



**INSTITUTO POTOSINO DE INVESTIGACIÓN
CIENTÍFICA Y TECNOLÓGICA, A.C.**

**POSGRADO EN CIENCIAS EN BIOLOGÍA
MOLECULAR**

**ANÁLISIS FILOGENÉTICO-FUNCIONAL DE REGIONES
DE CONTROL REPLICATIVO Y TRANSCRIPCIONAL EN
VIRUS DE DNA DE CADENA SENCILLA**

Tesis que presenta
María Aurora Londoño Avendaño

Para obtener el grado de
Doctor en Ciencias en Biología Molecular

Directores de la tesis
Dr. Gerardo Argüello Astorga Dra. Lina Riego Ruiz

San Luis Potosí, S.L.P., marzo de 2010



CONSTANCIA DE APROBACIÓN DE LA TESIS

La tesis “Análisis filogenético-funcional de regiones de control replicativo y transcripcional en virus de DNA de cadena sencilla” presentada para obtener el Grado de de Doctor en Ciencias en Biología Molecular fue elaborada por **María Aurora Londoño Avendaño** y aprobada el **19 de marzo de 2010** por los suscritos, designados por el Colegio de Profesores de la División de Biología Molecular del Instituto Potosino de Investigación Científica y Tecnológica, A.C.

Dr. Gerardo Rafael Arguéllo Astorga
(Director de la tesis)

Dra. Lina Raquel Riego Ruiz
(Director de la tesis)

Dra. Irene Beatriz Castaño Navarro
(Asesor de la tesis)

Dra. Laura Silva Rosales
(Sinodal externo)



CRÉDITOS INSTITUCIONALES

Esta tesis fue elaborada en el Laboratorio de Biología Molecular de Plantas de la División de Biología Molecular del Instituto Potosino de Investigación Científica y Tecnológica, A.C., bajo la codirección de los doctores Gerardo Arguello Astorga y Lina Riego Ruiz.

El trabajo de investigación se realizó con apoyo financiero del Consejo Nacional de Ciencia y Tecnología, a través de los proyectos con clave SEP-2003-42639-Q y SEP-CONACYT-2005-49039.

Durante la realización del trabajo la autora recibió una beca académica a partir de recursos propios del Instituto Potosino de Investigación Científica y Tecnológica, A.C, apoyo económico por participar en los proyectos SEP-2003-42639-Q y SEP-CONACYT-2005-49039, y además una beca académica para estudios de doctorado del Consejo Nacional de Ciencia y Tecnología (No. de registro 211758).



Instituto Potosino de Investigación Científica y Tecnológica, A.C.

Acta de Examen de Grado

El Secretario Académico del Instituto Potosino de Investigación Científica y Tecnológica, A.C., certifica que en el Acta 026 del Libro Primero de Actas de Exámenes de Grado del Programa de Doctorado en Ciencias en Biología Molecular está asentado lo siguiente:

En la ciudad de San Luis Potosí a los 19 días del mes de marzo del año 2010, se reunió a las 12:00 horas en las instalaciones del Instituto Potosino de Investigación Científica y Tecnológica, A.C., el Jurado integrado por:

Dra. Irene Beatriz Castaño Navarro	Presidenta	IPICYT
Dra. Lina Raquel Riego Ruiz	Secretaria	IPICYT
Dr. Gerardo Rafael Argüello Astorga	Sinodal	IPICYT
Dra. Laura Silva Rosales	Sinodal externo	CINVESTAV

a fin de efectuar el examen, que para obtener el Grado de:

DOCTORA EN CIENCIAS EN BIOLOGÍA MOLECULAR

sustenta la C.

María Aurora Londoño Avendaño

sobre la Tesis intitulada:

Análisis filogenético-funcional de regiones de control replicativo y transcripcional en virus de DNA de cadena sencilla

que se desarrolló bajo la dirección de


Dra. Lina Raquel Riego Ruiz
Dr. Gerardo Rafael Argüello Astorga

El Jurado, después de deliberar, determinó

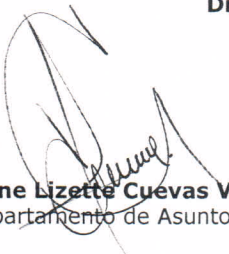
APROBARLA

Dandose por terminado el acto a las 14:00 horas, procediendo a la firma del Acta los integrantes del Jurado. Dando fe el Secretario Académico del Instituto.

A petición de la interesada y para los fines que a la misma convengan, se extiende el presente documento en la ciudad de San Luis Potosí, S.L.P., México, a los 19 días del mes de marzo de 2010.


Dr. Marcia Borja María
Secretario Académico




Mtra. Ivonne Lizette Cuevas Vélez
Jefa del Departamento de Asuntos Escolares

A

Barbariccia, Scamiglione, Alichino y Beator

John H. Campbell. 1993. Mem Ass Aust Palaeontol 15:43-50

AGRADECIMIENTOS

Al doctor Gerardo Arguello y a los diferentes miembros que integraron el Comité de Becas del IPICYT en los años 2006-2008, por creer en mí y darme la oportunidad de realizar este doctorado.

A la doctora Lina Riego por su guía en el proceso de investigación y publicación de resultados, pero también por su paciencia y mediación en los momentos de crisis.

Nuevamente al doctor Arguello por las enseñanzas que me transmitió y que muchas veces sólo adquirí a fuerza de golpes.

A los QFBs Mariana Cantú-Iris y Amando Mauricio-Castillo, y al BQ Bernardo Bañuelos-Hernández por su apoyo en la parte experimental e igualmente al biólogo Salvador Ambriz-Granados por facilitar el trabajo de laboratorio.

A las doctoras Irene Castaño y Laura Silva por sus comentarios y aportes.

A todos los integrantes del Laboratorio de Biología Molecular de Plantas por brindarme su amistad durante estos años.

A la señora Martha Gallegos por aceptarme como parte de su familia.

RESUMEN

La mayoría de los virus con genomas circulares de DNA de cadena sencilla se replican por el mecanismo de círculo rodante. Esta característica hace que todos codifiquen una proteína iniciadora de la replicación que tiene actividad de endonucleasa, usualmente conocida como Rep. Las relaciones evolutivas entre los virus que codifican esas proteínas Rep no son claras. Se usó un análisis teórico con una aproximación heurística para analizar el origen de replicación y la respectiva proteína Rep de tres familias virales (*Geminiviridae*, *Nanoviridae* y *Circoviridae*), con el fin de detectar similitudes funcionales que indiquen relaciones evolutivas entre ellas; los resultados muestran que en todos los casos la proteína Rep tiene dos regiones con la misma configuración espacial que están involucradas en la unión específica a secuencias repetidas, llamadas iterones, presentes en el origen de replicación viral, y esto hace que las tres familias resulten más emparentadas de lo que antes se pensaba. Por otro lado se identificaron huellas biogeográficas en la proteína Rep de los virus del género *Curtovirus* (familia *Geminiviridae*) y señales de eventos de recombinación que indican que los miembros típicos de éste género probablemente se diversificaron en Norteamérica, tras adquirir un segmento genómico de un begomovirus (otro género de los geminivirus) que permaneció aislado por millones de años en Sudamérica. Adicionalmente, se estableció un sistema de preparación de protoplastos a partir de células vegetales cultivadas en suspensión, el cual se estandarizó midiendo la actividad β -glucuronidasa generada por promotores de begomovirus fusionados al gen *uidA*; dicho sistema sirve para hacer experimentos de relevancia en los campos de la virología y biología molecular de plantas, por ejemplo aquellos surgidos de los análisis teóricos.

Palabras claves: círculo rodante, geminivirus, circovirus, nanovirus, iterones, proteína Rep, curtovirus, huella bio-geográfica, recombinación, protoplastos, promotor, β -glucuronidasa.

ABSTRACT

Most viruses with circular single-stranded DNA genome replicate by the rolling circle mechanism. Because of this characteristic they encode a rolling circle initiator protein with endonuclease activity usually called Rep. The evolutionary relationships between the viruses codifying Rep proteins are poorly understood. Here we used a theoretical analysis with and heuristic approach to analyze the replication origin and the respective Rep protein of viruses from three viral families (*Geminiviridae*, *Nanoviridae* and *Circoviridae*), aimed to detect some functional similitude indicative of relationships between the families; the results show that in all cases the Rep protein has two regions in the same spatial configuration that are involved in the specific binding of repeated DNA sequences, called iterons, present in the replication origin; this finding makes the studied viral families more related than it was believed before. Additionally the evolution of the genus *Curtovirus* in the family *Geminiviridae* was reviewed; with data from recombination analyses, detection of bio-geographical finger prints and phylogenetic reconstruction it was got evidence suggesting that the typical members of this genus likely diversified in North America, after having acquired a genomic segment from a begomovirus (another genus of *Geminiviridae*) who stayed in isolation during millions of years in South America. It was also standardized an experimental system to prepare protoplasts from plant cells cultures; the system was tested measuring the β -glucuronidase activity from molecular constructs containing begomoviral promoters fused the *uidA* gene; it will let to perform experiments in the areas of virology and molecular biology of plants, like those derived from the theoretical analyses.

Key words: rolling circle, geminivirus, circovirus, nanovirus, iteron, Rep protein, curtovirus, bio-geographical print, recombination, protoplasts, promoter, β -glucuronidase.

ÍNDICE

	Pág.
I. CONSTANCIA DE APROBACIÓN DE LA TESIS.....	ii
II. CRÉDITOS INSTITUCIONALES	iii
III. ACTA DE EXÁMEN DE GRADO.....	iv
IV. DEDICATORIA.....	v
V. AGRADECIMIENTOS.....	vi
VI. RESUMEN.....	vii
VII. ABSTRACT.....	viii
VII. TRABAJO DE INVESTIGACIÓN.....	x
1. Introducción general.....	11
1.1. Virus ssDNA.....	11
1.2. Replicación por círculo rodante.....	14
1.3. Geminivirus y sus DNAs satélites.....	17
1.4. Nanovirus.....	24
1.5. Circovirus.....	27
1.6. Literatura citada.....	30
2. Estudio teórico de la proteína iniciadora de la replicación por círculo rodante.....	37
2.1. Antecedentes.....	37
2.2. Material y métodos.....	40
2.3. Resultados.....	52
2.4. Referencias.....	53
3. Historia evolutiva del género <i>Curtovirus</i>	56
3.1. Antecedentes.....	56
3.2. Métodos experimentales.....	60
3.3. Análisis de secuencias.....	60
3.4. Resultados.....	71
3.5. Referencias.....	72
4. Estandarización de un sistema experimental para analizar promotores de begomovirus.....	75
4.1. Antecedentes.....	75
4.2. Material y métodos.....	78
4.3. Resultados.....	83
4.4. Discusión y perspectivas.....	90
4.5. Referencias.....	92
VIII. CONCLUSIONES GENERALES.....	95
IX. ANEXOS.....	97
1. Protocolos de laboratorio.....	97
2. Artículo aceptado en <i>Archives of Virology</i>	109

VIII. TRABAJO DE INVESTIGACIÓN

Este es un trabajo principalmente teórico en el que se analizan genomas que se replican por el mecanismo de círculo rodante, los cuales ocurren en los tres dominios de vida (bacterias, arqueobacterias y eucariotes), en la forma de virus y plásmidos pequeños (de 1-6 kb). El enfoque general del trabajo es entender las relaciones evolutivas, y/o similitudes funcionales que existen entre los diferentes linajes que usan este mecanismo replicativo, entendiendo previa, o paralelamente la evolución y naturaleza de cada linaje. La información que se plasma en esta tesis habla exclusivamente de los resultados obtenidos del análisis de tres familias virales cuyo genoma de ssDNA se multiplica por círculo rodante: *Geminiviridae*, *Nanoviridae* y *Circoviridae*, y está organizada en cuatro secciones. En la primera sección se describen las tres familias virales; en la segunda se expone el trabajo teórico de delimitación del dominio de especificidad de unión al DNA en la proteína iniciadora de la replicación de los virus de las familias *Nanoviridae* y *Circoviridae*, enfatizando las semejanzas que estos tienen con los geminivirus y los plásmidos bacterianos de la familia pMV158. En la tercera sección se muestra un trabajo donde se explora el enigmático origen del género *Curtovirus* de la familia *Geminiviridae*. En la última sección se menciona la parte experimental que se realizó, la cual consistió en establecer un sistema experimental para el análisis de secuencias génicas reguladoras en *cis*; este sistema permitirá hacer análisis funcionales y sacar ventajas de las observaciones obtenidas del estudio teórico, además de que aumenta la capacidad operativa del grupo de trabajo.

1. Introducción general

1.1. Virus de DNA de cadena sencilla

Los virus de DNA de cadena sencilla (ssDNA) constituyen cerca del 15% de los virus conocidos hasta el momento. Taxonómicamente corresponden a las familias *Inoviridae* (infectan bacterias y micoplasmas), *Microviridae* (infectan bacterias y espiroplasmas), *Geminiviridae* y *Nanoviridae* (infectan plantas), *Circoviridae* (infectan vertebrados), *Parvoviridae* (infectan vertebrados e invertebrados) y la recientemente propuesta familia *Anelloviridae* (infecta vertebrados). Los virus representativos de cada familia son el fago M13, el fago ϕ X174, el *virus del mosaico dorado del tomate* (TGMV), el *virus del amarillamiento necrótico del haba* (FBNYV), el *Circovirus porcino 1* (PCV1), el *virus adeno-asociado 2* (AAV2), y el *virus Torque teno* (TTV), respectivamente (Fauquet et al. 2005, Hino & Prasetyo 2009). Sus genomas pueden estar constituidos de una o varias moléculas de estructura circular, con excepción de los inovirus y microvirus, en los que en general el genoma cambia entre lineal y circular a lo largo del ciclo viral y la molécula que se empaqueta en virión es característica de las especies (Carter & Saunders, 2007), y los parvovirus que tienen genomas lineales. En todos estos virus el genoma se empaca en una cápside isométrica, excepto en los inovirus, que usan una cápside filamentosa, y ninguno de ellos adquiere envoltura a su salida de la célula (Fauquet et al. 2005).

Aunque predominan los genomas circulares y su proceso de replicación involucra la generación de un DNA intermediario de doble cadena, entre estas familias ocurren varios mecanismos replicativos. Los genomas lineales se multiplican mediante el mecanismo de replicación por horquilla rodante; aquí el mismo genoma funciona como iniciador para la polimerización del DNA, ya que en los extremos posee secuencias palindrómicas que le permiten formar una asa corta de doble cadena en la que se conserva un extremo OH-3' libre que le sirve de sustrato a la DNA polimerasa (Carter & Saunders, 2007). Para los

genomas circulares (con excepción de los anellovirus) se conocen dos mecanismos replicativos: la replicación tipo theta y la tipo sigma. El mecanismo predominante es la replicación de tipo sigma, también conocida como replicación por círculo rodante, pero se sabe que algunos virus pueden cambiar de un mecanismo al otro en determinadas circunstancias. Para ambos mecanismos se necesita una proteína iniciadora de la replicación que se encarga de generar el OH-3' sustrato de la polimerasa, mediante la acción de su actividad endonucleasa. En los anellovirus se desconoce cuál es el mecanismo de replicación ya que no se ha encontrado una proteína iniciadora, ni alguno de los otros elementos que participan en la replicación de genomas circulares pequeños; su mecanismo de replicación parece depender en gran parte de proteínas del hospedero (Hino & Prasetyo, 2009).

Todos estos virus de DNA de cadena sencilla tienen una tasa de mutación alta (en el intervalo de 10^{-4} sustituciones/sitio/año), comparable a la tasa de los genomas de RNA (van der Walt et al. 2008, Shackelton et al. 2005), lo cual hace que tengan una diversidad alta o potencial para diversificarse con rapidez si encuentran condiciones que favorezcan su dispersión. Ésta característica sólo se notó en experimentos recientes donde se estudió la evolución de éste tipo de virus mediante muestreos en el tiempo (Gibbs et al. 2010), lo cuales no se hicieron antes porque por décadas se había supuesto que como estos virus usan las polimerasas del huésped, debían tener una tasa de mutación acorde con la fidelidad de éstas enzimas (Duffy & Holmes 2009).

