al metabolismo del RNA.

Funciones metabólicas. Como ya se ha mencionado, la única categoría funcional COG dentro de las metabólicas fue *inorganic ion transport and metabolism* [P], ocupando el segundo lugar en el mayor número de ocurrencias dentro de todo el catálogo. Igualmente se hizo la observación de que en su totalidad las secuencias halladas pertenecen a la proteína phoH. La presencia de estas proteínas ha sido confirmada en diversas ocasiones en bacteriófagos, principalmente en fagos marinos (Goldsmith et al., 2011). Aunque su función no se conoce ni en bacterias ni virus, generalmente se reconoce como una ATPasa; algunos autores la han propuesto como una helicasa por presentarse asociada a RDB como el KH y PIN (Anantharaman et al, 2002), mientras que otros le han involucrado con el metabolismo de lípidos y modificación de RNA (Kazakov et al., 2003).

Un aspecto que define a los virus es la ausencia de metabolismo, por lo que no es de extrañar la poca diversidad de funciones y proteínas dentro de estas categorías. Aun así, los virus pueden llegan a tomar ciertos componentes clave de rutas metabólicas con las que pueden regular y manipular la maquinaria metabólica celular.

Procesos celulares y señalización. Éste grupo de categorías obtuvieron una menor frecuencia en las asignaciones. La función más destacada fue *Posttranscriptional modification, protein turnover, chaperones* [O], dentro de la cual las proteínas

asociadas a ubiquitinas fueron la mayoría. Estas proteínas regulan funciones y procesos celulares a través de modificaciones postraduccionales reversibles (Pelisch et al., 2013; Bellare et al., 2008), entre ellos la degradación de proteínas, tráfico de proteínas, transcripción, control del ciclo celular y señalización celular (Randow & Lehner, 2009). De igual forma, algunos autores les han asociado al metabolismo del RNA. La razón de esta asociación es porque pueden presentarse fusionadas a RBD (Anantharaman, et al., 2002) como en las ligasas de ubiquitina de unión a RNA (Hildebrandt et al., 2017) o algunos factores de *splicing* (Bellare et al., 2008), pero también porque se conocen algunos ejemplos donde el RNA funciona como regulador en el sistema de ubiquitina (Hildebrandt et al., 2017).

La presencia de estas proteínas en los virus ha sido reportada principalmente en herpesvirus y poxvirus, posiblemente adquirida de sus hospederos. Las ubiquitinas virales podrían ser empleadas para usar el sistema de ubiquitinación a su favor y así manipular el ciclo celular del hospedero, el tráfico de membranas, la reparación del DNA, regular apoptosis, así como para evadir el sistema inmune (Randow & Lehner, 2009). En nuestros resultados las proteínas homólogas de las ubiquitinas virales recuperadas fueron principalmente factores de *splicing* descritos por su capacidad de unión a RNA y con presencia de dominios tipo ubiquitina. Puede ser que las ubiquitinas virales se asocien a múltiples vías y procesos celulares, pero habría que considerar su posible participación también en regulación de la expresión genética a través de procesos de *splicing*.

Además de algunas posibles peptidasas, el otro tipo de proteína que se destacó en esta función [O], aunque en una mucho menor cantidad, fueron las proteínas con dominio J de baculovirus. Sus homólogos son proteínas eucariontes con dominio DnaJ, las cuales contienen un dominio RRM de unión a RNA. En general estas proteínas funcionan como chaperonas, relacionándose con procesos como el plegamiento, translocación, degradación y traducción de proteínas. Evidentemente la presencia de éste RBD les asocia a interacciones con el RNA, pero la función o mecanismos de estas interacciones son poco claras. Wang et al. (2002) ha propuesto que en los baculovirus su función pudiese ser estructural, pudiendo jugar algún papel en la transformación e inmortalización celular, pero también en la replicación del DNA.

Por último, queda mencionar que las otras categorías funcionales de éste grupo contuvieron escasas asignaciones, aunque algunas de ellas interesantes por su posible función dentro del ciclo replicativo viral. Algunas de estas proteínas fueron asociadas por contener dominios de unión a ATP o dominios RRM, otras posiblemente actúen como transportadores de mRNA. El grupo que presenta casos con mayor claridad funcional son los de la categoría [V], en donde se reunieron endonucleasas de restricción y metiltransferasas relacionadas a procesos de defensa ante procesos de infección.

Almacenamiento y procesamiento de la información. En el conjunto de categorías aquí reunidas se destacó la de replicación, recombinación y reparación [L], debido a la gran abundancia de retrotranscriptasas. Además, esta categoría incluyó entre otras proteínas, nucleasas, integrasas, helicasas y factores de replicación, casi en su totalidad con homología a familias de dominios eucariontes. Algunas de ellas, aunque muy posiblemente estén asociadas a DNA, fueron incluidos debido a la presencia de dominios como los de unión a ATP u otros generales de unión a nucleótidos que están presentes en las RBP.

Dentro de la categoría de modificación y procesamiento del RNA [A], se encuentran algunas nucleasas, nucleotidiltransferasas, helicasas de RNA, así como proteínas de regulación postranscripcional, pero las proteínas más abundantes fueron las RNA ligasas virales. Estas ligasas, comúnmente presentes en bacteriófagos, pueden estar relacionadas a reparación de RNA (RNA ligasas 1), específicamente reparando tRNA dañado por el mismo hospedero como mecanismo ante la infección viral, o en la edición de RNA (RNA ligasas 2) (Ho et al., 2004). Algunas de las otras proteínas mencionadas en esta categoría funcional posiblemente se encuentren actuando en modificaciones del RNA como adición de *capping* y poli(A), degradación o procesamientos en tRNA.

En relación a las proteínas asignadas a la función de traducción [J], estas fueron pocas pero muy relevantes debido a las funciones que realizan. Se detectaron factores de elongación o iniciación de la traducción, así como reguladores, como las proteínas CsrA que la inhiben (Liu et al., 1997). Estas proteínas, en general, presentan homología con proteínas celulares, mientras que hubo un grupo mayoritario de proteínas que aparentemente son exclusivas de virus. Una de estas proteínas es la 100k o *hexon assembly protein*, que inhibe la traducción del hospedero y promueve la

de sus propios mRNA, reclutando otras proteínas una vez unido al RNA (Xi et al., 2004). Otra de ellas es la proteína no estructural 3 de rotavirus, nsP3, por sus siglas en inglés, que actúa estimulando la traducción viral al unirse al 3' terminal de los mRNA no poliadenilados, facilitando la unión de otras proteínas del hospedero (Keryer-Bibens et al., 2009).

Respecto a las proteínas con función de transcripción [K], todas ellas actúan como reguladores, como factores de elongación, de anti-terminación, o activadores. Entre ellos se reportó la proteína Tat de retrovirus, que estimula la transcripción al unirse al mRNA viral en la región 5' terminal, actuando como factor de elongación para la RNA polimerasa II al estabilizarla (Gait & Karn, 1993).

Por último, las pocas proteínas asignadas con la función general de procesamiento de la información [X], fueron aquéllas cuyas descripciones fueron insuficientes para relacionarlas con procesos más específicos, pero que sin duda participan en alguna de las categorías del procesamiento de la información. Principalmente se trató de helicasas, nucleasas y metiltransferasas.

Procesos virales. Finalmente, con la inclusión de las cuatro categorías aquí propuestas para funciones virales se logró integrar dicha información a un panorama general de los procesos en que actúan las RBP virales, tomando en cuenta también el contexto celular.

Las proteínas con funciones difíciles de definir a causa de su multifuncionalidad se les agrupó en dos categorías distintas, procesos virales [VX] y poliproteínas virales [VY]. En la primera se incluyeron proteínas que pueden participar en diversas tareas a lo largo del ciclo replicativo viral, pero que además son proteínas exclusivas de virus, por ejemplo, las fosfoproteínas P que actúa durante la replicación, transcripción y en la inmuno evasión. La segunda categoría [VY] incluye poliproteínas, en su totalidad gag-pro-pol de retrovirus, en donde no fue claro identificar la ubicación o función del dominio particular encontrado por la búsqueda de perfiles. Aunque algunas de ellas pueden ser retrotranscriptasas, en otros casos la homología fue con diversas proteínas celulares de unión a RNA como activadores de la transcripción o proteínas con RBD como dedos de zinc o *cold-shock*, complicando una asignación más precisa.

Dentro de la categoría [Vc] de estructura de cápside o membrana, se enlistan proteínas que llevan a cabo funciones meramente estructurales o de encapsidación del

genoma. La fosfoproteína p10 del virus de la leucemia Murina de Rauscher (*Retroviridae*), por ejemplo, es descrita comúnmente como estructural de la cápside, pero su capacidad de unión a RNA o DNA de cadena simple ya ha sido descrita (Sen & Todaro, 1977). Otras proteínas estructurales son gag, las nucleoproteínas o las proteínas de la matriz. Cabe destacar que las nucleoproteínas N, típicas de virus con genoma ssRNA (-), pese a que comúnmente son asociadas a estructura o protección del genoma, en realidad cumplen diversas funciones que aún permanecen sin ser totalmente descritas. Esta capacidad multifuncional deriva de la presencia de dominios de unión a RNA y de unión a proteínas, lo que le permite participar en la transcripción y replicación, así como en el tráfico de la ribonucleoproteína que conforma junto al RNA genómico viral, durante la entrada y salida del núcleo celular (Portela & Digard, 2002).

La última categoría por mencionar, [Vp] de patogenicidad viral, está conformada por distintas proteínas que interfieren con la respuesta inmune del hospedero, evitando así la producción de interferón, de enzimas APOBEC o inhibiendo la apoptosis. Para llevar a cabo su objetivo actúan sobre los mRNA de dichas vías, produciendo su degradación. Por ejemplo, la proteína Vif (*virion infective factor*), cumple con la doble función de degradar las enzimas APOBEC mediante el reclutamiento de ligasas de ubiquitina, pero también interactúa con el mRNA de estas enzimas para evitar que se traduzcan (Davey et al., 2011).

5. Distribución de funciones en los grupos virales

La distribución de las funciones asignadas fue dispar entre los distintos grupos virales. A grandes rasgos se logró discernir dos grupos principales, los virus con genoma dsDNA, que presentan la mayor diversidad funcional, y los demás tipos virales que están restringidos a funciones esenciales, como lo son el procesamiento de la información genética y procesos virales. La diferencia entre estos dos grupos puede ser vinculada con las características intrínsecas de los tipos de genomas. Los genomas de dsDNA en virus son más estables, pueden llegar a ser de un gran tamaño, y por lo mismo pueden generar más transferencias de genes y mantenerlos. En contraste, los demás grupos virales con genomas de ssRNA, dsRNA y ssDNA, tienen genomas pequeños y menos estables, además de que presentan restricciones en el tamaño de las cápsides, debido a los mecanismos de ensamblaje y empaquetamiento (Flint et al.,

2009), así como por las distintas estrategias de entrada a las células de sus hospederos (Shukla et al., 2018), lo que evita la incorporación de genes adicionales.

Virus dsDNA. A pesar que las familias de la base de datos contienen a pocas familias, los virus de éste grupo fueron los que presentaron el mayor número de datos, de familias virales y diversidad de funciones. La mayoría de datos se obtuvieron por su homología con organismos celulares. Las categorías funcionales con mayor recurrencia fueron las pobremente caracterizadas, las de procesamiento de información, así como modificaciones postraduccionales [O] y metabolismo y transporte de iones inorgánicos [P]. Casi todos aquellos *hits* que no han sido asignados a una familia viral fueron virus dsDNA, en los cuales se destacan las mismas categorías funcionales que en todo el grupo I.

Los virus que infectan arqueas fueron muy escasos, reflejando el desconocimiento que se tiene de estos virus. Por otro lado, los bacteriófagos fueron uno de los grupos que obtuvieron mayor diversidad de funciones, así como un gran número de ocurrencias. La función más destacada fue [P], que comprende a proteínas phoH del regulón de fosfato, las cuales, como ya ha sido mencionado, son producto de la transferencia horizontal de genes. Asimismo, se ha demostrado que en viromas marinos, donde abundan los fagos, se encuentran una gran cantidad de genes metabólicos como genes de la respiración, del metabolismo de ácidos nucleicos, de carbohidratos, proteínas, entre otros (Rohwer & Thurber, 2009). Por su parte, las proteínas asociadas a funciones generales [R] presentaron descripciones muy vagas como unión a RNA, a ATP o nucleótidos, por lo que no fueron muy informativas.

Los virus gigantes, NCLDV, por sus siglas en inglés (*nucleocytoplasmic large DNA viruses*), fueron el otro grupo de virus con mayor cantidad de funciones y ocurrencias. En todo ellos se destacó la presencia de la función [R], que en éste caso la constituyen proteínas en repetición como las ankirinas, F-box, y otras que ya han sido descritas en el apartado anterior. La gran cantidad de proteínas con dominios en repetición presentes en los genomas de los NCLDV está correlacionada con el tamaño de su genoma, mismos que posiblemente han sido adquiridos de bacterias y sus hospederos eucariontes. Estas proteínas han sido relacionadas con muchas funciones debido a su capacidad de interactuar con distintas proteínas. También se ha sugerido que permiten la expansión de genoma y la adaptación al hospedero, debido a que pueden sufrir de ciclos de relajación y endurecimiento por presión purificadora, lo que

lleva a nuevas funciones y al aumento en la adecuación a su ambiente celular (Shukla et al., 2018).

La familia *Phycodnaviridae,* a diferencia de los otros virus gigantes, obtuvo abundantes asignaciones en la función [P]. La homología con secuencias bacterianas de la proteína phoH en phycodnavirus nos hace suponer que su presencia se debe a un proceso de transferencia horizontal de genes, lo cual no sería extraño en estos virus, de los que se conocen distintos casos de adquisición de genes bacterianos (Fileé et al., 2007). La cuestión radicaría en la función que podría estar desempeñando. La mayoría de bacteriófagos que contienen éste mismo gen son marinos; pudiera ser que estos genes son predominantes en phycodnavirus que comparten mismo ambiente, por lo que su función podría estar relacionada, como ya se ha mencionado, a la limitación de fosfato en ambientes marinos, actuando en la activación de toma de fosfato.

La otra constante en los virus gigantes es la función [O], que está conformada casi en su totalidad por ubiquitinas, y las funciones de procesamiento del RNA, que incluye ligasas, RNasas, helicasas de RNA, y enzimas de modificación del RNA, como transferasas. Sin embargo, la faltante fueron las funciones relacionadas a traducción [J]. Distintas proteínas del aparato de traducción han sido reportadas, como factores de liberación y aminoacil tRNA sintetasas (Schulz et al., 2017; Jeudy et al., 2012). El no haber encontrado genes relacionados a traducción puede deberse, por un lado a la falta de datos en la base de datos, y por otro a la falta de señal evolutiva a nivel de secuencia por parte de las proteínas celulares, aunque se ha reportado que los factores de liberación presentan homología con los de eucariontes y arqueas (Jeudy et al., 2012).

Por último, los distintos virus que infectan animales fueron los que menos datos presentaron, con la excepción de los herpesvirus que estaban presentes en la base de datos, y los baculovirus que presentaron la mayor diversidad funcional, en su mayoría halladas por homología con secuencias eucariontes.

En éste trabajo hemos podido asociar por primera vez, a las proteínas con función desconocida [S], con la capacidad de unión a RNA, pero estas mismas deben ser sometidas a estudios más meticulosos de tipo experimental para poder asociarlas a funciones más específicas o confirmar si en realidad son RBP.

De forma general se puede observar que los virus con genoma dsDNA tienen la capacidad de adquirir y mantener una mayor cantidad de genes relacionados con una gran diversidad de funciones. Esto posiblemente se deba a las características de su genoma, como la estabilidad, así como la similitud con el genoma de sus hospederos.