En este trabajo nos enfocamos en las familias que se replican por círculo rodante (CR), las cuales a pesar de compartir este mecanismo no tienen, aparentemente, una relación filogenética directa. De hecho, lo único que tienen en común es que usan una proteína iniciadora de la replicación por CR, y como se observa en la figura 1.1, aún en esta proteína tienen grandes diferencias, ya que solo comparten el dominio endonucleasa. Este dominio a su vez tiene algunas variaciones que hacen que el inicio de la replicación por círculo rodante no ocurra exactamente de la misma manera en todas las familias. Es importante resaltar además que varias familias de plásmidos bacterianos, entre ellas las familias pMV158 y pT181, comunes en bacterias Gram-positivas (Khan

2003, del Solar et al. 1998), y varios plásmidos de arqueobacterias (Soler et al. 2007, Marsin & Forterre 1999) también usan el mecanismo de círculo rodante para multiplicar sus genomas y de igual manera, en ellos la proteína iniciadora de la replicación contiene un dominio endonucleasa que le confiere particularidades al proceso.

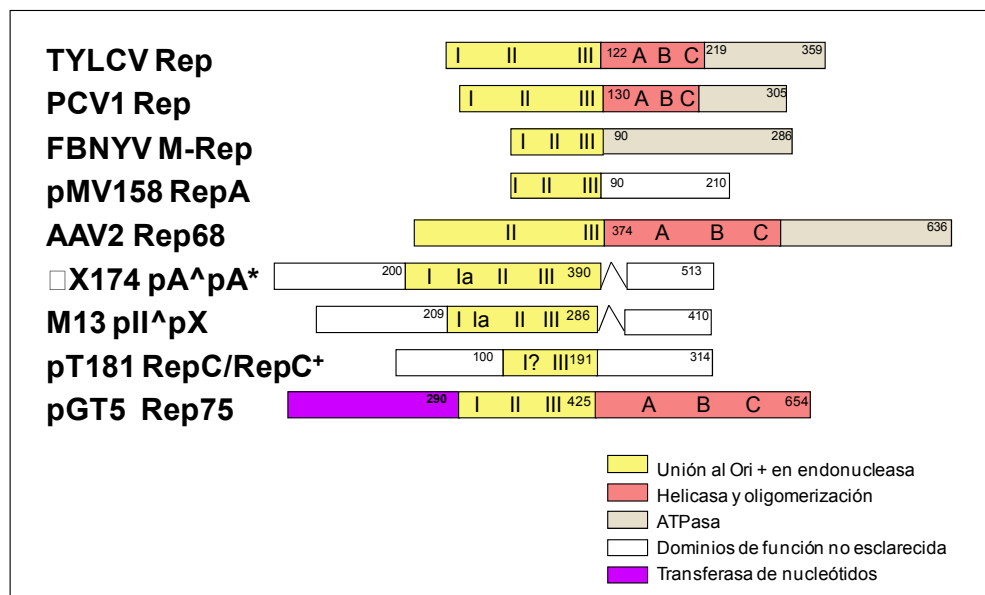


Figura 1.1. Organización de dominios en la proteína iniciadora de la replicación por círculo rodante de los representantes de las familias virales con genomas de ssDNA mencionados al principio de éste capítulo, del plásmido pMV158 de *Streptococcus agalactiae*, del plásmido pGT5 de la arqueobacteria *Pyrococcus abyssus* y del plásmido pT181 de *Staphylococcus aureus*. En los fagos phiX174 y M13 se producen dos proteínas a partir del transcrito primario, la que participa en la iniciación de la replicación en phiX174 se llama pA* y la de M13 es el producto pX. +La proteína RepC es una versión de RepC que lleva unido un oligodesoxiribonucleótido que participa en la regulación de la actividad replicativa.

Las diferencias en el inicio de la replicación CR tienen que ver entonces con los dominios funcionales que complementan la actividad endonucleasa y con las propiedades del dominio endonucleasa en sí mismo. Así, estos dos aspectos han dado origen a varias clasificaciones para las proteínas iniciadoras de RCR. La primera de ellas fue propuesta por Ilyna y Koonin en 1992, quienes establecieron dos superfamilias de acuerdo al arreglo de tres motivos conservados en el dominio endonucleasa (Ilyna & Koonin 1992). Dos de esos

motivos tienen funciones concretas en la unión y corte del DNA, siendo así que el motivo II (consenso xpHuHuuux, u= L, I, M, V, Y,F, W, T, A) posee dos histidinas que unen cationes divalentes necesarios para la función endonucleasa, y el motivo III (uxxYuxKxx) tiene uno o dos residuos de tirosina que son el sitio activo de corte (Campos-Olivas 2002), mientras que el primer motivo (consenso FuTLTxxx) parece ser meramente estructural ya que no se le ha asignado una función bioquímica. Según esta primera clasificación, todas las proteínas que poseen los tres motivos conservados, independientemente de la localización del dominio endonucleasa, pertenecen a la superfamilia Rep1-2-3 (Ilyna & Koonin 1992, Koonin & Ilyna 1993).

La clasificación en familias de proteínas que hacen las bases de datos ProDom y Profam (<http://pfam.sanger.ac.uk>) se basa en la arquitectura de la proteína completa, de tal manera que en los iniciadores de replicación CR los dominios adicionales al de endonucleasa juegan un papel importante en su agrupamiento (Finn et al. 2008). De acuerdo a estas bases de datos se reconocen al menos las siguientes familias: Gemini_AL1 (PF00799) para geminivirus; Viral_Rep (PF02407) para circovirus y nanovirus; Rep _N (PF08724) para parvovirus; Rep_1 (PF01446) para los plásmidos tipo pGT5; Rep_3 (PF01051) para los plásmidos del tipo pMV158; Phage_GPA (PF05840) para los inovirus, y Phage_CRI (PF05144) para los microvirus.

1.2. Replicación por círculo rodante

Esta tesis está especialmente enfocada en algunos pasos de la replicación por círculo rodante, por lo que es necesario dedicar una sección completa a describir este proceso; la figura 1.2 ilustra los pasos básicos.

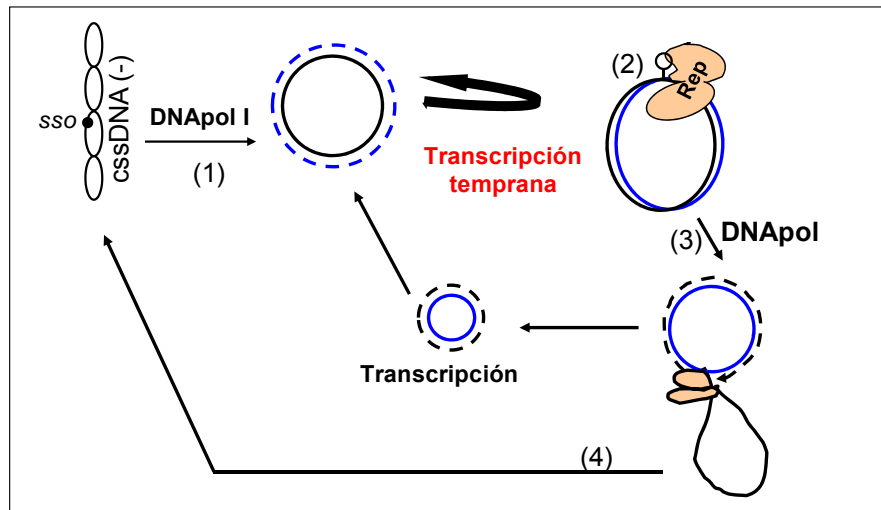


Figura 1.2. Pasos básicos de la replicación por círculo rodante: 1) Paso de DNA circular de cadena sencilla (cssDNA) a DNA circular de cadena doble (cDsDNA); 2) Corte de la estructura tallo-asa para generar OH-3' libre; 3) Elongación de la cadena naciente; 4) Religación de moléculas y liberación cssDNA y cDsDNA.

1.2.1 Conversión del cssDNA en cDsDNA

Lo primero que sucede con los genomas que se replican por este proceso es el paso de moléculas circulares de ssDNA a círculos de dsDNA. Ésto se hace a través de un origen de replicación de DNA de cadena sencilla (sso), el cual tiene en una estructura secundaria estable e incluye uno de dos elementos alternativos: una asa amplia a la que se une un oligonucleótido de unos 80 pb, el cual puede estar disponible como parte del ácido nucleico que se empaqueta en la cápside viral y se conoce como iniciador de ssDNA (Gutiérrez 2000), ó bien, una región de reconocimiento para una RNA polimerasa ó DNA-primasa del hospedero, proteínas que una vez ubicadas sintetizan un “primer” para generar el OH-3' que servirá de sustrato a la DNA-polimerasa (Khan 2005, Khan 2003). Posteriormente la replicación procede de una manera discontinua y la nueva molécula de dsDNA sirve como molde para la transcripción del gen que codifica a la proteína iniciadora de la replicación.

1.2.2 Generación del sustrato OH-3'

Una vez que se producen las proteínas iniciadoras, estas son reclutadas en el origen de replicación de CR, donde reconocen el sitio de corte de acuerdo a ciertas propiedades que éste posee según el linaje de replicón. En general el

inicio de CR, también conocido como origen de replicación para doble cadena, o *dso* por sus siglas en inglés, se caracteriza por ser una región genómica con potencial para formar una estructura tallo-asa en la cual el asa, rica en A y T, está formada por una secuencia de al menos nueve nucleótidos, de los cuales los últimos cinco son conservados entre todos los miembros del linaje en cuestión. El sitio de corte se encuentra entre los nucleótidos 8 y 9 del nona-nucleótido del asa (Khan 2003, Gutiérrez 2000).

Según el modelo más aceptado (y que aplica a varios linajes con replicación CR, incluyendo geminivirus y familias de plásmidos como la de pMV158 y la de pT181) (Ruiz-Masó et al. 2007, Khan 2003, Gutiérrez 2000), para guiar a la proteína a su adecuado posicionamiento, el *dso* posee unas secuencias repetidas cercanas a la región que forma la estructura tallo-asa, a las cuales se une la proteína Rep de manera secuencial, desplazándose hacia el tallo-asa gracias a interacciones entre monómeros de sí misma (Singh et al. 2008, Khan 2005). Otros replicones RCR carecen de estructura tallo-asa y algunos no tienen secuencias repetidas, y esto se relaciona con la organización de dominios de la proteína Rep. En los casos en que hay estructura tallo-asa en el *dso*, una vez que una proteína Rep alcanza el sitio de corte, corta la región de ssDNA que forma el asa de la cadena positiva (molécula ssDNA empaquetada en el virión), a través de un ataque nucleofílico al enlace fosfodiéster del DNA por el residuo de tirosina contenido en el motivo conservado III del dominio endonucleasa (Khan 2005, Campos-Olivas 2002).

1.2.3 Elongación

La proteína iniciadora de RCR permanece unida al extremo 5' de la molécula de DNA en forma de tirosil-éster, mientras una DNA polimerasa del huésped extiende el extremo 3' usando como molde a la cadena negativa, generada en el paso de creación del dsDNA.

1.2.4 Religación y liberación de moléculas cssDNA

Aunque la terminación de un ciclo de RCR no ha sido descrita en detalle en los virus, para los plásmidos se conoce que una vez que se ha copiado todo el genoma y se regenera el *dso*, la misma proteína iniciadora se encarga de

finalizar el proceso (Ruiz-Masó et al. 2007, Khan 2005). Por un lado el dominio de unión de DNA de la proteína Rep que estaba unida al extremo 5' de la molécula de cadena positiva vuelve a reconocer los sitios del *dso* recién regenerado, poniendo cerca un extremo del otro; luego por un proceso aún poco claro en el que se cree participa la misma tirosina que mantiene el enlace tirosil-éster se da la ligación y liberación de la primera molécula cortada, y se deja a la vez una nueva molécula partida para que el ciclo se repita. Es importante resaltar que en este proceso no se generan copias en tándem del genoma replicado, como ocurre con el proceso de amplificación por círculo rodante que se hace con la polimerasa del fago phi29.

1.3. Generalidades de los Geminivirus

1.3.1. Taxonomía y distribución

La familia *Geminiviridae* está conformada actualmente por los géneros *Mastrevirus*, *Curtovirus*, *Begomovirus* y *Topocuvirus*. Se dice que ésta es una familia de origen monofilético ya que todos los virus que la conforman se caracterizan por poseer una cápside isométrica que forma una estructura geminada mediante la unión de dos semi-icosaedros (Rybicki 1994). La agrupación de los geminivirus en cuatro géneros se basa en el insecto vector y el tipo de plantas hospederas. Así, los mastrevirus son transmitidos por algunas especies de la familia *Cicadellidae* (*Cicadulina mbila* usualmente) e infectan plantas mono y dicotiledóneas; los curtovirus los transmiten las chicharritas de la especie *Circulifer tenellus* (*Cicadellidae*) e infectan plantas dicotiledóneas, y a los begomovirus los transmite la mosquita blanca (*Bemisia tabaci*, *Aleyrodidae*) a plantas dicotiledóneas (Fauquet et al. 2005, Fauquet & Stanley 2005).

En cuanto a su distribución, los mastrevirus están restringidos a Europa, Asia y África (el Viejo Mundo) (Nahid 2008), pero de los begomovirus y los curtovirus se han encontrado representantes tanto en el Viejo Mundo como en las Américas (Padidam et al. 1999, Ha et al. 2006, Baliji et al. 2004); el género *Topocuvirus* sólo tiene una especie, encontrada en Norteamérica, la cual ha

sido poco estudiada, aunque se sabe que es transmitida por el saltahojas chupador (“treehopper”) *Micrutalis malleifera* (*Membracidae*) a varias especies de plantas dicotiledóneas (Bridson et al. 1996).

1.3.2. Organización genómica

La organización genómica del representante típico de cada uno de los géneros se indica en la figura 1.3. Se trata de genomas entre 2.5 y 3.0 kb que contienen varios genes, en un arreglo que maximiza el almacenamiento de información en la molécula de DNA mediante la codificación de genes en ambas cadenas y el solapamiento de genes. Así pues, la organización genómica en general se divide en la región que codifica los genes desde la cadena +, o genes en sentido del virión, la de los genes codificados de la cadena -, también conocidos como genes en sentido complementario, y las regiones no codificantes o intergénicas, que poseen elementos reguladores transcripcionales y de la replicación, especialmente la región intergénica mayor.

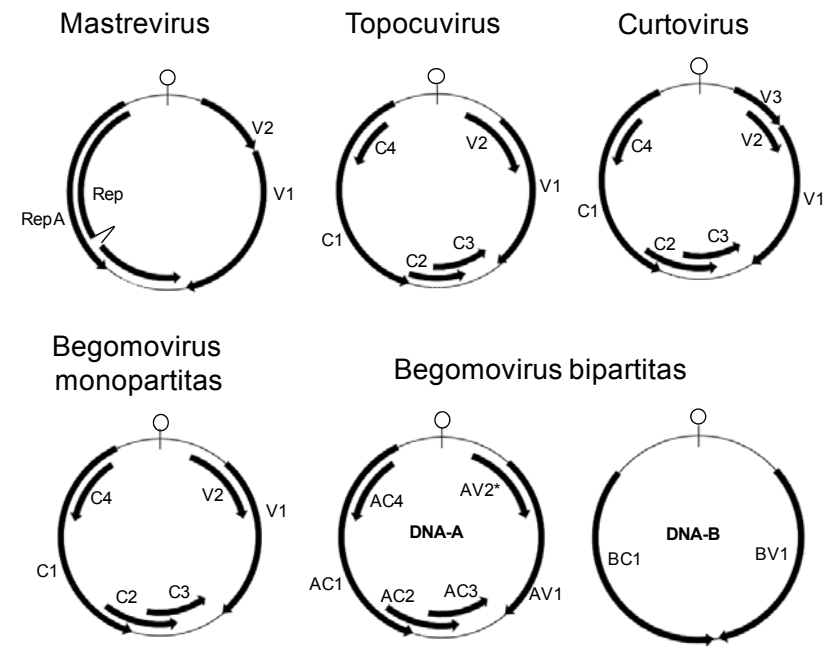


Figura 1.3. Organización genómica de los cuatro géneros de la familia *Geminiviridae*. Las flechas con la punta en el sentido de las manecillas del reloj indican los genes en sentido del virión, y las flechas en dirección contraria a los genes del sentido complementario. *El gen AV2 sólo se encuentra en los begomovirus del Viejo Mundo.