Virus ssDNA. Algunas de las principales características de estos virus es el tener tasas de mutación muy elevadas y genomas muy pequeños, que pueden llegar a codificar un par de proteínas, la de cápside y otra para la replicación. Además, infectan a una amplia gama de hospederos pertenecientes a los tres dominio celulares, comúnmente integrando su genoma al de su hospedero mediante integrasas o nucleasas (Kuprovic & Forterre, 2015). De las nueve familias que integran a éste grupo viral, en nuestros resultados únicamente obtuvimos unos pocos datos relacionados a dos familias, *Circoviridae* e *Inoviridae*. Todas las proteínas pertenecientes a estos virus fueron asociadas a funciones de replicación, recombinación y reparación [L]. Para el caso de los inovirus estas fueron integrasas de DNA, las cuales contienen dominios de nucleasas con semejanza a los dominios de RNasas. Por su parte, la única proteína de circovirus fue una replicasa, que a pesar de tener funciones primordialmente de unión a DNA, por el tipo de genoma con el que interactúa también presenta una actividad reducida como RNA ligasa (Stedman, 2013).

La escasez de datos encontrados podría estar relacionada con el hecho de que estos virus no han sido tan estudiados como otros (Stedman, 2013), y tal vez también por el tamaño tan pequeño de sus genomas y las pocas proteínas que estos contienen. Por tanto, podría tratarse del grupo viral con menor cantidad de RBP y de interacciones con el metabolismo del RNA.

Virus dsRNA. La diversidad funcional asociada a las proteínas de éste grupo viral se concentró en categorías del procesamiento de la información y procesos virales, sobre todo en funciones asocias a replicación, traducción, inmuno evasión y formación de cápside. Los datos recuperados pertenecen a tres de las ocho familias en que están clasificados estos virus: *Reoviridae*, *Endornaviridae* y *Cystoviridae*. La heterogeneidad funcional entre estas familias (figura 22) puede deberse a escasez de información en las bases de datos, pero también a su alta tasa de mutación, lo que causa que proteínas conservadas tengan poca similitud entre distintos géneros. Incluso

ha sido reportado que es complicado determinar funciones en proteínas de virus dsRNA por comparación de secuencias (Mertens, 2004), como aquí se ha hecho.

Aun con los pocos datos obtenidos para éste grupo viral, se puede decir que se cubrió, de forma general, el total de funciones en que pueden estar implicadas sus RBP. Estos virus contienen pocas proteínas, tanto por las características de su genomas, como por restricciones físicas de la cápside. El manejo de la información genética es realizado por enzimas virales dentro de la cápside para evitar la detección y acción de las defensas del hospedero. El otro conjunto de proteínas sobresalientes en los virus dsRNA, además de las estructurales, son aquellas que tienen como objetivo interferir con la activación del sistema inmune como la proteína nsP1 de rotavirus, que bloquea la producción de interferón mediante degradación de proteínas; aunque es ampliamente reconocida la capacidad de unión a RNA en esta proteína, particularmente a mRNA viral, su actividad aún permanece indefinida (Arnold & Patton, 2009).

Virus ssRNA, grupos IV y V. Estos grupos presentaron una distribución similar en las funciones, destacándose aquellas asociadas a cápside y procesos virales. Las pocas categorías funcionales asociadas reflejan el reducido número de proteínas codificadas por los virus de ssRNA. Además del genoma ssRNA, algo que relaciona a los dos grupos virales es la presencia de proteínas multifuncionales y péptidos precursores que les permiten superar el problema del bajo número proteínas que codifican. Estas proteínas interactúan y explotan diversos componentes del hospedero, con lo cual pueden evitar el sistema inmune, promover la replicación o regular la expresión genética, permitiéndoles desarrollar el ciclo replicativo viral (Chatterjee et al., 2016; Nagy & Pogany, 2012; Ahlquist et al., 2003). Otro mecanismo descrito en virus ssRNA para reclutar factores celulares es la presencia de elementos estructurales en el RNA, tanto genómico como mensajero (Gutiérrez-Escolano, 2014). Por tanto, la aparente escasez funcional de las RBP en virus de ssRNA, se debe al reducido número de proteínas que contienen, lo que solucionan reclutando proteínas celulares.

Para los virus de ssRNA (+), la gran mayoría de los resultados pertenecen a las familias *Coronaviridae* y *Togaviridae*. Esto se debe a que nuestra base de datos sólo contiene información de estas dos familias virales, de manera que las secuencias de las familias de dominios se hallaron a sí mismas en los proteomas virales. En nuestra

base de datos se tuvo una limitada diversidad de secuencias de éste grupo viral pese a que es uno de los más diversos, afectando de manera evidente el resultado obtenido.

La situación para los virus de ssRNA (-) fue un tanto distinta, ya que nuestra base de datos contenía una mayor diversidad de secuencias de este grupo de, aunque casi todo pertenecientes al Orden Mononegavirales. La distribución de las funciones fue homogénea, acumulándose en su mayoría en funciones relacionadas a cápside [Vc]; sólo las familias *Paramyxoviridae, Filoviridae* y *Pneumoviridae* tuvieron datos asociados a otras categorías como transcripción [K] y procesos virales [VX], lo cual coincide con las descripciones que califican a estas mismas familias virales como las de mayor complejidad entre los Mononegavirales por el número de proteínas que codifican (Chatterjee et al., 2016).

Un patrón interesante en los datos obtenidos en los virus de ssRNA (+), fue que todas las funciones asignadas en éste grupo viral distintas de [Vc] tienen homología con secuencias eucariontes, incluidas funciones [VX] de procesos virales. Estas secuencias celulares homólogas incluyen RNA helicasas que actúan en procesos de *splicing*, nucleasas con plegamiento *P-loop ATP binding* y dedos de zinc C3H1 asociados al decaimiento de mRNA, cuyas contrapartes virales se asociaron a helicasas, una replicasa y poliproteínas replicasa 1ab. Únicamente dos casos de proteínas de la nucleocápside tuvieron homología con secuencias celulares; una perteneciente a Astroviridae con un *E- value* cercano al umbral considerado como significativo y con homología a factores de *splicing* SRA1E con dominio RRM, y otra de Coronaviridae, igualmente relacionada con factores de *splicing* U2 con dominio RRM, pero con un valor de *E-value* significativo. De forma similar, en los virus de ssRNA (-) se presentaron grupos de proteínas con homología celular, como la de matriz M2-1, y otros aparentemente exclusivos de virus.

Nos parece excepcional haber logrado identificar relaciones de homologías entre éste tipo de proteínas virales y proteínas celulares, debido a la naturaleza cambiante de los virus ssRNA y las limitantes en la metodología aquí empleada. Por su parte, las proteínas exclusivas de virus, como las de cápside, pudieron haber perdido identidad a nivel de secuencia a causa de la alta tasa de mutación que los caracteriza; aunque el análisis de la estructura terciaria de estas mismas proteínas ha sugerido que su origen proviene del reclutamiento de proteínas del hospedero (Krupovič & Koonin, 2017), poniendo en duda la existencia de los llamados *hallmark* genes de virus.

Para poder comprender la evolución y la diversidad de estos grupos virales se requiere de un aumento en la exploración del contenido genético de las muchas familias virales que han sido poco estudiadas. De igual manera, generar estudios a nivel de estructura terciaria de estas y otras proteínas virales permitiría esclarecer puntos clave en el debate de la evolución y origen de los virus.

Virus con retrotranscripción. En el caso de los virus con retrotranscripción, tanto de ssRNA como de dsDNA, obtuvieron asignaciones relacionadas a funciones similares, procesos virales [VX], formación de cápside o membrana [Vc] y replicación, recombinación y reparación [L]. Únicamente en el caso de los retrovirus se encontró una mayor diversidad funcional, debido a la proteína Tat que actúa como factor de elongación para la polimerasa II durante la transcripción (Smith et al., 2000; Gait & Karn, 1993), y la proteína Vif involucrada en patogenicidad que media la degradación de proteínas mediante el reclutamiento de ligasas de ubiquitina del hospedero (Daey et al., 2011).