Se puede decir que el genoma de todos los geminivirus posee una región con potencial para generar una estructura tallo-asa, en la que el asa está compuesta por el nonanonucleótido TAATATTA'C y el tallo se caracteriza por ser una secuencia palindrómica rica en GC; la comilla (') antes del último nucleótido de la secuencia indica el sitio específico donde la endonucleasa Rep hace el corte que permite generar el sustrato para la DNA polimerasa de la planta.

Todos los geminivirus tienen un gen que codifica para la proteína iniciadora de la replicación (C1 ó Rep) y otro para la proteína de la cápside (V1 ó CP), pero hay además una serie de genes adicionales en los miembros de cada género, los cuales no pueden ser considerados característicos de cada uno de los cuatro linajes, ya que no se conoce con precisión como se originaron los cuatro géneros, ni el flujo que han tenido los diferentes componentes genómicos entre ellos (Rybicki 1994, Varsani et al. 2009). La tabla 1.1 deja muy claro este punto, ya que muestra cómo los diferentes linajes comparten genes en la misma posición, los cuales incluso llevan el mismo nombre (homólogos posicionales), pero la función del gen no es la misma en todos ellos (no son homólogos funcionales u homólogos verdaderos, es decir, tienen orígenes distintos).

Tabla 1.1. Función de los genes geminivirales

Gen	Proteína	Función
V1, AV1	CP	Encapsidar el genoma viral
V2, AV2	MP	Movimiento del genoma y supresión del silenciamiento
V2 curto		Regulador acumulación de moléculas de DNA
V3	MP	Movimiento del genoma
C1, AC1, Rep	Rep	Inicio de replicación
C2, AC2	TrAP	Transactivador de genes tardíos Supresor de silenciamiento
C3, AC3	REn	Aumento de la replicación
C4, AC4		Movimiento (begomovirus monopartitas), Supresor del silenciamiento,
RepA	RepA	Modificación del ciclo celular
BV1	NSP	Entrada y salida de genomas virales hacia el núcleo
BC1	MP	Movimiento

Un trabajo reciente propone un quinto género en la familia, que sería el género *Ecuvirus*, para ubicar allí al *Virus del rayado de Eragrostis curvula* (ECSV), descrito hace poco (Varsani et al. 2009), el cual infecta plantas monocotiledóneas y tiene una organización genómica en la que los genes en sentido contrario del virión incluyen un gen C1 que codifica una proteína similar a la Rep de los begomovirus y un gen C2 cuyo producto proteico es ligeramente parecido a la proteína TrAP, y los del sentido del virión están organizados al estilo mastrevirus, aunque la similitud entre las proteínas CP de ambos grupos es baja.

1.3.4. Ciclo infeccioso

El ciclo infeccioso de estos virus es el siguiente: una vez que un insecto vector se alimenta del floema de una planta infectada se lleva consigo los viriones contenidos en la savia. En general los viriones solo transitan a través del sistema digestivo, entran al hemocele, y regresan al aparato bucal sin modificaciones aparentes, para luego ser inyectados en el floema de la próxima planta de la que el insecto se alimenta (Rosell et al. 1999). Dentro de la planta, el virión es transportado al núcleo gracias a una señal de localización nuclear que contiene la proteína de la cápside; una vez allí el DNA se libera y se inicia el proceso de replicación, se transcriben y sintetizan las proteínas del virus (Gutierrez 2000).

La primera proteína producida es Rep, que se encarga de modificar el ciclo de la célula vegetal haciéndola entrar en endo-replicación, que es un ciclo en el que se multiplica el material genético celular sin que haya citocinesis; luego se sintetizan la proteína TrAP que es necesaria para la síntesis posterior de las proteínas del movimiento y de la cápside y la proteína REn que sirve para aumentar las copias del genoma viral (Shimada-Beltran & Rivera-Bustamante 2007). Las proteínas del movimiento transportan las moléculas de ssDNA a través de los plasmodesmos y lo introducen al núcleo de las nuevas células hospederas, esparciendo así el virus a través de la planta (Rojas et al. 2001), mientras otro tanto de moléculas de DNA están siendo empaquetadas en las cápsides bigeminadas y listas para pasar al siguiente vector. La mayoría de los

geminivirus solo se mueven entre las células del floema y prefieren replicarse en las células completamente diferenciadas.

1.3.5. Problemas que generan

En años recientes los virus de la familia *Geminiviridae* se han convertido en amenazas para la producción agrícola de las regiones tropicales y subtropicales del mundo. Las primeras especies de geminivirus se descubrieron en la década de los 70's (Goodman 1977, Galvez & Castaño 1976) y desde entonces la diversidad conocida ha ido en aumento, de tal manera que en un lapso de diez años el número de especies conocidas llegó a cuadruplicarse (Fauquet & Stanley 2005, Padidam et al. 1995). También se ha incrementado la frecuencia de infecciones virales y epidemias agrícolas causadas por geminivirus debido a los cambios demográficos recientes, los cuales han modificado los sistemas agrícolas tradicionales y la distribución de los insectos vectores (Martin et al. 2000, Seal et al. 2006a).

Básicamente lo que ocurre es que un biotipo de la mosquita blanca (*Bemisia tabaci*, Hemiptera: Aleyrodidae) que es más hábil en la transmisión de los virus y que antes habitaba en la región mediterránea se ha dispersado a los demás continentes y allí ha sido capaz de transmitir los geminivirus presentes en plantas silvestres, a plantas de interés agrícola (Polston et al. 1997, Seal et al. 2006a). De esta manera especies de virus que existían en malezas desde tiempos remotos, quizá sin causarles mucho daño, se hacen evidentes una vez infectan un cultivo (Mansoor et al. 2006). También el aumento de las extensiones cultivadas y la siembra en monocultivos contribuyen al fenómeno, dado que se limita la diversidad de especies de plantas disponibles como alimento y hospedaje del vector (Seal et al. 2006b, Brown & Bird 1995).

Con unas pocas excepciones, todas las especies de geminivirus conocidas hasta ahora tienen la capacidad de infectar plantas de varias familias, muchas de importancia económica como la familia de las Solanaceas (papa, chile, tomate, tabaco) que es especialmente sensible, y las familias *Fabaceae* (frijol, haba, soya), *Gramineae* (maíz, arroz, trigo), *Chenopodiaceae* (remolacha o

betabel) y *Cucurbitaceae* (melón, sandía, calabaza). Las plantas infectadas presentan una sintomatología que incluye la aparición de mosaicos, el “enchinamiento”, enrollamiento, o deformación de las hojas, detención del crecimiento (enanismo) y producción de frutos manchados, pequeños y/o deformes (Seal et al. 2006b, Creamer et al. 2005, Garzón-Tiznado et al. 2002).

Este tipo de síntomas se han venido observando en México desde 1970 en los cultivos de chile y jitomate, especialmente en los estados de Sinaloa y Jalisco, pero otra serie de cultivos hortícolas también se han visto afectados (Hernández-Zepeda et al. 2007, Garzón-Tiznado et al. 2002, Torres-Pacheco et al. 1996, Brown et al. 1993). Todos los virus identificados en estos cultivos han sido del tipo Begomovirus, pero en meses pasados se identificó por primera vez un virus del género *Curtovirus* en México, el cual se encontró en cultivos de chile en Villa de Arista, SLP, y resultó ser una variante del *virus moderado de la punta rizada de la remolacha* (BMCTV), que ya se había reportado como causante de epidemias en Estados Unidos (Creamer et al. 2005, Stenger & McMahon 1997) y su hallazgo representa un dato de alerta sobre el esparcimiento de estos agentes fitopatógenos.

Para evitar los brotes de epidemias se utilizan estrategias que previenen la entrada del virus a las plantas, los cuales van dirigidos al control del insecto vector, o a la regulación de los ciclos de siembra con el fin de que al campo salgan plantas más vigorosas (Seal et al. 2006a, Rampersad 2003). También gracias a lo que se conoce de la biología básica de estos virus se han podido considerar algunos mecanismos para frenar la infección una vez que el virus ha entrado a la planta, y que evitarían el abandono o quema de las cosechas, que es la medida que se suele tomar en caso de epidemias. Dentro de estos mecanismos de control post-infección se cuentan las plantas transgénicas que expresan proteínas o pequeñas moléculas capaces de disminuir la replicación viral o de aumentar la capacidad de respuesta de la planta (Bonfim et al. 2007, Vanderschuren et al. 2007, López-Ochoa et al. 2006). Sin embargo, el uso de estas opciones no se ha popularizado por varias razones, entre ellas que se puede ver afectado el rendimiento de la planta, o también por los problemas sociales que se generan alrededor de la introducción de transgénicos.

1.3.6. Genomas satélite

En las infecciones por begomovirus monopartitas se han encontrado además unas moléculas de ssDNA más pequeñas que se han llamado genomas satélites y de los cuales hay dos grupos, denominados alfa y beta-satélites (1.2 y 0.6 kb, respectivamente) (Briddon & Stanley 2006). Los alfa-satélites son replicones de círculo rodante ya que tienen un gen en sentido del virión cuyo producto proteico pertenece a la familia Viral-Rep y además tienen una secuencia con potencial de formar una estructura tallo-asa y secuencias repetidas adyacentes a ésta. Los beta-satélites carecen de las regiones iteradas adyacentes a la estructura tallo-asa y codifican para una proteína en el sentido complementario del virión, llamada β C1 que no tiene ninguna similitud con las de los geminivirus, pero que su presencia puede resultar ventajosa para el establecimiento de la infección en algunos casos (Guo et al. 2008).

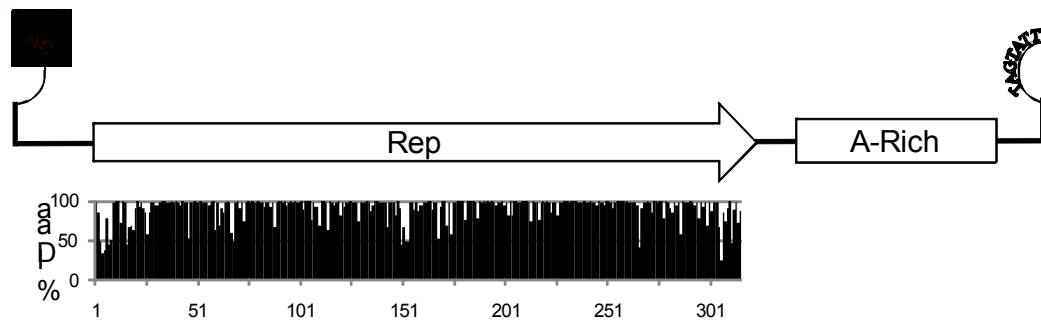


Figura 1.4 Organización del genoma circular de los alfa-satélites y gráfica del porcentaje de identidad de la proteína Rep de este grupo. La figura muestra un esquema 5'-3' del genoma, el cual inicia en el sitio de corte para la proteína Rep contenido en la estructura tallo-asa; el marco de lectura de la proteína Rep está representado con una flecha y la región rica en adenina por un rectángulo.

Como se observa en la figura 1.4, el genoma de los alfa-satélites consiste en tres regiones que son: una estructura tallo-asa, el marco de lectura de la proteína Rep y una región rica en adenina. La región rica en adenina es una característica distintiva del linaje, pero también lo son las proteínas Rep ya que son bastante idénticas entre sí y siempre forman un grupo definido cuando se comparan con las de otros linajes.

Los alfa-satélites, que también se conocen como satélites de tipo nanovirus, o DNAs¹, tienen la capacidad de auto-replicarse pero dependen del begomovirus al que están asociados para ser movidos y encapsidados. Los beta-satélites son más dependientes ya que no tienen proteína iniciadora de la replicación, y más aún, hasta ahora no se sabe exactamente cómo es que se replican, dado que no comparten con el begomovirus “patrocinador” los elementos en *cis* indispensables para la especificidad de la replicación por círculo rodante.

1.4. Generalidades de los Nanovirus

1.4.1. Distribución y taxonomía

Los nanovirus son virus de angiospermas que poseen genomas multipartitas compuestos por moléculas circulares de ssDNA de 0.9 a 1.2 Kb, encapsuladas en cápsides icosaédricas muy pequeñas (diámetro ~18 nm), están restringidos al Viejo Mundo y todos ellos son transmitidos por áfidos. La familia, *Nanoviridae*, se divide en dos géneros: *Nanovirus* y *Babuvirus*. El primero incluye tres especies que infectan leguminosas: el *virus del amarillamiento necrótico del haba* (FBNYV), el *virus del enanismo del Astragalus* (MVDV) y el *virus del enanismo del trébol subterráneo* (SCSV) (Gronenborn 2004). Las dos especies que integran el género *Babuvirus* infectan al plátano y especies relacionadas, y se denominan *virus del arracimamiento apical del plátano* (BBTV) y *virus del arracimamiento apical del abacá* (ABTV) (Sharman et al. 2008). Existe un miembro de esta familia viral que ocupa una posición taxonómica incierta, éste es el *virus de la defoliación del coco* (CFDV), que se diferencia de los otros por tener un genoma monopartita de unos 1300 pb (Merits et al. 2000), y que no puede ser clasificado como un genoma satélite porque no posee la región rica en adenina característica de éstos y además de Rep el genoma codifica para una segunda proteína en el sentido del virión, en otro marco de lectura.

La distribución de estos virus en el Viejo Mundo no es muy amplia, siendo así que los babuvirus se limitan al Sudeste Asiático y algunas islas del Pacífico Sur; los que infectan leguminosas se encuentran en algunos países del Medio Oriente (FBYNV) y alrededor de la cuenca del Mediterráneo, en Japón (MVDV) y Australia (SCSV), y CFDV sólo se ha observado en Vanuatu, en el Pacífico Sur.

1.4.2. Organización genómica

Con excepción de CFDV, todos los nanovirus tienen genomas con múltiples componentes que se empaquetan en cápsides individuales y son transmitidos por el vector de manera independiente. En la figura 1.5 se clasifican los componentes virales en aquellos que son indispensables para el establecimiento y proliferación de la infección, y aquellos que pueden ser considerados componentes genómicos no-esenciales; estos últimos replicones codifican proteínas Rep y por lo tanto se comportan como entidades satélites con capacidad de auto-replicarse (Bell et al. 2002).

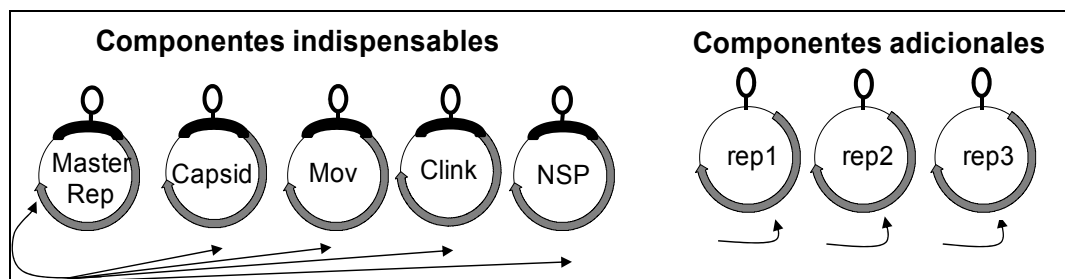


Figura 1.5. Organización genómica de los nanovirus multipartitas. Las líneas con punta de flecha debajo del esquema del replicón indican si se trata de un genoma que se auto-replica, o si depende una proteína codificada en otro genoma. Las proteínas Rep, de la cápside y Mov se encargan del inicio de la replicación, de la contención y el movimiento del genoma viral, respectivamente. La función de la proteína Clink equivale a la del dominio de oligomerización y unión a la proteína retinoblastoma que se encuentra en la región media de la proteína Rep de los geminivirus; dicha función consiste en la modificación del ciclo celular para haya un proceso de replicación del genoma sin hacer el ciclo división celular completo (Lageix et al. 2007). La proteína NSP por su parte se encarga de transportar moléculas del genoma viral y del virión desde y hacia el núcleo celular.