Se obtuvieron datos de todas las familias que conforman a estos dos grupos de virus. Es de destacar que en nuestra base de datos únicamente se contaba con secuencias pertenecientes a *Retroviridae* (retrovirus de ssRNA), y ninguna secuencia relacionada a retrovirus de dsDNA. Todos los resultados relacionados a las familias *Hepadnaviridae* y *Caulimoviridae* fueron halladas por similitud con secuencias celulares. De manera similar, la mayor cantidad de datos asociados con retrovirus fueron obtenidos por similitud con secuencias celulares, incluyendo proteínas como gag, retrotranscriptasas, proteínas de envoltura y otras poliproteínas.

Los virus dsDNA con retrotranscripción contienen genomas pequeños y codifican para pocas proteínas (Dill et al., 2016; Hull et al., 1987), las cuales son muy divergentes entre los distintos géneros, pero con dominios funcionales muy conservados (Dill et al., 2016). Tanto los hepadnavirus que infectan animales, como los caulimovirus que infectan plantas, se han encontrado insertados en los genomas de sus hospederos como elementos endógenos, pese a que la integración al genoma hospedero no es parte de ciclo replicativo. Esto ha demostrado una estrecha relación entre virus y hospedero mostrando procesos de coevolución (Diop et al., 2018; Dill et al., 2016).

A diferencia de estos últimos, los retrovirus requieren integrarse en el genoma hospedero para poder ser transcritos y llevar a cabo su ciclo replicativo, por lo que es

común hallarlos como elementos endógenos en los genomas de vertebrados. Al igual que los retrovirus de dsDNA, estos presentan una estrecha relación evolutiva con sus hospederos (Shi et al., 2018). Esta relación es tan estrecha que incluso están considerados como un tipo más de retroelementos, y se ha propuesto que los retrovirus provienen de retrotransposones de vertebrados que adquirieron la proteína env, lo que les permitió diseminarse entre distintas células (Hayward, 2017; Kim et al., 2004).

La reducida diversidad funcional en estos virus indudablemente está asociada a lo reducido de su genoma y al empleo de las proteínas del hospedero para realizar una parte importante de los procesos requeridos para su replicación. Las pocas proteínas incluidas en el catálogo que no presentan una homología celular podrían ser objeto de estudio para comprobar si realmente se trata de proteínas exclusivas virales. Un primer acercamiento sería la utilización de las familias de dominios de nuestra base de datos que contienen a estas proteínas y realizar una búsqueda con perfiles pero en los proteomas celulares.

6. Descripción general de las secuencias virales

En primer lugar, se observó que las secuencias homólogas virales presentan una reducción en el número de aminoácidos en comparación con su contrapartes celulares. Cuando se trató de familias de dominios virales las dimensiones se mantuvieron constantes. Esto coincide con la reducción de los genomas virales, en donde las proteínas multidominio llegan incluso a sobreponer dominios. Un ejemplo de esto es la proteína Rev que contiene un dominio de unión a proteínas y otro a RNA sobrepuestos, funcionando como un *switch* molecular (Davey et al., 2011). Aún así, no se descarta la posibilidad de que lo observado pueda presentar alguna otra explicación.

La descripción de la arquitectura de las proteínas virales puede dar pistas de cómo podrían funcionar o regularse. Hay que recordar que generalmente las RBP son multidominio, al menos en proteínas celulares, particularmente en las eucariontes. Contrario a lo que se esperaba, en virus dsDNA las proteínas con función *[R]*, obtuvieron resultados similares. Posiblemente las proteínas asociadas a unión a RNA o ATP, puedan contener un solo RBD, mientras que las multidominio estarían representadas por las proteínas con dominios en repeticiones. Otro caso no esperado

es el de las proteínas phoH. Por su parte, las proteínas de procesamiento de RNA aparecen como monodominio, lo que pudiera reflejar esa compactación general en los genomas virales.

Para los otros grupos virales, la mayoría de las funciones tiene muy pocos datos, por lo que no podemos deducir conclusiones generales. Pero las funciones [Vc] y [L] que aparentemente son multidominio, en realidad puede deberse a que muchas de estas secuencias pertenecen a poliproteínas virales. Pudiera ser que las proteínas [VX] por deberse a proteínas multifuncionales, sean de carácter multidominio. En general se esperaba observar patrones más definidos en relación con las funciones asignadas, lo cual tal vez sea posible si se analizan de una manera más detallada.

Para concluir, un resultado interesante que se desprende del análisis de estructura secundaria fue el elevado porcentaje de regiones sin una estructura definida, los cuales obtuvieron valores altos en el índice de predicción arrojado por Jpred4. Esto podría apoyar las observaciones acerca de lo común que son las regiones desordenadas en las RBP y en particular en los RBD, y aún más en virus. A manera de perspectiva, el realizar análisis de predicción de desorden en las secuencias encontradas permitiría hacer una descripción más precisa y completa de los datos aquí presentados.

VI. CONCLUSIÓN

El estudio de las RBP es fundamental para entender las distintas etapas del ciclo replicativo viral y la manera en que interactúa con los distintos componentes y vías celulares de sus hospederos. El avance en su conocimiento puede tener implicaciones en el entendimiento de los procesos evolutivos virales y en el desarrollo de estrategias que busquen combatir procesos infecciosos virales.

Para éste fin, el empleo de la genómica comparada en el estudio de los virus es un enfoque potente que permite descubrir características funcionales compartidas y deducir historias evolutivas (Mihara et al., 2016). En el presente trabajo, a pesar de las limitantes en el método empleado, se logró visualizar de manera general la composición funcional de las RBP virales a partir de la homología de dominios, así

como generar un catálogo curado de RBD virales que puede proveer de una fuente de consulta de la cual partir en el planteamiento de futuras investigaciones.

La subestimación de la diversidad de proteínas y dominios se evidenció en algunos casos, como fue el de la falta de funciones de traducción en virus gigantes, por lo que muy posiblemente otros grupos proteínicos virales con interacción con el metabolismo del RNA tampoco pudieron ser recuperados. Sin embargo, considerando que ésta es una primera aproximación al problema, podemos decir que el objetivo de describir, de forma general, la gran diversidad funcional en los virus, se cumplió. Los resultados reflejaron la naturaleza y particularidades de cada grupo viral, además de que se obtuvieron datos puntuales que permiten generar nuevas preguntas y proyectos a realizar.

El haber empleado una base de datos con una mayoría de secuencias celulares nos permitió establecer un enlace evolutivo entre los genes virales y sus contrapartes celulares, ayudando a determinar la función intrínseca de cada grupo de proteínas, así como a suponer cómo los virus interactúan con las vías celulares de sus hospederos.

Uno de los puntos más relevantes del trabajo fue confirmar la estrecha relación virus-hospedero, la cual se alimenta a través de una constante transferencia horizontal de genes. Una fracción mayoritaria de los datos obtenidos fueron obtenidos por homología con secuencias celulares, incluso algunas proteínas virales conservadas. Esto demuestra que la interacción viral con el metabolismo de RNA celular está mediado por proteínas con dominios adquiridos de sus hospederos, pero también que muchos procesos virales son mediados por proteínas que contienen estos mismos dominios, posiblemente porque es más fácil reusar que reinventar dominios funcionales.

Es importante mencionar la necesidad de explorar la diversidad de virus que infectan grupos de organismos poco comunes para poder completar los vacíos informacionales existentes. También se requiere de estudios que permitan comprender las interconexiones entre las distintas vías y sistemas que los virus emplean para el secuestro celular, por ejemplo el metabolismo del RNA y el sistema de ubiquitinación.

La información creciente sobre éstas proteínas, como su localización, regulación o ligandos, permitirá un entendimiento sistémico de su función y participación dentro de los procesos tanto celulares como virales. De igual forma, estudios exploratorios

como el que hemos presentado pueden contribuir a ubicar posibles RBP no descritas y con funciones inesperadas, contribuyendo en el robustecimiento del conocimiento de las RBP y su papel en la biología de los virus.

REFERENCIAS

Afonso, C. L., Tulman, E. R., Lu, Z., Zsak, L., Sandybaev, N. T., Kerembekova, U. Z., Zaitsev, V. L., Kutish, G. F. & Rock, D. L. (2002). The genome of camelpox virus. Virology, 295(1), 1-9.