Cada uno de los componentes indispensables codifica una proteína diferente en el sentido del virión y todos tienen una secuencia con potencial de formar una estructura tallo-asa, que en este caso tiene como nonanucleótido consenso de la región del asa a la secuencia BAKTATT'AC. Un detalle importante es que aunque en una planta infectada se pueden encontrar varios genomas que codifican Reps, para que los otros genomas se multipliquen hace falta un componente Rep-codificante que comparta con ellos las secuencias iteradas asociadas a la estructura tallo-asa; esta proteína Rep es por lo tanto indispensable para el ciclo viral, y se conoce con el nombre de Rep Maestra (Timchenko et al. 2000).

1.4.3. Ciclo infeccioso

Como cada componente codifica para una sola proteína, para que una planta se infecte es necesario el concurso de al menos cinco componentes: el de la proteína Rep, las proteínas Clink y NSP, la proteína de la cápside y la proteína del movimiento (Grigoras et al. 2009, Timchenko et al. 2006). Al igual que en los geminivirus, el ciclo consiste en inoculación de la planta por un insecto vector que contiene en su sistema digestivo viriones adquiridos al alimentarse de otra planta infectada (Oweis et al. 2005). La función de las proteínas codificadas por cada componente indispensable se indica en el pie de la figura 1.5; en general las proteínas tienen funciones muy semejantes a las de los geminivirus, haciendo que la patogénesis proceda de una manera similar, esto es, que al principio el virus manipule el ciclo celular, luego se replique en sus células preferidas y posteriormente se desplace a través de la planta gracias a las proteínas del movimiento.

1.4.4. Problemas que causan

En sus áreas de distribución los nanovirus provocan pérdidas económicas cuya gravedad va asociada a la importancia de la planta hospedera como producto agrícola; por ejemplo, ABTV y BBTV afectan seriamente la producción de fibra de Manila y plátanos de exportación en Filipinas (Sharman et al. 2008), pero no se reporta como problema significativo otros países de su área de distribución. De la misma manera, se han reportado epidemias de FBNYV en Egipto y Siria

(Makkouk & Kumari 2009, Oweis et al. 2005), pero en otros países la enfermedad pasa desapercibida; SCSV no es considerado como una amenaza para las fuentes de forraje en los suelos áridos australianos y las plantas del género *Astragalus*, que se utilizan como plantas medicinales en países orientales por sus propiedades inmuno-estimulantes, diuréticas y anti-cancerígenas, no se cultivan a gran escala (How & Jia 2004).

Aunque estos virus tienen un área de distribución estrecha y un número de hospederos naturales bajo, las infecciones experimentales han demostrado que la cantidad de plantas hospederas puede ser mayor; la interpretación preocupante de éste hecho está asociada con la naturaleza multipartita de sus genomas: la probabilidad de que se junten los cinco componentes indispensables es baja, pero si se crean condiciones en que éstas posibilidades aumenten, como que crezcan las poblaciones de los insectos vectores (por el calentamiento climático, por ejemplo), no sería sorprendente que las especies expandieran su distribución geográfica y/o que se presenten epidemias.

1.5. Generalidades de los Circovirus

1.5.1. Taxonomía y distribución

Todos los miembros de la familia *Circoviridae* tienen genomas monopartitas; la familia se divide en dos géneros: el género *Circovirus* que consiste en doce especies y el género *Gyrovirus* en el que sólo se conoce al *Virus de la anemia del pollo* (CAV), pero que se ha propuesto sea movido a la familia *Anelloviridae* por sus similitudes con el *virus Torque teno* (Hino & Prasetyo 2009). El género *Circovirus* se puede dividir en dos subgrupos, el que infecta mamíferos, al que pertenecen las especies *Circovirus porcino 1* y *2* (PCV1) y (PCV2), respectivamente, y el que infecta aves, al que pertenecen el *virus de la enfermedad del pico y de las plumas de los psitácidos* (BFDV), y los *circovirus de canarios* (CaCV), *de columbidos* (CoCV), *de patos* (DuCV), *de los gorriones* (FiCV), *de gansos* (GoCV), *de gaviotas* (GuCV), *de los cuervos* (RaCV), *de estorninos* (StCV) y *de cisnes* (SwCV) (Halami et al. 2008, Fauquet et al. 2005).

Los circovirus porcinos tienen una distribución mundial, relacionada con las granjas de cría de cerdos, en tanto que en la mayoría de los circovirus de aves sólo se conoce un reporte de la especie, y los otros casos tienen el mismo patrón de distribución del ave hospedera, es decir, éstas especies virales parecen ser muy específicas en cuanto a su huésped.

1.5.2. Organización genómica

Todos los circovirus tienen genomas monopartitas pequeños, de 1.7 a 2.1 Kb en los que se codifican dos proteínas base, la iniciadora de la replicación, y la de la cápside (Halami et al. 2008, Fauquet et al. 2005). Algunos circovirus tienen uno o varios marcos de lectura adicionales, de los cuales solo el ORF C3 de PCV2 ha sido caracterizado funcionalmente; se ha visto que la proteína producto de este ORF promueve la apoptosis en una línea de células epiteliales de riñón de cerdo (Liu et al. 2007). En esta familia la proteína Rep se codifica en el sentido del virión y la CP en sentido complementario.

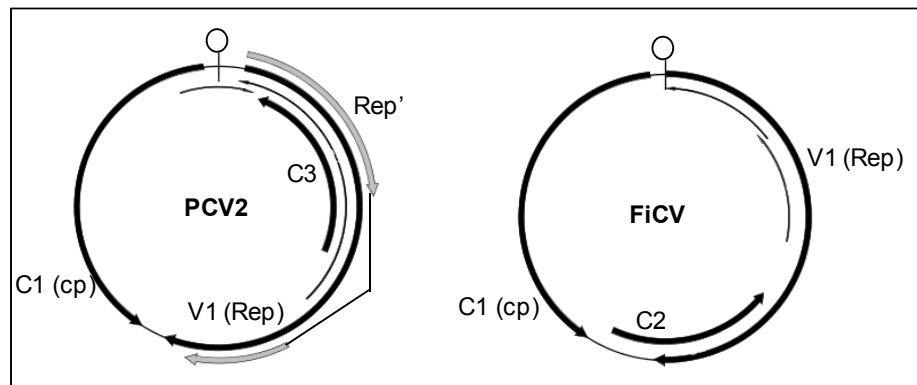


Figura 1.6. Organización genómica de los circovirus. En PCV2 las líneas más delgadas indican marcos de lectura que no se ha confirmado si se transcriben, y la proteína Rep' (168 aa, indicada por flechas grises) se produce mediante el corte de un intrón contenido en el ORF de la proteína Rep (Mankertz & Hillenbrand 2001); en FiCV las líneas más delgadas indican marcos de lectura que generan productos mayores a 80 residuos y que no se reportan en la descripción original del virus.

1.5.3. Ciclo infeccioso

Los circovirus se transmiten de un individuo a otro a través de secreciones corporales. En todos los hospederos las infecciones son más prevalentes en

los juveniles y además parece que en todos los casos las células blanco son las de la línea mononuclear/macrófagos y que la replicación del virus en ellas no es un proceso inocuo, ya que puede inducir apoptosis, causando depleción de la línea linfóide (Finsterbusch & Mankertz 2009, Todd et al. 2007). No hay muchos estudios sobre las funciones adicionales a la replicación y la encapsidación en las proteínas codificadas por estos virus, por lo que el proceso de patogénesis a nivel molecular sigue siendo un misterio. Básicamente solo se especula de la capacidad apoptótica de la proteína C3, no obstante, en un estudio reciente se identificaron una serie de proteínas del tejido del bazo del cerdo que interactúan con Rep (miembros del complejo de “splicing”, varios reguladores transcripcionales y un factor angiogénico) y CP (todas relacionadas con el transporte a través de los microfilamentos, localización nuclear y tráfico por endosomas) (Finsterbusch et al. 2009); el estudio detallado de la interacción de Rep y CP con estas proteínas del hospedero generará información sobre cómo es que se afectan las células linfoides.

1.5.4. Problemas que generan

Los datos epidemiológicos más abundantes sobre esta familia tratan sobre el síndrome de desgaste post-destete de los cerdos, causado solo por PCV2, ya que PCV1 se considera inocuo; se postula que el producto del gen C3 en PCV2 sería el determinante de su patogenicidad porque en PCV1 no hay una proteína homóloga a ésta, aunque los resultados no han sido concluyentes (Finsterbusch & Mankertz 2009). El síndrome consiste en la pérdida progresiva de peso en los cerdos jóvenes, asociada a desórdenes digestivos y respiratorios, y en el tejido linfóide se observa infiltración de macrófagos, formación de sincicios y cuerpos de inclusión. Se considera una enfermedad multifactorial cuya morbilidad puede alcanzar el 50% cuando los cerdos se mantienen en condiciones de estrés y hacinamiento, y la letalidad puede llegar al 90% ya que los individuos entran en condiciones de inmunosupresión y quedan expuestos a otra serie de agentes patógenos.

En los últimos años se han introducido al mercado dos tipos de vacuna contra PCV2, uno de ellos consiste en virus atenuados y el otro en partículas

semejantes a viriones (VLPs, producidas en baculovirus). Estas vacunas están diseñadas para darlas a las hembras de cría, o a los lechones, en ambos casos con el fin de reforzar el sistema inmune. La primera aplicación busca prevenir las infecciones por el virus mediante los anticuerpos maternos y la segunda solo reducir la fuerza de la infección mediante los anticuerpos propios del lechón; en ambos casos se ha visto que la mortalidad se reduce en alrededor del 50%.

En cuanto a la avifauna, en general las infecciones por circovirus cursan con una sintomatología que incluye letargo, depresión y anemia, y que luego progresa a pérdida de peso, distrofia y pérdida de las plumas y deformación del pico y de las uñas (Heath et al. 2004). No se conocen muchos estudios sobre la prevalencia de circovirus en las poblaciones de aves hospedadoras, excepto en el caso del *virus de la enfermedad del pico y de las plumas* (BFDV). Este virus infecta a los psitaciformes (pericos, cacatúas y parientes) y los datos indican que en algunas especies de cotorros la prevalencia en aves en cautiverio puede ser hasta del 8.5% (Bert et al. 2005), mientras que en cacatúas silvestres se han encontrado prevalencias de hasta el 28% (Ha et al. 2007). Dado que varias especies del orden psitaciformes están entre los animales más frecuentemente sacados de sus hábitats naturales e introducidos en todas partes del mundo, esto representa un riesgo sanitario. Debido a éste riesgo y a que de los circovirus de aves el BFDV es el más estudiado, ya se han hecho los primeros esfuerzos por producir versiones recombinantes de la proteína de la cápside de este circovirus para usarse como una vacuna (Bonne et al. 2009), la cual podría aplicarse al menos en los criaderos de aves y exigirse a los comercializadores de aves exóticas.

1.6. Literatura citada

Baliji S, Black MC, French R, Stenger D, Sunter G. 2004. Spinach curly top virus: A newly described Curtovirus species from southwest Texas with incongruent gene phylogenies. *Phytopathology* 94:772-779.

- Bell KE, Dale JL, Ha CV, Vu MT, Reville PA. 2002. Characterisation of Rep-encoding components associated with banana bunchy top nanovirus in Vietnam. *Arch Virol.* 147:695-707.
- Bert E, Tomassone L, Peccati C, Navarrete MG, Sola SC. 2005. Detection of beak and feather disease virus (BFDV) and avian polyomavirus (APV) DNA in psittacine birds in Italy. *J Vet Med B Infect Dis Vet Public Health.* 52(2):64-8.
- Bonfim K, Faria JC, Nogueira EO, Mendes EA, Aragão FJ. 2007. RNAi-mediated resistance to Bean golden mosaic virus in genetically engineered common bean (*Phaseolus vulgaris*). *Mol Plant Microbe Interact.* 20:717-26.
- Bonne N, Shearer P, Sharp M, Clark P, Raidal S. 2009. Assessment of recombinant beak and feather disease virus capsid protein as a vaccine for psittacine beak and feather disease. *J Gen Virol.* 90(Pt 3):640-7.
- Briddon RW, Bedford ID, Tsai JH, Markham PG. 1996. Analysis of the nucleotide sequence of the treehopper-transmitted geminivirus, tomato pseudo-curly top virus, suggests a recombinant origin. *Virology.* 219:387-94.
- Briddon RW, Stanley J. 2006. Subviral agents associated with plant single-stranded DNA viruses. *Virology.* 344:198-210.
- Brown J, Bird J. 1995. Variability within the *Bemisia tabaci* species complex and its relation to new epidemics caused by Geminiviruses. *CEIBA* 36:73-80.
- Brown JK, Idris AM, Fletcher DC. 1993. Sinaloa tomato leaf curl virus, a newly described geminivirus of tomato and pepper in west coastal Mexico. *Plant Dis.* 77:1262.
- Campos-Olivas R, Louis JM, Clerot D, Gronenborn B, Gronenborn AM. 2002. The structure of a replication initiator unites diverse aspects of nucleic acid metabolism. *Proc Natl Acad Sci USA.* 99:10310-5.
- Carter John & Saunders Venetia. *Virology: principles and applications.* John Wiley & Sons Ltd, West Sussex, England, 2007.
- Creamer R, Hubble H, Lewis A. 2005. Curtovirus infection of chile plants in New Mexico. *Plant Disease.* 89:480-486.
- del Solar G, Giraldo R, Ruiz-Echevarria MJ, Espinosa M, Diaz-Orejas R. 1998. Replication and control of circular bacterial plasmids. *Microbiol Mol Biol Rev.* 62:434-64.

- Duffy S, Holmes EC. 2009. Validation of high rates of nucleotide substitution in geminiviruses: phylogenetic evidence from East African cassava mosaic viruses. *J Gen Virol.* 6:1539-47.
- Fauquet CM, Mayo MA, Maniloff J, Desselberger U, Ball LA (eds). *Virus Taxonomy: The eighth report of the international committee on taxonomy of viruses.* 2005. Elsevier/Academic Press. London, UK, pp. 301-326.
- Fauquet CM, Stanley J. 2005. Revising the way we conceive and name viruses below the species level: a review of geminivirus taxonomy calls for new standardized isolate descriptors. *Arch Virol.* 150(10):2151-79.
- Finn RD, Tate JJ, Mistry PC et al. 2008. The Pfam protein families database. *Nucl Ac Res. Database Issue* 36:D281-D288.
- Finsterbusch T, Mankertz A. 2009. Porcine circoviruses--small but powerful. *Virus Res.* 143:177-83.
- Galvez GE, Castaño MJ. 1976. Purification of the whiteflytransmitted bean golden mosaic virus. *Turrialba* 26:205-207.
- Garzón-Tiznado JA, Acosta-García G, Torres-Pacheco I, et al. 2002. Presencia de los geminivirus, huasteco del chile (PHV), texano del chile variante tamaulipas (TPV-T), y chino del tomate (VCdT) en los estados de Guanajuato, Jalisco y San Luis Potosí, México. *Rev Mex Fitopatol.* 20:45-52.
- Gibbs AJ, Fargette D, García-Arenal F, Gibbs MJ. 2010. Time--the emerging dimension of plant virus studies. *J Gen Virol.* 91:13-22.
- Goodman RM. 1977b. A new kind of virus is discovered. *Illinois Research* 19:5.
- Grigoras I, Timchenko T, Katul L, Grande-Pérez A, Vetten HJ, Gronenborn B. 2009. Reconstitution of authentic nanovirus from multiple cloned DNAs. *J Virol.* 83:10778-87.
- Gronenborn B. 2004. Nanoviruses: genome organisation and protein function. *Vet Microbiol.* 98:103-9.
- Guo W, Jiang T, Zhang X, Li G, Zhou X. 2008. Molecular variation of satellite DNA beta molecules associated with *Malvastrum* yellow vein virus and their role in pathogenicity. *Appl Environ Microbiol.* 74:1909-13.
- Gutierrez C. 2000. DNA replication and cell cycle in plants: learning from geminiviruses. *EMBO J.* 19:792-9.
- Ha C, Coombs S, Revill P, Harding R, Vu M, Dale J. 2006. *Corchorus* yellow vein virus, a New World geminivirus from the Old World. *J Gen Virol.* 87:997-1003.