Ahlquist, P., Noueiry, A. O., Lee, W. M., Kushner, D. B., & Dye, B. T. (2003). Host factors in positive-strand RNA virus genome replication. Journal of virology, 77(15), 8181-8186.

Al-Khodor, S., Price, C. T., Kalia, A., & Kwaik, Y. A. (2010). Functional diversity of ankyrin repeats in microbial proteins. Trends in microbiology, 18(3), 132-139.

Anantharaman, V., Koonin, E. V., & Aravind, L. (2002). Comparative genomics and evolution of proteins involved in RNA metabolism. Nucleic acids research, 30(7), 1427-1464.

Arnold, M. M., & Patton, J. T. (2009). Rotavirus antagonism of the innate immune response. Viruses, 1(3), 1035-1056.

Baltimore, D. (1971). Expression of animal virus genomes. Bacteriological reviews, 35(3), 235.

Bamford, D. H., Burnett, R. M., & Stuart, D. I. (2002). Evolution of viral structure. Theoretical population biology, 61(4), 461-470.

Ban, T., Zhu, J. K., Melcher, K., & Xu, H. E. (2015). Structural mechanisms of RNA recognition: sequence-specific and non-specific RNA-binding proteins and the Cas9-RNA-DNA complex. Cellular and molecular life sciences, 72(6), 1045-1058.

Bellare, P., Small, E. C., Huang, X., Wohlschlegel, J. A., Staley, J. P., & Sontheimer, E. J. (2008). A role for ubiquitin in the spliceosome assembly pathway. Nature structural & molecular biology, 15(5), 444.

Björklund, Å. K., Ekman, D., Light, S., Frey-Skött, J., & Elofsson, A. (2005). Domain rearrangements in protein evolution. Journal of molecular biology, 353(4), 911-923.

Blanco, C., Bayas, M., Yan, F., & Chen, I. A. (2018). Analysis of Evolutionarily Independent Protein-RNA Complexes Yields a Criterion to Evaluate the Relevance of Prebiotic Scenarios. Current Biology, 28(4), 526-537.

Borneman, A. R., Gianoulis, T. A., Zhang, Z. D., Yu, H., Rozowsky, J., Seringhaus, M. R., Wang, L. Y., Gerstein, M., & Snyder, M. (2007). Divergence of transcription factor binding sites across related yeast species. *Science*, *317*(5839), 815-819.

Boyle, J. F., & Holmes, K. V. (1986). RNA-binding proteins of bovine rotavirus. Journal of virology, 58(2), 561-568.

Bru, C., Courcelle, E., Carrère, S., Beausse, Y., Dalmar, S., & Kahn, D. (2005). The ProDom database of protein domain families: more emphasis on 3D. *Nucleic acids research*, *33*(suppl 1), D212-D215.

Burd, C. G., Matunis, E. L., & Dreyfuss, G. (1991). The multiple RNA-binding domains of the mRNA poly (A)-binding protein have different RNA-binding activities. Molecular and Cellular Biology, 11(7), 3419-3424.

Burd, C. G., & Dreyfuss, G. (1994). Conserved structures and diversity of functions of RNA-binding proteins. *Science-AAAS-Weekly Paper Edition*, *265*(5172), 615-620.

Campillo-Balderas, J. A., Lazcano, A., & Becerra, A. (2015). Viral genome size distribution does not correlate with the antiquity of the host lineages. Frontiers in Ecology and Evolution, 3, 143.

Cech, T. R. (2012). The RNA worlds in context. Cold Spring Harbor perspectives in biology, 4(7), a006742.

Cermakian, N., Ikeda, T. M., Cedergren, R., & Gray, M. W. (1996). Sequences homologous to yeast mitochondrial and bacteriophage T3 and T7 RNA polymerases are widespread throughout the eukaryotic lineage. Nucleic acids research, 24(4), 648-654.

Chatterjee, S., Basler, C. F., Amarasinghe, G. K., & Leung, D. W. (2016). Molecular mechanisms of innate immune inhibition by non-segmented negative-sense RNA viruses. Journal of molecular biology, 428(17), 3467-3482.

Chavali, P. L., Stojic, L., Meredith, L. W., Joseph, N., Nahorski, M. S., Sanford, T. J., Sweeney, R. T., Krishna, B. A., Hosmillo, M., Firth, A. E., Bayliss, R., Marcelis, C. L., Lindsay, S., Goodfellow, I., Woods, C. G., & Gergely, F. (2017). Neurodevelopmental protein Musashi 1 interacts with the Zika genome and promotes viral replication. Science, eaam9243.

Cook, K. B., Kazan, H., Zuberi, K., Morris, Q., & Hughes, T. R. (2010). RBPDB: a database of RNA-binding specificities. Nucleic acids research, 39(suppl_1), D301-D308.

Cuff, J. A., & Barton, G. J. (1999). Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. *Proteins: Structure, Function, and Bioinformatics*, *34*(4), 508-519

Davey, N. E., Travé, G., & Gibson, T. J. (2011). How viruses hijack cell regulation. Trends in biochemical sciences, 36(3), 159-169.

D'haeseleer, P. (2006). What are DNA sequence motifs?. *Nature biotechnology*, *24*(4), 423-425.

di Fagagna, F. D. A., Weller, G. R., Doherty, A. J., & Jackson, S. P. (2003). The Gam protein of bacteriophage Mu is an orthologue of eukaryotic Ku. EMBO reports, 4(1), 47-52.

Dill, J. A., Camus, A. C., Leary, J. H., Di Giallonardo, F., Holmes, E. C., & Ng, T. F. F. (2016). Distinct viral lineages from fish and amphibians reveal the complex evolutionary history of hepadnaviruses. Journal of virology, JVI-00832.

Diop, S. I., Geering, A. D., Alfama-Depauw, F., Loaec, M., Teycheney, P. Y., & Maumus, F. (2018). Tracheophyte genomes keep track of the deep evolution of the Caulimoviridae. Scientific reports, 8(1), 572.

Domingo, E., & Perales, C. (2014). Virus evolution. *eLS*.

Drozdetskiy, A., Cole, C., Procter, J., & Barton, G. J. (2015). JPred4: a protein secondary structure prediction server. *Nucleic acids research*, *43*(W1), W389-W394.

Dupressoir, A., Lavialle, C., & Heidmann, T. (2012). From ancestral infectious retroviruses to bona fide cellular genes: role of the captured syncytins in placentation. Placenta, 33(9), 663-671.

Eddy, S. (2003). HMMER User's Guide. Biological sequence analysis using profile hidden Markov models.

Filée, J. (2014). Multiple occurrences of giant virus core genes acquired by eukaryotic genomes: The visible part of the iceberg?. Virology, 466, 53-59.

Filée, J., Siguier, P., & Chandler, M. (2007). I am what I eat and I eat what I am: acquisition of bacterial genes by giant viruses. TRENDS in Genetics, 23(1), 10-15.

Fischer, M. G., Allen, M. J., Wilson, W. H., & Suttle, C. A. (2010). Giant virus with a remarkable complement of genes infects marine zooplankton. Proceedings of the National Academy of Sciences, 201007615.

Fisher, S. (2010). Are RNA viruses vestiges of an RNA world?. In *Darwinism, Philosophy, and Experimental Biology* (pp. 67-87). Springer Netherlands.

Flint, S. J., Enquist, L. W., Racaniello, V. R., Skalka, A. M., & Barnum, S. R. (2009). Principles of Virology: Vol. 1. Molecular Biology.

Forterre, P. (2006). The origin of viruses and their possible roles in major evolutionary transitions. *Virus research*, *117*(1), 5-16.

Fox, G. E. (2010). Origin and evolution of the ribosome. Cold Spring Harbor perspectives in biology, a003483.

Gait, M. J., & Karn, J. (1993). RNA recognition by the human immuno-deficiency virus Tat and Rev proteins. Trends in biochemical sciences, 18(7), 255-259.

Gerstberger, S., Hafner, M., & Tuschl, T. (2014). A census of human RNA-binding proteins. Nature Reviews Genetics, 15(12), 829.

Giudice, G., Sánchez-Cabo, F., Torroja, C., & Lara-Pezzi, E. (2016). ATtRACT—a database of RNA-binding proteins and associated motifs. Database, 2016.

Glisovic, T., Bachorik, J. L., Yong, J., & Dreyfuss, G. (2008). RNA-binding proteins and post-transcriptional gene regulation. *FEBS letters*, *582*(14), 1977-1986.

Goldsmith, D. B., Crosti, G., Dwivedi, B., McDaniel, L. D., Varsani, A., Suttle, C. A., Weinbauer, M. G., Sanda, R. & Breitbart, M. (2011). Pho Regulon Genes in Phage: Development of phoH as a Novel Signature Gene for Assessing Marine Phage Diversity. Applied and environmental microbiology, AEM-05531.

Gutiérrez-Escolano, A. L. (2014). Host-cell factors involved in the calicivirus replicative cycle. Future Virology, 9(2), 147-160.

Hayward, A. (2017). Origin of the retroviruses: when, where, and how?. Current opinion in virology, 25, 23-27.

Hennig, J., Gebauer, F., & Sattler, M. (2014). Breaking the protein-RNA recognition code. Cell Cycle, 13(23), 3619-3620.

Hildebrandt, A., Alanis-Lobato, G., Voigt, A., Zarnack, K., Andrade-Navarro, M. A., Beli, P., & König, J. (2017). Interaction profiling of RNA-binding ubiquitin ligases reveals a link between posttranscriptional regulation and the ubiquitin system. Scientific reports, 7(1), 16582.

Ho, C. K., Wang, L. K., Lima, C. D., & Shuman, S. (2004). Structure and mechanism of RNA ligase. Structure, 12(2), 327-339.

Hoffman, M. M., Khrapov, M. A., Cox, J. C., Yao, J., Tong, L., & Ellington, A. D. (2004). AANT: the amino acid–nucleotide interaction database. *Nucleic acids research*, *32*(suppl 1), D174-D181.

Holmes, E. C. (2009). *The evolution and emergence of RNA viruses*. Oxford University Press.

Holmqvist, E., & Vogel, J. (2018). RNA-binding proteins in bacteria. Nature Reviews Microbiology, 1.

Hull, R., Covey, S. N., & Maule, A. J. (1987). Structure and replication of caulimovirus genomes. J Cell Sci, 1987(Supplement 7), 213-229.

Hunter, P. (2017). Viral taxonomy: The effect of metagenomics on understanding the diversity and evolution of viruses. EMBO reports, e201744982.

Inouye, S., Hsu, M. Y., Eagle, S., & Inouye, M. (1989). Reverse transcriptase associated with the biosynthesis of the branched RNA-linked msDNA in Myxococcus xanthus. Cell, 56(4), 709-717.

Jeong, E., Kim, H., Lee, S. W., & Han, K. (2003). Discovering the interaction propensities of amino acids and nucleotides from protein-RNA complexes. Molecules and cells, 16(2), 161-167.

Jeudy, S., Abergel, C., Claverie, J. M., & Legendre, M. (2012). Translation in giant viruses: a unique mixture of bacterial and eukaryotic termination schemes. PLoS genetics, 8(12), e1003122.

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution*, *30*(4), 772-780.

Kazakov, A. E., Vassieva, O., Gelfand, M. S., Osterman, A., & Overbeek, R. (2003). Bioinformatics classification and functional analysis of PhoH homologs. In silico biology, 3(1, 2), 3-15.

Kelley, L. A., & Sternberg, M. J. (2015). Partial protein domains: evolutionary insights and bioinformatics challenges. *Genome biology*, *16*(1), 100.

Keryer-Bibens, C., Legagneux, V., Namanda-Vanderbeken, A., Cosson, B., Paillard, L., Poncet, D., & Osborne, H. B. (2009). The rotaviral NSP3 protein stimulates translation of polyadenylated target mRNAs independently of its RNA-binding domain. Biochemical and biophysical research communications, 390(2), 302-306.

Kim, F. J., Battini, J. L., Manel, N., & Sitbon, M. (2004). Emergence of vertebrate retroviruses and envelope capture. Virology, 318(1), 183-191.

Krupovič, M., & Bamford, D. H. (2008). Virus evolution: how far does the double β-barrel viral lineage extend? Nature Reviews Microbiology, 6(12), 941.

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

Krupovič, M., & Koonin, E. V. (2017). Multiple origins of viral capsid proteins from cellular ancestors. Proceedings of the National Academy of Sciences, 114(12), E2401-E2410.

Levin, R. A., Voolstra, C. R., Weynberg, K. D., & van Oppen, M. J. H. (2017). Evidence for a role of viruses in the thermal sensitivity of coral photosymbionts. The ISME journal, 11(3), 808.

Li, W., Jaroszewski, L., & Godzik, A. (2001). Clustering of highly homologous sequences to reduce the size of large protein databases. Bioinformatics, 17(3), 282-283.

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

Li, Z., & Nagy, P. D. (2011). Diverse roles of host RNA binding proteins in RNA virus replication. RNA biology, 8(2), 305-315.

Lindell, D., Sullivan, M. B., Johnson, Z. I., Tolonen, A. C., Rohwer, F., & Chisholm, S. W. (2004). Transfer of photosynthesis genes to and from Prochlorococcus viruses. Proceedings of the National Academy of Sciences, 101(30), 11013-11018.

Liu, M. Y., Gui, G., Wei, B., Preston, J. F., Oakford, L., Yüksel, Ü., Giedroc, D. P. & Romeo, T. (1997). The RNA molecule CsrB binds to the global regulatory protein CsrA and antagonizes its activity in Escherichia coli. Journal of Biological Chemistry, 272(28), 17502-17510.

Lloyd, R. E. (2015). Nuclear proteins hijacked by mammalian cytoplasmic plus strand RNA viruses. *Virology, 479*, 457-474.

Luscombe, N. M., Laskowski, R. A., & Thornton, J. M. (2001). Amino acid–base interactions: a three-dimensional analysis of protein–DNA interactions at an atomic level. Nucleic acids research, 29(13), 2860-2874.

Maris, C., Dominguez, C., & Allain, F. H. T. (2005). The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression. *Febs Journal*, *272*(9), 2118-2131.

Matia-González, A. M., Laing, E. E., & Gerber, A. P. (2015). Conserved mRNA-binding proteomes in eukaryotic organisms. Nature structural & molecular biology, 22(12), 1027.

McGeoch, A. T., & Bell, S. D. (2005). Eukaryotic/archaeal primase and MCM proteins encoded in a bacteriophage genome. Cell, 120(2), 167-168.

Mertens, P. (2004). The dsRNA viruses. Virus research, 101(1), 3-13.

Michaely, P., & Bennett, V. (1992). The ANK repeat: a ubiquitous motif involved in macromolecular recognition. Trends in cell biology, 2(5), 127-129.

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

Miles, M. C., Janket, M. L., Wheeler, E. D., Chattopadhyay, A., Majumder, B., DeRicco, J., Schafer, E. A. & Ayyavoo, V. (2005). Molecular and functional characterization of a novel splice variant of ANKHD1 that lacks the KH domain and its role in cell survival and apoptosisc. The FEBS journal, 272(16), 4091-4102.

Monier, A., Pagarete, A., de Vargas, C., Allen, M. J., Claverie, J. M., & Ogata, H. (2009). Horizontal gene transfer of an entire metabolic pathway between a eukaryotic alga and its DNA virus. Genome research.

Na, D., Son, H., & Gsponer, J. (2014). Categorizer: a tool to categorize genes into user-defined biological groups based on semantic similarity. BMC genomics, 15(1), 1091.

Nagy, P. D., & Pogany, J. (2012). The dependence of viral RNA replication on co-opted host factors. Nature Reviews Microbiology, 10(2), 137.

Nicholls, R., & Gray, T. (2004). Cellular source of the poxviral N1R/p28 gene family. Virus genes, 29(3), 359-364.

O'Malley, M. A. (2016). The ecological virus. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *59*, 71-79.