- Ha HJ, Anderson IL, Alley MR, Springett BP, Gartrell BD. 2007. The prevalence of beak and feather disease virus infection in wild populations of parrots and cockatoos in New Zealand. *N Z Vet J.* 55:235-8.
- Halami MY, Nieper H, Müller H, Johne R. 2008. Detection of a novel circovirus in mute swans (*Cygnus olor*) by using nested broad-spectrum PCR. *Virus Res.* 132:208-12.
- Heath L, Martin DP, Warburton L, Perrin M, Horsfield W, Kingsley C, Rybicki EP, Williamson AL. 2004. Evidence of unique genotypes of beak and feather disease virus in southern Africa. *J Virol.* 78:9277-84.
- Hernández-Zepeda C, Idris A M, Carnevali G, Brown JK, Moreno-Valenzuela OA. 2007. Molecular characterization and phylogenetic relationships of two new bipartite begomovirus infecting malvaceous plants in Yucatan, Mexico. *Virus Genes.* 35:369–377.
- Hino S, Prasetyo AA. 2009. Relationship of Torque teno virus to chicken anemia virus. *Curr Top Microbiol Immunol.* 331:117-30.
- Hou SW, Jia JF. 2004. Plant regeneration from protoplasts isolated from embryogenic calli of the forage legume *Astragalus melilotoides* Pall. *Plant Cell Rep.* 22:741-6.
- Ilyina TV, Koonin EV. 1992. Conserved sequence motifs in the initiator proteins for rolling circle DNA replication encoded by diverse replicons from eubacteria, eucaryotes and archaebacteria. *Nucleic Acids Res.* 20:3279-85.
- Khan SA. 2003. DNA-protein interactions during the initiation and termination of plasmid pT181 rolling circle replication. *Prog Nuc Acid Res and Mol Biol.* 75:113-133.
- Khan SA. 2005. Plasmid rolling-circle replication: highlights of two decades. *Plasmid.* 53:26-136.
- Koonin EV, Ilyina TV. 1993. Computer-assisted dissection of rolling circle DNA replication. *Biosystems.* 30:241-68.
- Lageix S, Catrice O, Deragon JM, Gronenborn B, Pélissier T, Ramírez BC. 2007. The nanovirus-encoded Clink protein affects plant cell cycle regulation through interaction with the retinoblastoma-related protein. *J Virol.* 81:4177-85.
- Liu J, Zhu Y, Chen I, Lau J, He F, Lau A, et al. 2007. The ORF3 protein of porcine circovirus type 2 interacts with porcine ubiquitin E3 ligase Pirh2 and facilitates p53 expression in viral infection. *J Virol.* 81:9560-7.
- Lopez-Ochoa L, Ramirez-Prado J, Hanley-Bowdoin L. 2006. Peptide aptamers that bind to a geminivirus replication protein interfere with viral replication in plant cells. *J Virol.* 80:5841-53.

- Makkouk KM, Kumari SG. 2009. Epidemiology and integrated management of persistently transmitted aphid-borne viruses of legume and cereal crops in West Asia and North Africa. *Virus Res.* 141:209-18.
- Mankertz A, Hillenbrand B. 2001. Replication of porcine circovirus type 1 requires two proteins encoded by the viral rep gene. *Virology.* 279:429-38.
- Mansoor S, Zafar Y, Briddon RW. 2006. Geminivirus disease complexes: the threat is spreading. *TRENDS in Plant Science* 11: 209-212.
- Marsin S, Forterre P. 1999. The active site of the rolling circle replication protein Rep75 is involved in site-specific nuclease, ligase and nucleotidyl transferase activities. *Mol Microbiol.* 33:537-45.
- Martin JH, Mifsud D, Rapisarda C. 2000. The whiteflies (Hemiptera: Aleyrodidae) of Europe and the Mediterranean Basin. *Bull Entomol Res.* 90: 407-448.
- Merits A, Fedorkin ON, Guo D, Kalinina NO, Morozov SY. 2000. Activities associated with the putative replication initiation protein of coconut foliar decay virus, a tentative member of the genus Nanovirus. *J Gen Virol.* 81:3099-106.
- Nahid N, Amin I, Mansoor S, Rybicki EP, van der Walt E, Briddon RW. 2008. Two dicot-infecting mastreviruses (family Geminiviridae) occur in Pakistan. *Arch Virol.* 153:1441-51.
- Oweis T, Hachum A, Pala M. 2005. Faba bean productivity under rainfed and supplemental irrigation in northern Syria. *Agric Water Manag.* 73:57-72.
- Padidam M, Beachy RN, Fauquet CM. 1995. Classification and identification of geminiviruses using sequence comparisons. *J Gen Virol.* 76:249-63.
- Padidam M, Sawyer S, Fauquet CM. 1999. Possible emergence of new geminiviruses by frequent recombination. *Virology* 265: 218–225.
- Polston JE, Anderson PK. 1997. The emergence of whitefly-transmitted geminiviruses in tomato in the Western Hemisphere. *Plant Dis.* 81: 1358 – 1369.
- Rampersad SN. 2003. Proposed strategies for begomovirus disease management in tomato in Trinidad. *Plant Health Progress*, October: 1-5.
- Rojas MR, Jiang H, Salati R, Xoconostle-Cázares B, Sudarshana MR, Lucas WJ, Gilbertson RL. 2001. Functional analysis of proteins involved in movement of the monopartite begomovirus, Tomato yellow leaf curl virus. *Virology.* 291:110-25.

- Rosell RC, Torres-Jerez I, Brown JK. 1999. Tracing the geminivirus-whitefly transmission pathway by polymerase chain reaction in whitefly extracts, saliva, hemolymph, and honeydew. *Phytopathology*. 89:239-46.
- Ruiz-Masó JA, Lurz R, Espinosa M, del Solar G. 2007. Interactions between the RepB initiator protein of plasmid pMV158 and two distant DNA regions within the origin of replication. *Nucleic Acids Res*. 35:1230-44.
- Rybicki EP. 1994. A phylogenetic and evolutionary justification for three genera of Geminiviridae. *Arch. Virol*. 139: 49-77.
- Seal SE, van den Bosch F, Jeger MJ. 2006a. Factors influencing Begomovirus evolution and their increasing global significance: Implications for sustainable control. *Crit Rev Plant Sci*. 25:23–46.
- Seal SE, Jeger MJ, van den Bosch F. 2006b. Begomovirus evolution and disease management. *Adv Virus Res*. 67:297-316.
- Shackelton LA, Parrish CR, Truyen U, Holmes EC. 2005. High rate of viral evolution associated with the emergence of carnivore parvovirus. *Proc Natl Acad Sci USA*. 102:379-84.
- Sharman M, Thomas JE, Skabo S, Holton TA. 2008. Abaca´ bunchy top virus, a new member of the genus Babuvirus (family Nanoviridae). *Arch Virol*. 153:135–147.
- Shimada-Beltrán H, Rivera-Bustamante RF. 2007. Early and late gene expression in pepper huasteco yellow vein virus. *J Gen Virol*. 88:3145-53.
- Singh DK, Malik PS, Choudhury NR, Mukherjee SK. 2008. MYMIV replication initiator protein (Rep): roles at the initiation and elongation steps of MYMIV DNA replication. *Virology*. 380:75-83.
- Soler N, Justome A, Quevillon-Cheruel S, Lorieux F, Le Cam E, Marguet E, Forterre P. 2007. The rolling-circle plasmid pTN1 from the hyperthermophilic archaeon *Thermococcus nautilus*. *Mol Microbiol*. 66:357-70.
- Stenger D, McMahon CL. 1997. Genotypic variability of beet curly top virus populations in Western United States. *Phytopathology* 87:737-744.
- Timchenko T, Katul L, Aronson M, Vega-Arreguín JC, Ramirez BC, Vetten HJ, Gronenborn B. 2006. Infectivity of nanovirus DNAs: induction of disease by cloned genome components of Faba bean necrotic yellows virus. *J Gen Virol*. 87:1735-43.
- Timchenko T, Katul L, Sano Y, de Kouchkovsky F, Vetten HJ, Gronenborn B. 2000. The master rep concept in nanovirus replication: identification of missing genome components and potential for natural genetic reassortment. *Virology*. 274:189-95.

- Todd D, Scott AN, Fringuelli E, Shivraprasad HL, Gavier-Widen D, Smyth JA. 2007. Molecular characterization of novel circoviruses from finch and gull. *Avian Pathol.* 36:75-81.
- Torres-Pacheco, J.A. Garzón-Tiznado, J.K. Brown, A. Becerra-Flora, R.F. Rivera-Bustamante. 1996. Detection and distribution of geminiviruses in Mexico and the southern United States. *Phytopathology.* 86:1186-1192.
- van der Walt E, Martin DP, Varsani A, Polston JE, Rybicki EP. 2008. Experimental observations of rapid Maize streak virus evolution reveal a strand-specific nucleotide substitution bias. *Virology J.* 5:104.
- Vanderschuren H, Stupak M, Fütterer J, Grisse W, Zhang P. 2007. Engineering resistance to geminiviruses--review and perspectives. *Plant Biotechnol J.* 5:207-20.
- Varsani A, Shepherd DN, Dent K, Monjane AL, Rybicki EP, Martin DP. 2009. A highly divergent South African geminivirus species illuminates the ancient evolutionary history of this family. *Virology J.* 6:36.

2. Delimitación teórica de los determinantes de especificidad de las proteínas iniciadoras de la replicación por círculo rodante

2.1. Antecedentes

Las proteínas Rep son proteínas multifuncionales que pertenecen a diferentes familias de acuerdo a la distribución de sus dominios funcionales. La familia Gemini_AL1, por ejemplo, incluye a las proteínas Rep de los geminivirus, las cuales tienen actividad de endonucleasa-ligasa en la región N-terminal y su dominio C-terminal posee actividad de helicasa/topoisomerasa (Campos-Olivas et al. 2002). Con base en la arquitectura del dominio endonucleasa, Ilyna & Koonin (1992) agruparon las proteínas Rep en superfamilias caracterizadas por el arreglo de tres motivos conservados, involucrados en la unión y el corte del DNA (Ilyna & Koonin 1992, Koonin & Ilyna 1993). El motivo I (consenso FuTLTxx) parece ser meramente estructural ya que no se le ha asignado una función bioquímica concreta; el motivo II (xpHuHuuux, u= L, I, M, V, Y,F, W, T, A) incluye dos residuos de histidina, separados por un aminoácido no polar, a los cuales se unen cationes divalentes (Mg^{2+} y/o Mn^{2+}), necesarios para la función endonucleolítica de la proteína, y el motivo III (uxxYuxKxx) tiene uno o dos residuos de tirosina que participan directamente en el corte del DNA (Campos-Olivas 2002). Todas las proteínas que poseen los tres motivos conservados, independientemente de la localización del dominio endonucleasa, pertenecen a la superfamilia N1-2-3C descrita por Koonin e Ilyna en su trabajo de 1993.

Se ha acumulado evidencia teórica y experimental que demuestra que existen similitudes en el inicio de la replicación por círculo rodante (RCR) entre los plásmidos de la familia pMV158 y los geminivirus. En ambos sistemas el inicio de RCR, o *dso*, contiene el sustrato endonucleolítico de la proteína Rep, que es una región genómica donde se forma una estructura tallo-asa en la cual

el asa tiene secuencias reconocibles por la proteína. Además en ambos linajes existen una serie de secuencias repetidas adyacentes a la región donde se forma el tallo-asa, las cuales son distintivas de cada especie y son específicamente reconocidas por la proteína Rep afín (Arguello-Astorga et al. 1994, Fontes et al. 1994, Behjatnia & Rezaian 1998, Khan 2005, Ruiz-Masó et al. 2007).

El dominio de la proteína Rep de los geminivirus que está involucrado en la unión al DNA se identificó de manera experimental dentro de la región 1-116 de la proteína (Jupin et al. 1995) y mediante un acercamiento teórico se predijo que el dominio responsable de la especificidad de unión a los repetidos comprendía los primeros 15 residuos de la proteína, precisamente a la izquierda del motivo conservado I (Arguello-Astorga et al. 2001). Esta predicción coincidió con datos experimentales generados por otros grupos (Chatterji et al. 1999, Campos-Olivas 2002, Singh et al. 2008).

En los últimos años ha aumentado el número de replicones CR en las bases de datos, algunos de ellos conformando familias virales recientemente descritas. Con estos nuevos replicones han surgido controversias acerca de su origen y las relaciones filogenéticas entre las proteínas Rep y los genomas que las codifican (Niagro et al. 1998, Gibbs & Weiller 1999, Campos-Olivas 2002). Por ejemplo, las Rep de algunos linajes parecen carecer de uno o dos de los motivos conservados por los iniciadores RCR o tienen una organización atípica (Gibbs et al. 2006) (remitirse a Figura 1.1). Los circovirus y nanovirus que son agentes patógenos de animales (aves y cerdos) y plantas, respectivamente, se cuentan como replicones RCR recientemente conocidos (Todd et al. 2007, Johne et al. 2006, Stewart et al. 2006, Fauquet et al. 2005, Gronenborn 2004). Otros replicones RCR nuevos son los alfa-satelites ó DNAs-1, que se transmiten en asociación con algunos begomovirus monopartitas del Viejo Mundo (Briddon & Stanley 2006).

De los nanovirus y circovirus se sabe que usan el mecanismo de CR para multiplicar sus genomas (Timchenko et al. 1999, Steinfeldt et al. 2001, Cheung 2004, Gronenborn 2004) y que comparten con los geminivirus un *dso* similar,

estos es, con una región capaz de generar una estructura tallo-asa y con secuencias repetidas asociadas a ésta (Steinfeldt et al. 2006, Herrera-Valencia et al. 2006). Los DNAs¹ se consideran replicones CR porque poseen una región con las propiedades del *dso*, pero las formas replicativas intermedias características de la RCR no se han observado experimentalmente.

En los geminivirus las secuencias repetidas asociadas a la región que forma el tallo-asa se conocen como iterones. La forma como son reconocidos estos iterones sigue siendo un motivo de investigación, ya que se busca reconocer a los aminoácidos de la proteína Rep responsables del reconocimiento específico de una secuencia repetida determinada. Por las similitudes entre los geminivirus y los nuevos replicones RCR mencionados arriba, los datos al respecto que se obtengan de cualquiera de estos linajes pueden servir para establecer qué propiedades de la interacción DNA-proteína Rep son útiles a la hora de diseñar estrategias de control de éstos patógenos (Vanderschuren et al. 2007). En este trabajo se identifican los aminoácidos de la proteína Rep que determinan su capacidad para reconocer las secuencias repetidas particulares del *dso* de las especies de nanovirus, circovirus y alfa-satélites mediante un enfoque teórico. La identificación de estos residuos mejora significativamente el entendimiento del mecanismo replicativo usado por los virus de estos linajes y sirve para guiar experimentos de mutagénesis sitio-dirigida para la caracterización de la proteína Rep.

Aunque en la literatura se reportan varias formas de evaluar la importancia de un aminoácido en la unión a una secuencia nucleotídica, ninguno de ellos resultó adecuado para hacer una definición extensiva de los residuos determinantes de la especificidad de unión al DNA en las proteínas iniciadoras de CR. Los métodos experimentales, que incluyen las mutaciones sitio-dirigidas y la reconstrucción tridimensional de los cristales obtenidos de complejos proteína-DNA fueron descartados por ser costosos y consumir mucho tiempo. Los métodos computacionales, por ejemplo aquellos que predicen los sitios de unión a DNA en proteínas, tienen la desventaja de que dependen de la disponibilidad de una estructura tridimensional de la proteína unida a su DNA cognado, o de que la secuencia de unión sea conservada, ya que los

algoritmos más comunes para identificar sitios de unión a DNA son aquellos basados en las huellas filogenéticas (Wu et al. 2009); ésta última característica de los predictores de sitios de unión a DNA los hace poco indicados para los fines de este trabajo, ya que por lo que se conoce de los geminivirus, se espera que los iterones de los nuevos replicones CR varíen según la especie, al igual que los residuos que los reconocen de manera específica. Para lidiar con el problema de la variabilidad esperada, en este trabajo se usa una estrategia que inicia con el supuesto heurístico de que los replicones que tienen la misma secuencia repetida comparten aminoácidos en una región de la proteína, que son los que determinantes de la especificidad.

2.2. Material y métodos

Los datos utilizados en este trabajo consisten en la secuencia nucleotídica de los replicones de círculo rodante pertenecientes a cuatro linajes virales: begomovirus y alfa-satélites, nanovirus y circovirus. Las secuencias se bajaron de la base de datos GenBank hasta el 15 de Agosto de 2009 y se usaron para hacer varios análisis *in silico*, tanto a nivel de la secuencia nucleotídica del DNA, como de la secuencia de aminoácidos de las proteínas Rep.

Para identificar los dominios de unión al DNA en las proteínas Rep, y con base en lo que se conoce de los sistemas RCR mas estudiados, se asumió lo siguiente: 1) Que los iterones son las secuencias de DNA específicamente reconocidas por la proteína Rep. 2) Que el dominio endonucleasa, el cual posee los motivos conservados de los iniciadores RCR de la Superfamilia Rep1-2-3, posee un dominio discreto de unión al DNA en el cual algunos residuos son responsables del reconocimiento específico de las secuencias iteradas y pueden llamarse Determinantes de Especificidad (DEs). 3) Que en un linaje pueden existir varias especies virales que poseen iterones con la misma secuencia, las cuales constituyen conjuntos de replicones con la misma especificidad replicativa; tales conjuntos de aquí en adelante se llamarán grupos de especificidad o grupos iso-específicos.

Bajo estas suposiciones se plantearon las siguientes hipótesis de trabajo:

- 1) Las proteínas Rep de los virus del mismo grupo de especificidad contienen residuos similares en el dominio involucrado en la unión a los iterones, independientemente de su distancia evolutiva, y 2) Las proteínas de diferentes grupos de especificidad divergen en secuencia en ese mismo dominio, aún cuando sean muy similares de manera global. A partir de esas hipótesis heurísticas se desarrolló una estrategia de trabajo que comprende tres pasos generales que son: a) Identificación de los iterones en cada uno de los replicones identificados, conformación de los grupos de especificidad y clasificación de las proteínas Rep de éstos replicones de acuerdo al grupo iso-específico, b) aplicación de un método comparativo de análisis de secuencias de proteína, y c) reforzamiento de los resultados arrojados por el método comparativo.

2.2.1. Determinación de las características de los orígenes de replicación y creación de los grupos iso-específicos

El primer paso para lograr el objetivo indicado arriba fue la identificación de los elementos de inicio de replicación en cada una de las secuencias a analizar. Para esto, se requiere identificar la secuencia que forma el tallo-asa del *dso* como elemento de posicionamiento, y a partir de allí se identifican otros elementos conservados como la caja TATA del gen *rep* y el inicio de transcripción del mismo. Éstos tres primeros elementos se pueden identificar en varias secuencias a la vez ya que tienen un consenso que se puede buscar incluso mediante una búsqueda simple en un texto. Posteriormente se identifican los elementos iterados, que se definen como secuencias de cinco a ocho nucleótidos que se encuentran repetidas en forma directa o invertida y pueden estar localizados tanto a la izquierda como a la derecha del tallo-asa. Los iterones difícilmente pueden identificarse mediante programas computacionales ya que su característica es variar entre las diferentes especies, y pueden hacerlo incluso en cuanto a posición relativa, como se puede observar en los ejemplos que se muestran en las figuras 2.1 y 2.6.

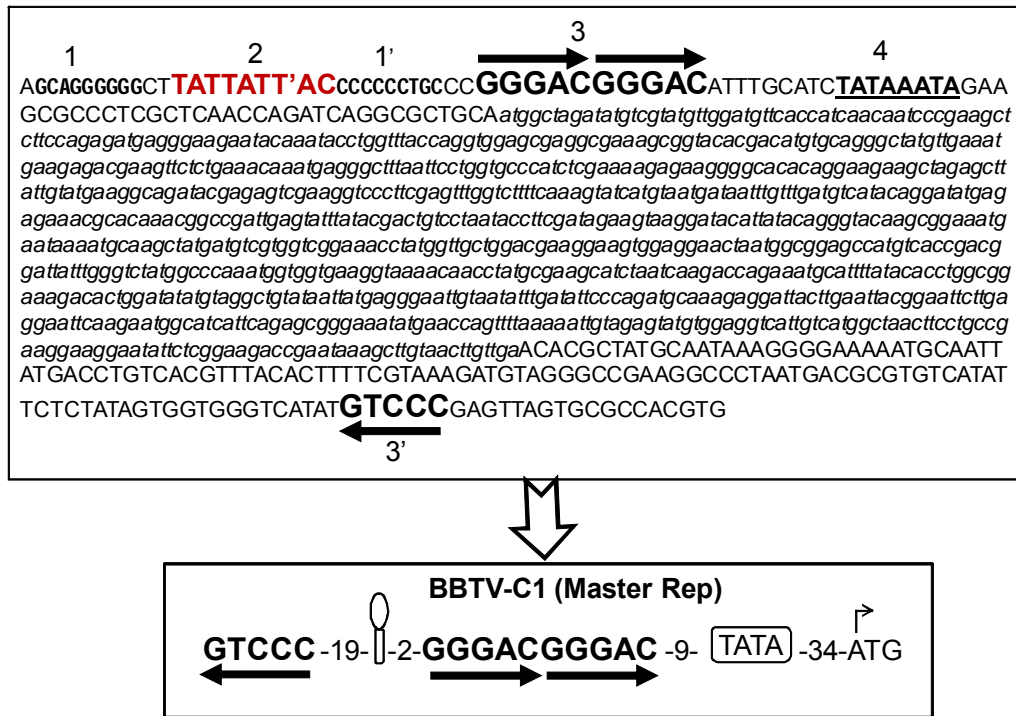


Figura 2.1. Ejemplo de la identificación del origen de replicación en un genoma nanoviral. 1 y 1') secuencia del tallo en donde 1 y 1' son complementarios entre sí; 2) nonanucleótido conservado (sitio de corte de Rep), 3 y 3') iterones, note que 3' es el elemento invertido; 4) caja TATA. En itálicas se muestra el marco de lectura de la proteína Rep. El cuadro de abajo es una representación simple de los elementos identificados.

Una vez que se identificó la organización de iterones de cada una de las secuencias, se pasó a clasificarlas en conjuntos de replicones de un mismo linaje que poseen el mismo arreglo y secuencia de iterones. Los conjuntos obtenidos se enlistan en la tabla 2, junto con algunas estadísticas sobre la cantidad de replicones que había en cada grupo. Cada uno de los replicones en un grupo de iso-especificidad presumiblemente codifica una proteína Rep con la misma especificidad de unión al DNA que los demás miembros de su grupo.

Tabla 2. Listado de grupos iso-específicos

	DNAs-1		Nanovirus		Circovirus	
Genomas analizados	90		46		12	
Genomas codif. Rep	90		22		12	
Arreglos de iterones*	3		5		4	
Iso-grupos	Secuencia	n	Secuencia	n	Secuencia	n
	1)GGMACCC	23	1)TGAC-TCAG [†]	17	1)GGGGCAC	2
	2)GGWTCCC	21	2)GGGAC [‡]	14	2)GGGGCCAT	1
	3)GAGACCC	12	3) CTCCCCCT	1	3)GGAGCCAC	4
	4)CGACCCT	4	4) CTMMCCCC	1	4)GGAACCAC	1
	5)GGCTACC	2	5) GGMGCCC	1	5)GTACTCC	2
	6)TAGACCC	3	6) TCATCCCT	1	6)STACTAC	1
	7)CGTGCTCT	1	7) GTGCTCCC	1	7)CGGCAG	2
	8)TGTCCCCT	1	8) GTTACAC	2		
	9)TGGCCCCT	1	9) GGAACAC	1		
	10)TCCACAC	1	10) CCTCGCCCT	2		
			11) CGCTTCCC	1		
			12) CCTCGGAAC	1		
			13) CCTCCGCGC	1		
			14) TGCTAA [§]	1		
			15) CCTTGGA	1		

*Se refiere a la variación en posición relativa, orientación y número de los repetidos

[†]Iterones de los componentes indispensables de FBNYV, MVDV y SCSV

[‡]Iterones de los componentes indispensable de ABTV y BBTV

[§]Secuencia iterada del nanovirus monopartita CFDV

2.2.2. Identificación de residuos determinantes de especificidad mediante métodos comparativos de secuencias de proteína

La identificación de dominios de unión a DNA en las proteínas puede hacerse por diferentes metodologías dependiendo de la cantidad de información que se conozca sobre la proteína implicada. Cuando se carece de información cristalográfica o de resonancia magnética nuclear del complejo DNA-proteína, el tipo de análisis a *grosso modo* que procede es el análisis de secuencias en busca de dominios de unión a DNA. De los estudios de sistemas RCR bien caracterizados se ha llegado a establecer que las proteínas Rep carecen de los motivos de unión al DNA más comunes entre los reguladores transcripcionales y otros reguladores nucleares, que son los dedos de zinc y los dominios hélice-vuelta-hélice y por lo tanto descartamos la posibilidad de identificar el dominio

de unión específico al DNA en los iniciadores RCR mediante una búsqueda de consensos de dominios clásicos de unión a DNA.

Por otra parte, en nuestro caso el método comparativo debía ajustarse a la necesidad de identificar un dominio de unión a secuencias de DNA variables. Se usó entonces una estrategia comparativa que detecta los residuos conservados entre iso-grupos de proteínas homólogas y los compara con los residuos que comparten los otros iso-grupos, la cual se aplicó con dos enfoques alternativos, dependiendo de la abundancia de proteínas Rep no redundantes en cada linaje:

1)Análisis Comparado de Grupos de Proteínas Homólogas iso-específicas (CAGHIP)

Este enfoque parte de nuestra segunda hipótesis heurística *-Las proteínas Rep de diferentes grupos de especificidad divergen en secuencia en el dominio que está implicado en la especificidad de unión a DNA, aún cuando sean muy similares de manera global-*, y luego pasa a la primera hipótesis de trabajo *- Las proteínas del mismo grupo de especificidad contienen residuos similares en el dominio involucrado en la unión a los iterones-*.

El análisis se desarrolla mediante comparaciones secuenciales, como se describe en las figuras 2.2, 2.3 y 2.4; el primer paso es comparar proteínas muy similares que pertenecen a distintos grupos iso-específicos para detectar los residuos diferenciales, los cuales luego se contrastan contra la variabilidad interna de cada grupo de especificidad, permitiendo así una estrategia de descarte en la que los residuos con mayor probabilidad de ser determinantes de la especificidad son aquellos que se comparten entre el mismo grupo, pero divergen con respecto a otro conjunto de especificidad.

Iteron GGM ACCC		vs	Iteron CGT GCTCT			
1) ToYLCCV-DNA1 (AJ888449)			2) MiYLCV-DNA1 (DQ641719)			
1) MPSV	T	SVFWCFTVFFTSATAPDLVPVFENTHVSYACWQEEESPTTKRRHLQGYLQLKG	K	RTLNQVK	SL	F
	*		*		**	
2) MPSV	A	SVFWCFTVFFTSATAPDLVPVFENTHVSYACWQEEESPTTKRRHLQGYLQLKG	R	RTLNQVK	AI	F
70	GDLKPHLEKQRARKTDEA	C	DYCMKEETRVSGPFEGDYCP	SGSHRRRQRESVIRSPVRM	S	E 130
	*		*		*	
70	GDLKPHLEKQRARKTDEA	R	DYCMKEETRVSGPFEGDYCP	SGSHKRRQRESVIRSPVRM	A	E 130

Figura 2.2. Paso 1 del Método CAGHIP. Se compara la secuencia completa del dominio catalítico de las dos proteínas con mayor porcentaje de identidad entre dos grupos iso-específicos y las diferencias entre el par representan posiciones candidatas a ser el determinante de especificidad. En esta figura el alfa-satélite asociado al *Virus del enchinamiento de la hoja del Tomate de China* con número de acceso AJ888449 es un miembro del iso-grupo con iterones GGMACC y su dominio endonucleasa difiere en seis residuos del equivalente en la proteína Rep del alfa-satélite asociado al *Virus del enchinamiento y amarillamiento de la hoja de la Mimosa* (con secuencia de iterones CGTGCTCT y acceso DQ641719), los cuales se marcan con un asterisco y están encerrados en un recuadro.

Iteron GGM ACCC		vs	Iteron CGT GCTCT				
1) TbCSV-DNA1 (AJ579346)			3) MiYLCV-DNA1 (DQ641719)				
2) ToYLCCV-DNA1 (AJ888449)							
1)	#	T	SVFWCFTVFFTSATAPDLVPVFENTHVSYACWQEEESPTTKRRHLQGYLQLKG	K	RTLNQVK	AI	F
2) MPSV		T		*	*	**	
3) MPSV		A	SVFWCFTVFFTSATAPDLVPVFENTHVSYACWQEEESPTTKRRHLQGYLQLKG	R	RTLNQVK	AI	F
70	GDLKPHLEKQRARKTDEA	R	DYCMKEETRVSGPFEGDYCP	SGSHRRRQRESVIRSPVRM	S	E 130	
	*		*		*		
70	GDLKPHLEKQRARKTDEA	R	DYCMKEETRVSGPFEGDYCP	SGSHKRRQRESVIRSPVRM	A	E 130	

Figura 2.3. Paso 2 del método CAGHIP. Para descartar algunas de las seis posiciones candidatas del ejemplo anterior se agrega una segunda proteína perteneciente al grupo iso-específico GGMACCC, y se analizan solo los residuos de interés. Las posiciones candidatas en las que el mismo residuo ocurre en el grupo iso-específico opuesto quedan descartadas, y las posiciones que quedan (aquí marcadas con #) corresponden al probable determinante de especificidad.

La aplicabilidad de este enfoque depende de la variabilidad interna de cada grupo de especificidad replicativa y pudo aplicarse sólo entre los alfa-satélites, ya que en éste grupo se contaba con más de 90 proteínas Rep distintas y la

probables (marcadas con #) tras el descarte de posiciones por las estrategias ya indicadas, pero hay una posición (marcada con ':') que no puede ser descartada porque varía en ambos grupos.

Posiciones como la que no se puede descartar en el ejemplo ocurren cuando hay pocos representantes no-redundantes de al menos uno de los iso-grupos (en el ejemplo sólo se contaba con la secuencia de cuatro DNAs¹ que unen CGACCCT y dos de ellos no difieren en los residuos de interés), y esto genera posiciones inciertas y mapeos falsos.

2)Enfoque alternativo a CAGHIP (cuando la variabilidad entre proteínas es alta, y el número de miembros no-redundantes en cada grupo de especificidad es bajo).

Este fue el enfoque que se aplicó para los análisis comparativos de las proteínas Rep de los nanovirus y los circovirus. Aquí las comparaciones se hicieron para buscar regiones conservadas entre los miembros del mismo grupo de especificidad, es decir, sólo se hace uso de la hipótesis de trabajo 1, y lo que se espera encontrar son “dominios convergentes” entre proteínas poco similares entre sí. Los datos que se obtienen son secciones de la proteína compartidas entre los miembros de cada iso-grupo, las cuales son candidatas a contener los residuos que determinan la especificidad. Los dos casos donde hay más de dos proteínas en el mismo grupo de especificidad se muestran en la figura 2.6.