Parte, A.C. (2018). LPSN – List of Prokaryotic names with Standing in Nomenclature (bacterio.net), 20 years on. International Journal of Systematic and Evolutionary Microbiology, 68, 1825-1829; doi: 10.1099/ijsem.0.002786

Pearson, W. R. (2013). An introduction to sequence similarity ("homology") searching. *Current protocols in bioinformatics*, 3-1.

Pedulla, M. L., Ford, M. E., Houtz, J. M., Karthikeyan, T., Wadsworth, C., Lewis, J. A., ... & Hatfull, G. F. (2003). Origins of highly mosaic mycobacteriophage genomes. Cell, 113(2), 171-182.

Pelisch, F., Risso, G., & Srebrow, A. (2012). RNA metabolism and ubiquitin/ubiquitin-like modifications collide. Briefings in functional genomics, 12(1), 66-71.

Portela, A., & Digard, P. (2002). The influenza virus nucleoprotein: a multifunctional RNA-binding protein pivotal to virus replication. Journal of general virology, 83(4), 723-734.

Randow, F., & Lehner, P. J. (2009). Viral avoidance and exploitation of the ubiquitin system. Nature cell biology, 11(5), 527.

Ray, D., Kazan, H., Cook, K. B., Weirauch, M. T., Najafabadi, H. S., Li, X., ... & Hughes, T. R. (2013). A compendium of RNA-binding motifs for decoding gene regulation. Nature, 499(7457), 172.

Ren, B., Robert, F., Wyrick, J. J., Aparicio, O., Jennings, E. G., Simon, I., Zeitlinger, J., Schreiber, J., Hannet, N., Kanin, E., Volkert, T. L., Wilson, C. J., Bell, S. P., & Young, R. A. (2000). Genome-wide location and function of DNA binding proteins. *Science*, *290*(5500), 2306-2309.

Rice, S. A., & Lampson, B. C. (1995). Bacterial reverse transcriptase and msDNA. Virus genes, 11(2-3), 95-104.

Rohwer, F., & Thurber, R. V. (2009). Viruses manipulate the marine environment. Nature, 459(7244), 207.

Santos-Beneit, F. (2015). The Pho regulon: a huge regulatory network in bacteria. Frontiers in microbiology, 6, 402.

Schulz, F., Yutin, N., Ivanova, N. N., Ortega, D. R., Lee, T. K., Vierheilig, J., Daims, H., Horn, M., Wagner, M., Jensen, G. J., Kyrpides, N. C., Koonin, E. V., & Woyke, T. (2017). Giant viruses with an expanded complement of translation system components. Science, 356(6333), 82-85.

Sen, A., & Todaro, G. J. (1977). The genome-associated, specific RNA binding proteins of an avian and mammalian type C viruses. *Cell*, 10(1), 91-99.

Shi, M., Lin, X. D., Chen, X., Tian, J. H., Chen, L. J., Li, K., ... & Holmes, E. C. (2018). The evolutionary history of vertebrate RNA viruses. Nature, 556(7700), 197.

Shukla, A., Chatterjee, A., & Kondabagil, K. (2018). The number of genes encoding repeat domain-containing proteins positively correlates with genome size in amoebal giant viruses. Virus evolution, 4(1), vex039.

Sigrist, C. J., Cerutti, L., Hulo, N., Gattiker, A., Falquet, L., Pagni, M., Bairoch, A. & Bucher, P. (2002). PROSITE: a documented database using patterns and profiles as motif descriptors. *Briefings in bioinformatics*, *3*(3), 265-274.

Smith, C. A., Calabro, V., & Frankel, A. D. (2000). An RNA-binding chameleon. Molecular cell, 6(5), 1067-1076.

Sonnberg, S., Seet, B. T., Pawson, T., Fleming, S. B., & Mercer, A. A. (2008). Poxvirus ankyrin repeat proteins are a unique class of F-box proteins that associate with cellular SCF1 ubiquitin ligase complexes. Proceedings of the National Academy of Sciences, 105(31), 10955-10960.

Stedman, K. (2013). Mechanisms for RNA capture by ssDNA viruses: grand theft RNA. Journal of molecular evolution, 76(6), 359-364.

Steitz, J., Borah, S., Cazalla, D., Fok, V., Lytle, R., Mitton-Fry, R., Riley, K., & Samji, T. (2010). Noncoding RNPs of viral origin. Cold Spring Harbor perspectives in biology, a005165.

Stormo, G. D. (2000). DNA binding sites: representation and discovery. *Bioinformatics*, *16*(1), 16-23.

Traut, T.(2014). Multidomain Proteins. eLS.

Treger, M., & Westhof, E. (2001). Statistical analysis of atomic contacts at RNA–protein interfaces. Journal of Molecular Recognition, 14(4), 199-214.

Varadi, M., Zsolyomi, F., Guharoy, M., & Tompa, P. (2015). Functional advantages of conserved intrinsic disorder in RNA-binding proteins. PloS one, 10(10), e0139731.

Viktorovskaya, O. V., Greco, T. M., Cristea, I. M., & Thompson, S. R. (2016). Identification of RNA binding proteins associated with dengue virus RNA in infected cells reveals temporally distinct host factor requirements. PLoS neglected tropical diseases, 10(8), e0004921.

Walia, R. R., Xue, L. C., Wilkins, K., El-Manzalawy, Y., Dobbs, D., & Honavar, V. (2014). RNABindRPlus: a predictor that combines machine learning and sequence homology-based methods to improve the reliability of predicted RNA-binding residues in proteins. PLoS One, 9(5), e97725.

Wang, C., Uversky, V. N., & Kurgan, L. (2016). Disordered nucleiome: Abundance of intrinsic disorder in the DNA-and RNA-binding proteins in 1121 species from Eukaryota, Bacteria and Archaea. *Proteomics*, *16*(10), 1486-1498.

Wang, L., Yu, J., Yin, C., Li, Z., Hu, X., & Pang, Y. (2002). Characterization of a J domain gene of Spodoptera litura multicapsid nucleopolyhedrovirus. Virus Genes, 25(3), 291-297.

Wang, P., Xu, J., Wang, Y., & Cao, X. (2017). An interferon-independent lncRNA promotes viral replication by modulating cellular metabolism. Science, 358(6366), 1051-1055.

Wei, R. R., & Richardson, J. P. (2001). Identification of an RNA-binding Site in the ATP binding domain of Escherichia coli Rho by H2O2/Fe-EDTA cleavage protection studies. Journal of Biological Chemistry, 276(30), 28380-28387.

Xi, Q., Cuesta, R., & Schneider, R. J. (2004). Tethering of eIF4G to adenoviral mRNAs by viral 100k protein drives ribosome shunting. Genes & development, 18(16), 1997-2009.

Xiong, Y., & Eickbush, T. H. (1990). Origin and evolution of retroelements based upon their reverse transcriptase sequences. The EMBO journal, 9(10), 3353-3362.

APÉNDICE

Apéndice 1.1. Número de ocurrencias por Phyla y Clases, pertenecientes a las familias proteicas de unión a RNA que conforman la base de datos final. Éstas se encuentran separadas por Dominio celular. Los * presentes en la sección Eukaryota indican que ese Phylum fue agregado manualmente por no haber sido proporcionado por UniProt.