2.2.3 Robustecimiento de los resultados del análisis comparativo

Para saber si las posiciones mapeadas en las comparaciones tienen algún significado biológico los resultados se analizan en un contexto estructural y de acuerdo a lo que se conoce de los otros replicones CR. De esta manera, los residuos o motivos que mapean como posibles DEs se analizan considerando su ubicación con respecto a la de los motivos conservados del dominio endonucleasa de los iniciadores de RCR y posteriormente considerando su ubicación con respecto a estructuras secundarias en modelos tridimensionales de la proteína. Como un ejemplo en la figura 2.6 se indican las regiones del

dominio endonucleasa en donde con mayor frecuencia se mapearon potenciales DEs en los alfa-satélites.

1)MaYMV	## #		#			#					
2)SiLCV	MPSLKSTFWC	F	T	V	F	F	T	A	S	A	P
	** *	*	**	*	*	*	*	*	*	*	*
3)ToYLCCV	MPSITSVFWC	F	T	I	F	F	A	S	S	A	P
4)ToYLCTV	VT V	V	T	A	V						
5)MaYMV	# #		##			#					
6)SiLCV	MPSLKSTFWC	F	T	V	F	F	T	A	S	A	P
	* **	*	**	*	*	*	*	*	*	*	*
7)SiLCV	MPALKAQWNC	F	T	V	F	F	L	S	S	A	P
8)OkLCV	A SHW	L	S	T							
9)ToYLCTV	#####		#			#					
10)TbLCYV	PSVTSVF		T	T							
	*****	*	*	*	*	*	*	*	*	*	*
11)ACMV	MAALKGQWNC	F	T	I	F	F	L	S	A	P	D
12)AYVV	PALRGQW	L	T								
13)TbCSV	# #					#					
14)ToYLCCV	MPSVTSVFWC	F	T	V	F	F	T	S	A	P	D
	* **	*	**	*	*	*	*	*	*	*	*
15)ToYLCCV	MPCVQSQWNC	F	T	V	F	F	L	T	A	P	D
16)CLCuMV	Q QW	L	S	L	F						

Figura 2.5. Localización de los residuos que mapean como determinantes de especificidad con respecto a los Motivos conservados I y II (sombreados en verde) en cuatro comparaciones del método CAGHIP. Note que la constante es que en todos los casos se mapean residuos a la izquierda del Motivo I y a la derecha del Motivo II; las figuras 2.3 y 2.4 pueden proporcionar ejemplos adicionales.

La información conocida con anterioridad, más los datos preliminares, pueden ser usados para hacer nuevas hipótesis que permitan completar el análisis. Así, como en los alfa-satélites se detectaron dos regiones donde se concentran los residuos que mapean como DEs (una al lado izquierdo del Motivo I y otra al lado derecho del Motivo II), y la primera de ellas coincidió con lo reportado para los geminivirus (Singh et al. 2008, Campos-Olivas 2002, Arguello-Astorga et al. 2001, Chatterji et al. 1999), al asumir que la proteína Rep de los otros linajes de replicones CR que se están analizando se comportan de manera similar, se pueden identificar los dominios convergentes de los grupos iso-específicos de los nanovirus y circovirus con mayor probabilidad de estar involucrados en la unión a DNA.

indicadas por flechas en línea discontinua porque tienden a ser muy cortos y en un arreglo degenerado; en estas especies no se ha estudiado a nivel experimental la relevancia de los distintos elementos repetidos.

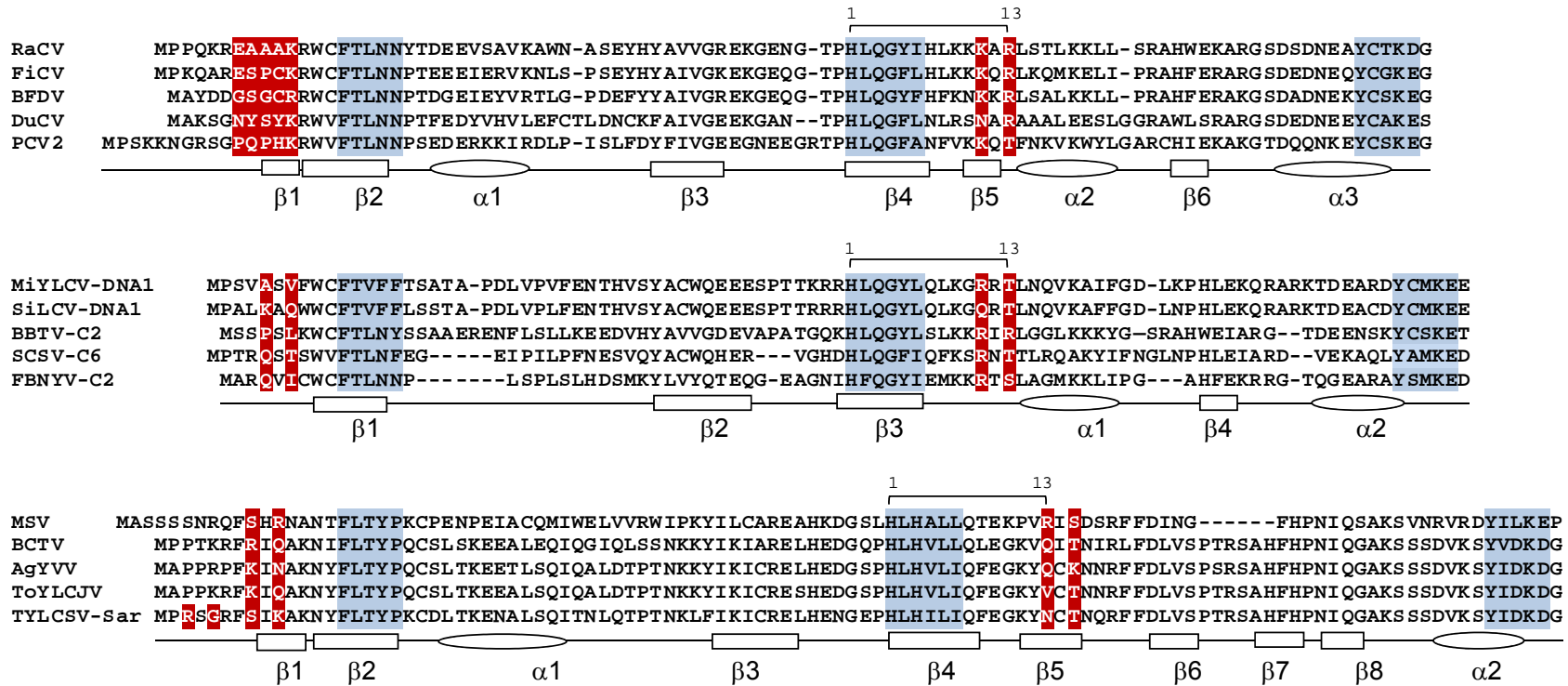


Figura 2.7. Una constante en la localización de los dominios que mapean como regiones con determinantes de especificidad confirma su significado biológico.

En los ejemplos de la figura 2.6 se muestra que hay varias regiones compartidas entre las proteínas Rep de de cada grupo iso-específico en nanovirus y circovirus. Al buscar similitudes con los alfa-satélites se encontró que en los nanovirus y los circovirus hay una región de la proteína Rep adyacente al Motivo I que siempre está compartida entre los miembros del mismo grupo iso-específico. Una segunda región al lado del Motivo II también estaba compartida en los grupos, aunque de extensión más corta. Para establecer si alguna de estas regiones contenía DEs, se consideró con más detalle su posición con respecto a los motivos I, II y los que no se ajustaban a los datos obtenidos previamente para los alfa-satélites y los geminivirus se descartaron.

La figura 2.7 contiene el resultado del proceso de selección de dominios “convergentes” dentro de los iso-grupos que son candidatos a poseer DEs. En esta figura se puede observar que la región asociada al motivo I de los nanovirus y circovirus que converge apenas se desvía en uno o dos residuos con respecto a la localización de los DEs previamente reportados en los geminivirus y de los identificados aquí para los alfa-satélites. En algunos casos las comparaciones permitieron proponer a dos aminoácidos del dominio convergente como los DEs, ya que la combinación de dichos residuos se encontraba exclusivamente en un grupo iso-específico. De la misma manera la secuencia aminoacídica asociada al motivo II resultó estar a la misma distancia entre todos los linajes analizados (incluyendo a los geminivirus, que fueron analizados en esta región tras los hallazgos en los alfa-satélites), y en ella también se encontraron combinaciones de aminoácidos que por su presencia en ciertos grupos de especificidad probablemente sean parte del conjunto de los aminoácidos que provocan la especificidad de reconocimiento del iterón.

El otro tipo de evidencia que se usó para reforzar las conclusiones derivadas de los datos obtenidos, y que permite confiar en que la segunda región con DEs es igualmente significativa, es la localización de las regiones mapeadas en un modelo tridimensional de las proteínas. En la figura 2.8 se muestran un par de ejemplos de éste tipo de evidencia y se observa como las regiones adyacentes a los Motivos I y II quedan físicamente cercanas en

modelos tridimensionales de las proteínas Rep realizados mediante modelado estructural con los datos obtenidos para representantes de tres familias de virus con ssDNA que se replican por círculo rodante (Vega-Rocha et al. 2007a, Vega-Rocha et al. 2007b, Campos-Olivas et al. 2002).

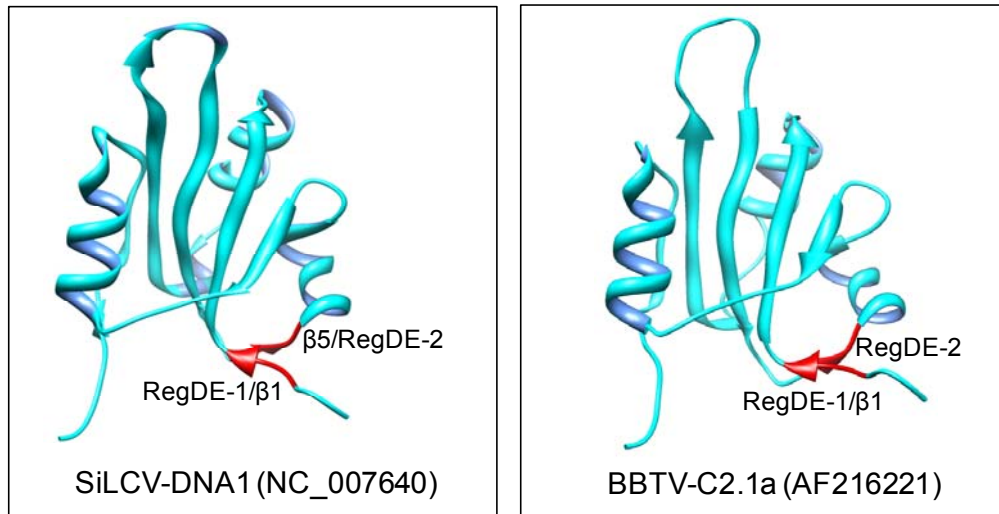


Figura 2.8. Localización de las regiones que contienen determinantes de especificidad en el modelo de la proteína Rep de un alfa-satélite y de un nanovirus, hechos con las herramientas del programa Swiss Model.

2.3. Resultados

La comparación de proteínas con diferentes especificidades de unión a secuencias iteradas mediante un análisis heurístico permitió delimitar de manera teórica los residuos aminoácidos que confieren la especificidad de unión al DNA en la proteína iniciadora de la replicación de los miembros de la familia *Nanoviridae*, del grupo de satélites auto-replicativos asociados a los geminivirus del Viejo Mundo (DNAs1 o alfa-satélites) y de los circovirus. Los detalles de los resultados obtenidos en este trabajo están contenidos en un artículo que fue aceptado para su publicación en la revista *Archives of Virology* y cuyo contenido en extenso se encuentra en el Anexo 2. La conclusión general a la que se llegó es que las proteínas iniciadoras de círculo rodante de estos tres linajes comparten entre sí, y con los geminivirus, más características de lo

que antes se pensó, lo cual además sugiere fuertemente que los virus de ssDNA que se replican por éste mecanismo tienen un origen común.

2.4. Referencias

- Arguello-Astorga GR, Guevara-Gonzalez RG, Herrera-Estrella LR, Rivera-Bustamante RF. 1994. Geminivirus replication origins have a group-specific organization of iterative elements: a model for replication. *Virology*. 203:90-100.
- Arguello-Astorga GR, Ruiz-Medrano R. 2001. An iteron-related domain is associated to Motif 1 in the replication proteins of geminiviruses: identification of potential interacting amino acid-base pairs by a comparative approach. *Arch Virol*. 146:1465-85.
- Behjatnia SAA, Dry IB, Rezaian MA. 1998. Identification of the replication-associated protein binding domain within the intergenic region of tomato leaf curl geminivirus. *Nucleic Acids Research* 26 :925-931
- Bridson RW, Stanley J. 2006. Subviral agents associated with plant single-stranded DNA viruses. *Virology*. 344:198-210.
- Campos-Olivas R, Louis JM, Clerot D, Gronenborn B, Gronenborn AM. 2002. The structure of a replication initiator unites diverse aspects of nucleic acid metabolism. *Proc Natl Acad Sci U S A*. 99:10310-5.
- Chatterji A, Padidam M, Beachy RN, Fauquet CM. 1999. Identification of replication specificity determinants in two strains of tomato leaf curl virus from New Delhi. *J Virol*. 73:5481-9.
- Cheung AK .2004. Palindrome regeneration by template strand-switching mechanism at the origin of DNA replication of porcine circovirus via the rolling-circle melting-pot replication model. *J Virol*. 78:9016-29.
- Eagle PA, Hanley-Bowdoin L. 1994. cis elements that contribute to geminivirus transcriptional regulation and the efficiency of DNA replication. *J Virol*. 71:6947-55.
- Fauquet CM, Mayo MA, Maniloff J, Desselberger U, Ball LA (eds) (2005). *Virus taxonomy. Classification and nomenclature of viruses*. 8th ICTV Report, Academic Press, Elsevier, 1217 pages.
- Fontes EP, Gladfelter HJ, Schaffer RL, Petty IT, Hanley-Bowdoin L. 1994. Geminivirus replication origins have a modular organization. *Plant Cell*. 6:405-16.

- Gibbs MJ, Smeianov VV, Steele JL, Upcroft P, Efimov BA. 2006. Two families of rep-like genes that probably originated by interspecies recombination are represented in viral, plasmid, bacterial, and parasitic protozoan genomes. *Mol Biol Evol.* 23:1097-100.
- Gronenborn B. 2004. Nanoviruses: genome organisation and protein function. *Vet Microbiol.* 98:103-9.
- Herrera-Valencia VA, Dugdale B, Harding RM, Dale JL. 2006. An iterated sequence in the genome of Banana bunchy top virus is essential for efficient replication *J Gen Virol.* 87(Pt 11):3409-12.
- Ilyina TV, Koonin EV. 1992. Conserved sequence motifs in the initiator proteins for rolling circle DNA replication encoded by diverse replicons from eubacteria, eucaryotes and archaeobacteria. *Nucleic Acids Res.* 20:3279-85.
- Johne R, Fernandez-de-Luco D, Hofle U, Muller H. 2006. Genome of a novel circovirus of starlings, amplified by multiply primed rolling-circle amplification. *J Gen Virol.* 87:1189-95.
- Jupin I, Hericourt F, Benz B, Gronenborn B. 1995. DNA replication specificity of TYLCV geminivirus is mediated by the amino-terminal 116 amino acids of the Rep protein *FEBS Lett.* 362:116-20.
- Khan SA. 2005. Plasmid rolling-circle replication: highlights of two decades *Plasmid.* 53:126-136.
- Koonin EV, Ilyina TV. 1993. Computer-assisted dissection of rolling circle DNA replication. *Biosystems.* 30:241-68.
- Niagro FD, Forsthoefel AN, Lawther RP, Kamalanathan L, Ritchie BW, Latimer KS, Lukert PD. 1998. Beak and feather disease virus and porcine circovirus genomes: intermediates between the geminiviruses and plant circoviruses. *Arch Virol.* 143:1723-44.
- Ruiz-Masó JA, Lurz R, Espinosa M, del Solar G. 2007. Interactions between the RepB initiator protein of plasmid pMV158 and two distant DNA regions within the origin of replication. *Nucleic Acids Res.* 35:1230-44.
- Singh DK, Malik PS, Choudhury NR, Mukherjee SK. 2008. MYMIV replication initiator protein (Rep): roles at the initiation and elongation steps of MYMIV DNA replication. *Virology.* 380:75-83.
- Steinfeldt T, Finsterbusch T, Mankertz A. 2001. Rep and Rep' protein of porcine circovirus type 1 bind to the origin of replication in vitro. *Virology.* 291:152-60.