Phylum	# Ocurrencias	Clase	# Ocurrencias
Archaea			
Crenarchaeota	4	Thermoprotei	4
Euryarchaeota	8	Archaeoglobi	4
		Halobacteria	2
		Methanobacteria	2
		Methanococci	4
		Methanomicrobia	5
		Methanopyri	2
		Thermoplasmata	2
		Thermococci	3
Nanoarchaeota	1		
Candidatus Korarchaeota	1		
Thaumarchaeota	3		
Bacteria			
Actinobacteria	20	Acidimicrobiia	5
		Actinobacteria	19
		Coriobacteriia	5
		Rubrobacteria	4
		Thermoleophilia	5
Acidobacteria	12	Acidobacteriia	10
		Blastocatellia	4
		Solibacteres	6
Proteobacteria	42	Acidithiobacillia	9
		Alphaproteobacteria	22
		Betaproteobacteria	21
		Deltaproteobacteria	21
		Epsilonproteobacteria	8
		Gammaproteobacteria	31
Chloroflexi	9	Anaerolineae	1
		Chloroflexia	6
		Dehalococcoidia	1
		Thermomicrobia	5
Aquificae	5	Aquificae	5
Firmicutes	34	Bacilli	24
		Clostridia	22
		Erysipelotrichia	4
		Negativicutes	8
		Tissierellia	6
Bacteroidetes	22	Bacteroidia	15
		Chitinophagia	11
		Cytophagia	15
		Flavobacteriia	12
		Saprospiria	9
		Sphingobacteriia	9
Chlamydiae	5	Chlamydiia	5
Chlorobi	6	Chlorobia	6
Chrysiogenetes	4	Chrysiogenetes	4
Deferribacteres	5	Deferribacteres	5
Deinococcus-Thermus	7	Deinococci	7
Dictyoglomi	4	Dictyoglomia	4
Elusimicrobia	4	Elusimicrobia	4
Fibrobacteres	4	Fibrobacteria	4
Fusobacteria	5	Fusobacteriia	5

Gemmatimonadetes	6	Gemmatimonadetes	6
Cyanobacteria	13	Gloeobacteria	4
Nitrospirae	4	Nitrospira	4
Planctomycetes	11	Planctomycetia	11
Spirochaetes	14	Spirochaetia	14
Synergistetes	5	Synergistia	5
Tenericutes	7	Mollicutes	7
Thermodesulfobacteria	5	Thermodesulfobacteria	5
Thermotogae	5	Thermotogae	7
Verrucomicrobia	7	Methylacidiphilae	3
		Opitutae	6
		Verrucomicrobiae	1
Candidatus Cloacimonetes	4		
Eukaryota			
Amoebozoa*	-	Archamoebae	3
		Dictyostelia	14
Apicomplexa	40	Aconoidasida	34
		Coccidia	18
Arthropoda	100	Arachnida	26
		Branchiopoda	29
		Insecta	88
		Malacostraca	2
Ascomycota	114	Dothideomycetes	43
		Eurotiomycetes	50
		Leotiomycetes	34
		Pezizomycetes	21
		Saccharomycetes	65
		Schizosaccharomycetes	28
		Sordariomycetes	53
Bacillariophyta	20	Bacillariophyceae	14
		Coscinodiscophyceae	14
Basidiomycota	37	Agaricomycetes	24
		Malasseziomycetes	2
		Microbotryomycetes	7
		Pucciniomycetes	12
		Tremellomycetes	25
		Ustilaginomycetes	16
Chlorophyta	42	Chlorophyceae	20
		Mamiellophyceae	28
		Trebouxiophyceae	22
Chordata	79	Actinopteri	41
		Amphibia	34
		Appendicularia	28
		Ascidiacea	7
		Aves	18
		Mammalia	58
Chytridiomycota	16	Chytridiomycetes	16
Cnidaria	23	Anthozoa	22
		Hydrozoa	2
Ctenophora	1	Tentaculata	1
Echinodermata	1	Asteroidea	1
		Echinoidea	1
Euglenozoa	14	Kinetoplastea	14
Hemichordata	1	Enteropneusta	1
Metamonada	4	Parabasalia	2
Microsporidia	9	-	-
Mollusca	7	Bivalvia	5
		Gastropoda	2
Nematoda	65	Chromadorea	64

Perkinsozoa*	-	Perkinsea	8
Phaeophyceae	24	Phaeophyceae	1
Placozoa	23	Heterolobosea	12
Platyhelminthes	39	Trematoda	38
Porifera	2	Demospongiae	2
Rotifera	2	Bdelloidea	2
Streptophyta	94	Bryopsida	26
		Liliopsida	69
		Magnoliopsida	4
-	-	Choanoflagellatea	6
-	-	Oomycetes	25
Bigyra*	-	Blastocystea	2
Ciliohora*	-	Oligohymenophorea	20
Ochrophyta*	-	Pelagophyceae	8

Apéndice 1.2. Número de ocurrencias halladas en la base de datos final divididas por tipo viral y familia viral.

Tipo viral	**# Ocurrencias**	**Famiia viral**	**# Ocurrencias**
virus			
Retro-transcribing viruses	10	Retroviridae	10
ssRNA negative-strand viruses	43	Arenaviridae	4
		Bunyaviridae	4
		Pneumoviridae	3
		Orthomyxoviridae	12
		Filoviridae	13
		Paramyxoviridae	7
dsRNA viruses	6	Reoviridae	4
		Endornaviridae	1
ssRNA positive-strand viruses, no DNA stage	3	Togaviridae	3
		Coronaviridae	3
dsDNA viruses, no RNA stage	19	Siphoviridae	1
		Herpesviridae	12
		Myoviridae	2
		Poxviridae	3
		Podoviridae	2

Apéndice 2. Número de hits encontrados en los proteomas virales asociados a las distintas familias taxonómicas de virus. Cuando no había una familia viral asociada se empleó el Orden viral.

dsDNA	
Adenoviridae	61
Ampullaviridae	1
Alloherpesviridae	6
Ascoviridae	4
Baculoviridae	172
Bicaudaviridae	2
Caudovirales (Orden)	23
Fuselloviridae	4
Herpesvirales (Orden)	3
Herpesviridae	165
Hytrosaviridae	2
Iridoviridae	58
Lipothrixviridae	4
Malacoherpesviridae	3
Marseilleviridae	26
Mimiviridae	395
Myoviridae	741
Nudiviridae	1
Phycodnaviridae	157
Podoviridae	61
Poxviridae	381
Siphoviridae	201
ssDNA	
Inoviridae	2
Circoviridae	1
dsRNA	
Endornaviridae	4
Reoviridae	9
Cystoviridae	2
ssRNA (+)	
Arteriviridae	1
Astroviridae	1
Caliciviridae	2
Coronaviridae	94
Roniviridae	1
Togaviridae	3
Virgaviridae	1
ssRNA (–)	
Arenaviridae	127
Filoviridae	93
Nairoviridae	8
Orthomyxoviridae	103
Paramyxoviridae	193
Pneumoviridae	25
Sunviridae	1
Retrovirus dsDNA	
Hepadnaviridae	11
Caulimoviridae	388
Retrovirus ssRNA	
Retroviridae	315

Apéndice 3.1. Número de ocurrencias de hospederos celulares que son infectados por las variedades virales que conforman el catálogo de dominios de unión a RNA. Se muestran los grupos taxonómicos superiores pertenecientes a Archaea y Bacteria.

Archaea	
Crenarchaeota	11
Euryarchaeota	17
Bacteria	
Actinobacteria	82
Bacterioidetes	3
Cyanobacteria	204
Deinococcus-Thermus	2
Firmicutes	188
Protobacteria	574
Tenericutes	2

Apéndice 3.2. Número de ocurrencias de hospederos celulares que son infectados por las variedades virales que conforman el catálogo de dominios de unión a RNA. Se muestran los grupos taxonómicos superiores pertenecientes al dominio Eukaryota. Los grupos no pertenecen a categorías taxonómicas equivalentes.

Eukaryota	
Fungi (Ascomycota)	3
Metazoa	1839
Viridiplantae	526
Protozoos	
Bicosoecida	17
Longamoebia	441
Algas	
Isochrysidales	5
Pelagophyceae	17
Phaeocystales	11
Phaeophyceae	14

Apéndice 3.3. Número de ocurrencias de hospederos celulares que son infectados por las variedades virales que conforman el catálogo de dominios de unión a RNA. Se muestran los grupos taxonómicos superiores que conforman a Metazoa y Viridiplantae incluidos en la tabla anterior.

Metazoa	
Amphibia	4
Actinopterygii	44
Aves	373
Mollusca	6
Insecta	269
Lepidosauria	20
Malacostraca	4
Mammalia	1073
Viridiplantae	
Streptophyta	
Asteridae	102
Caryophyllales	19
Liliopsida	125
Piperales	8
Rosidae	121
Saxifragales	11
Chlorophyta	
Chlamydomonadales	1
Chlorellales	57
Mamiellales	71