- Steinfeldt T, Finsterbusch T, Mankertz A. 2006. Demonstration of nicking/joining activity at the origin of DNA replication associated with the rep and rep' proteins of porcine circovirus type 1. *J Virol.* 80:6225-34.
- Stewart ME, Perry R, Raidal SR 2006. Identification of a novel circovirus in Australian ravens (*Corvus coronoides*) with feather disease. *Avian Pathol.* 35:86-92.
- Timchenko T, de Kouchkovsky F, Katul L, David C, Vetten HJ, Gronenborn B. 1999. A single rep protein initiates replication of multiple genome components of faba bean necrotic yellows virus, a single-stranded DNA virus of plants. *J Virol.* 73:10173-82.
- Timchenko T, Katul L, Sano Y, de Kouchkovsky F, Vetten HJ, Gronenborn B. 2000. The master rep concept in nanovirus replication: identification of missing genome components and potential for natural genetic reassortment. *Virology.* 274:189-95.
- Todd D, Scott AN, Fringuelli E, Shivradas HL, Gavier-Widen D, Smyth JA. 2007. Molecular characterization of novel circoviruses from finch and gull. *Avian Pathol.* 36:75-81.
- Vanderschuren H, Stupak M, Fütterer J, GUISSEM W, Zhang P. 2007. Engineering resistance to geminiviruses--review and perspectives. *Plant Biotechnol J.* 5:207-20.
- Vega-Rocha S, Byeon IJ, Gronenborn B, Gronenborn AM, Campos-Olivas R. 2007a. Solution structure, divalent metal and DNA binding of the endonuclease domain from the replication initiation protein from porcine circovirus 2. *J Mol Biol.* 367:473-87.
- Vega-Rocha S, Gronenborn B, Gronenborn AM, Campos-Olivas R. 2007b. Solution structure of the endonuclease domain from the master replication initiator protein of the nanovirus faba bean necrotic yellows virus and comparison with the corresponding geminivirus and circovirus structures. *Biochemistry.* 46:6201-12.
- Wu J, Liu H, Duan X, Ding Y, Wu H, Bai Y, Sun X. 2009. Prediction of DNA-binding residues in proteins from amino acid sequences using a random forest model with a hybrid feature. *Bioinformatics* 25:30-5.

3. Historia evolutiva del género curtovirus

3.1. Antecedentes

Los curtovirus constituyen uno de los cuatro géneros de virus de DNA de cadena sencilla que integran la familia *Geminiviridae*. Los miembros de éste género se caracterizan por poseer un genoma monopartita de aproximadamente 3000 pb, por ser transmitidos por chicharritas del género *Circulifer* (*Cicadellidae*) y por infectar un amplio rango de plantas dicotiledóneas, en las cuales causan enanismo, amarillamiento y deformación de las hojas, puntas rizadas y en algunos casos crecimientos anómalos en la superficie de las hojas (enaciones) (Stanley et al. 2005, Creamer et al. 2003, Bennett 1971). Se trata de un género poco diverso, compuesto por cinco especies reconocidas por el Comité Internacional de Taxonomía de Virus (ICTV), que son: *Beet curly top virus* (BCTV), *Beet mild curly top virus* (BMCTV), *Beet severe curly top virus* (BSCTV), *Horseradish curly top virus* (HrCTV) y *Spinach curly top virus* (SCTV) (Fauquet et al. 2008).

La enfermedad y las especies virales asociadas fueron descritas por primera vez en Estados Unidos de Norteamérica, pero casi al mismo tiempo que se detectaron los primeros miembros del género en ésta zona, se identificaron entidades virales similares en la región del Medio Oriente (Baliji et al. 2004, Bennet 1971); desde entonces se han hecho varias especulaciones acerca de las relaciones entre los curtovirus de los dos continentes, las cuales tratan de resolver la cuestión de si la enfermedad del rizado de las puntas del betabel se introdujo del Viejo Mundo a las Américas o viceversa.

La relación evolutiva entre los curtovirus y los demás géneros de la familia *Geminiviridae* no está completamente clara. En las reconstrucciones de la filogenia de la familia los curtovirus conforman un grupo intermedio entre los begomovirus y los mastrevirus (Varsani et al. 2009, Fauquet & Stanley 2003). Por su organización genómica y su secuencia nucleotídica estos virus tienen la

mitad de su genoma (la parte correspondiente a los cuatro genes en sentido complementario, ver figura 1.3 y/o 3.1) claramente relacionada al genoma de los begomovirus (con la excepción de HrCTV que solo se asemeja a estos en la primera porción de Rep) (Baliji et al. 2007, Klute et al. 1996). En la otra mitad tienen tres genes de los cuales solo el de la proteína de la cápside es homólogo a los de los demás géneros (Baliji et al. 2004, Klute et al. 1996, Rybicki 1994, Hormuzdi et al. 1993).

Las observaciones anteriores sugieren que los curtovirus se originaron por un evento de recombinación entre un begomovirus y un ancestro de tipo mastrevirus (Varsani et al. 2009, Padidam et al. 1995). Se ha sugerido que el geminivirus ancestral era como un mastrevirus y surgió en algún momento entre 200-100 millones de años atrás (maa) (Rojas et al. 2005, Ribicky 1994), en la zona norte de lo que hoy es África, y que la familia *Geminiviridae* se diversificó en asociación con las angiospermas (Rojas et al. 2005). Por la posición intermedia de los curtovirus en las filogenias y la especulación de que los cuatro géneros de ésta familia viral surgieron antes o durante la separación de los bloques que formaban Gondwana (130-80 maa), se ha postulado que el género *Curtovirus* se diversificó en el Viejo Mundo. El centro de origen sería la región del Medio Oriente y desde allí se habrían esparcido a varias latitudes en asociación con el cultivo de betabel (*Beta vulgaris*, Chenopodiaceae), llegando a América con los colonizadores (Briddon et al. 1998, Bennet 1971).

La hipótesis de la radiación de los curtovirus en el Viejo Mundo también se apoya en el hecho de que hay algún grado de resistencia a las infecciones por curtovirus en especies silvestres del género *Beta* del Medio Oriente. Por otra parte, el género *Circulifer*, que es el insecto vector, tiene su mayor diversidad en el área del Mediterráneo y el Medio Oriente (Briddon et al. 1998, Bennet 1971). En 1998 se identificó un aislado de curtovirus 97% idéntico a BSCTV en cultivos de remolacha en Irán (Briddon et al. 1998); dicho porcentaje de similitud más que indicar una relación de mucho tiempo entre los virus de ambos continentes, hace pensar en que un evento muy reciente (sólo décadas atrás, por ej. una introducción de un continente al otro) es lo que relaciona a los dos aislados.

Como argumento en contra de una radiación de los curtovirus en el Medio Oriente está la diversidad del género en las Américas: para los otros géneros de la familia *Geminiviridae* se considera que su centro de origen es la región donde se observa la mayor diversidad del grupo (Nawaz-ul-Rehman & Fauquet 2009), y al seguir esta línea de razonamiento los datos a simple vista indican que el género *Curtovirus* se ha diversificado en las Américas.

Trabajos recientes sobre la evolución del género *Begomovirus* pueden dar luz sobre la distribución de los curtovirus y su relación con el resto de la familia viral. Al igual que los curtovirus, los begomovirus se distribuyen en América y en el Viejo Mundo, pero hay una división clara entre los begomovirus del Nuevo y los del Viejo Mundo (Ha et al. 2008); esta división al parecer obedece al hecho de que los dos grandes bloques terrestres que hoy conforman el continente americano han estado aislados de las demás masas continentales desde hace unos 80 maa (Ribicky 1994). Los datos indican que los begomovirus americanos no han sido introducidos a este continente por los movimientos humanos recientes. Se postula que este tipo de geminivirus estaba presente en Gondwana y tras la ruptura del súper-continente pasaron a Sudamérica, para luego, hace unos 10-3 maa alcanzar el bloque Norteamericano, tras la formación del Istmo de Panamá.

La hipótesis anterior implica que los begomovirus americanos tuvieron varios millones de años para evolucionar en aislamiento geográfico. Se conoce al menos un par de evidencias que indican que así ocurrió; la primera es que todos los begomovirus del continente americano carecen de un gen AV2, mientras que la inmensa mayoría de los begomovirus del Viejo Mundo lo poseen (Stanley et al. 2005, Harrison et al. 2002), salvo un par de excepciones que parecen remanentes del linaje que se presume pasó a Sudamérica (Ha et al. 2008, Ha et al. 2006); en segundo lugar, en la parte N-terminal de la proteína CP de los begomovirus americanos hay una secuencia de aminoácidos que constituye una marca molecular que funciona como huella biogeográfica (Ha et al. 2008, Ha et al. 2006) y que sirve, junto con los registros de introducción de material vegetal, para identificar a los begomovirus que han circulado recientemente desde y hacia las Américas.

Hace poco se describieron dos nuevos geminivirus que son especies candidatas del género *Curtovirus*, uno originario de Arizona, *Pepper yellow dwarf virus* (PeYDV) (Lam et al. 2009) y el otro identificado en Irán, *Beet curly top Iran virus* (BCTIV) (Yazdi et al. 2008). Este último se considera nativo del Viejo Mundo pues su vector, *Circulifer haematoceps* no se ha encontrado en el continente americano, difiere de los otros curtovirus conocidos en que la organización de los genes en sentido complementario es similar a la de los mastrevirus, y sólo comparte con los curtovirus la región genómica que incluye los genes en sentido del virión, entre ellos el que codifica a la proteína de la cápside (CP), la cual tiene una similitud entre 71-75% con la de otros curtovirus.

Con las dos especies de curtovirus mencionadas aumenta el número de representantes del grupo, y también la posibilidad de encontrar evidencia molecular que permita discernir la ruta evolutiva que han seguido los miembros de éste género, y/o plantear una explicación alternativa que reconcilie las observaciones enunciadas anteriormente. Es evidente que a mayor número de especies conocidas del género y con un mejor conocimiento de su rango de distribución, más confiables serán las conclusiones a las que un estudio de éste tipo pueda conducir.

En este trabajo se hace un análisis exhaustivo del genoma de las cinco especies de curtovirus reconocidas por el ICTV y de las especies recién reportadas, con el fin de detectar en ellos marcas moleculares biogeográficas que den indicios del origen evolutivo de este linaje viral. El proyecto global del grupo de investigación incluye la identificación de curtovirus en México y así, en colaboración con otros miembros del grupo, se aislaron y caracterizaron dos curtovirus, uno que representa a una nueva especie, *Pepper curly top virus* (PepCTV), contenida en un extracto de DNA de plantas de Chile donado por la Dra. Rebecca Creamer de la Universidad Estatal de Nuevo México-EUA como control positivo para el proyecto, y cuya identidad no se conocía (Creamer et al. 2005), y una cepa de BMCTV detectada en cultivos de Chile en el municipio de Villa de Arista del Estado de San Luis Potosí. Los detalles de la caracterización molecular y biológica de éstos curtovirus no se discuten en esta tesis, pero los

datos contenidos en sus secuencias genómicas sí se incluyen en algunos de los análisis que aquí se muestran.

3.2. Métodos experimentales

Para aumentar la diversidad conocida del género se usaron varios métodos experimentales relacionados con el diagnóstico y caracterización de geminivirus. Dichos métodos se describen en el Anexo 1 de ésta tesis, una versión en español estará disponible pronto (Mauricio-Castillo 2010, en preparación) y además pueden ser discutidos mediante comunicación personal con los investigadores Mauricio-Castillo y/o Arguello-Astorga.

3.3. Análisis de secuencias

En general los análisis de secuencias se hicieron con dos propósitos:

- 1) Para obtener datos de las secuencias obtenidas en la parte de caracterización de curtovirus en México. En este caso el trabajo consistió en la detección de los marcos de lectura contenidos en el genoma viral y la comparación de éstos, de sus productos proteicos y de las regiones no codificantes con los de otros virus del mismo género, mediante las diferentes aplicaciones del programa Lasergene (DNASTAR, Madison, WI).
- 2) Para reconstruir la historia evolutiva del género; las herramientas informáticas concretas y las consideraciones teóricas utilizadas para éste caso se describen más adelante.

3.3.1. Algunos datos sobre PepCTV y BMCTV-Mex

El número de acceso en la base de datos GenBank del NCBI para el virus PepCTV es NC_009518 y el de la variante BMCTV-Mex es EU193175. La figura 3.1 muestra la organización genómica de PepCTV e indica que éste es un curtovirus típico, ya que contiene todos los marcos de lectura de la mayoría

de especies de este género. Dicha figura también muestra como un evento de recombinación entre BSCTV y una variante de BCTV produjo esta especie viral; la secuencia recombinante incluye la región de los genes en sentido complementario.

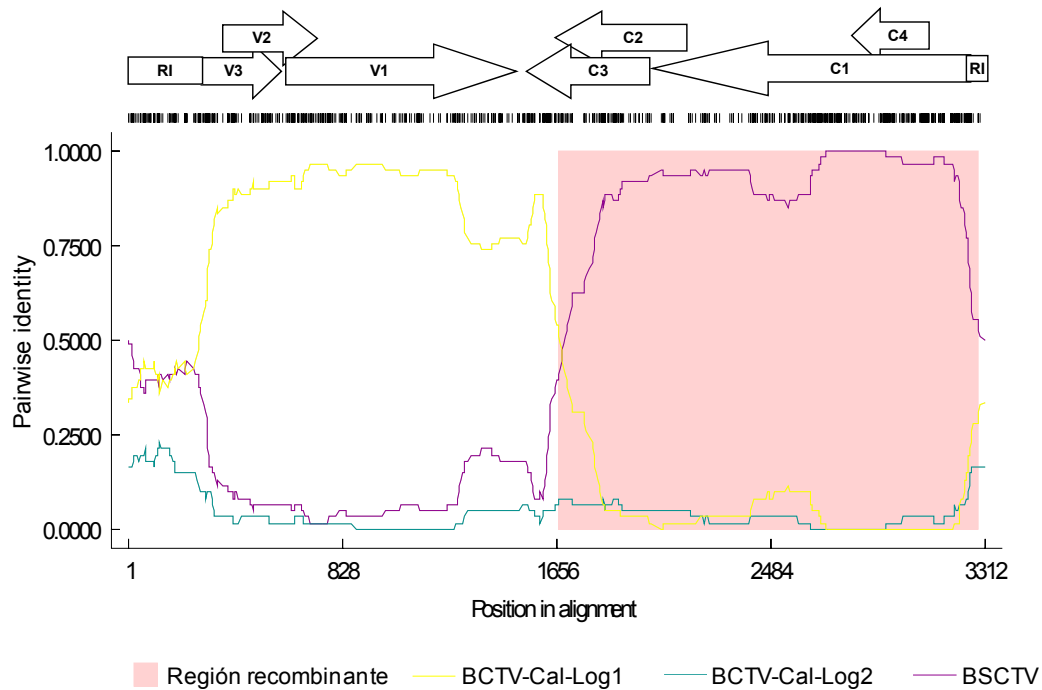


Figura 3.1. Un evento de recombinación dio origen a PepCTV. El análisis de recombinación se hizo con el programa RDP (Martin et al. 2005), tras realizar un alineamiento múltiple en el programa MEGA 4.0; los parámetros para la corrida fueron un valor de $p < 0.001$ y considerando como válidos sólo aquellos eventos detectados por al menos tres detectores de recombinación. Arriba se muestran los marcos de lectura de la nueva especie de curtovirus y debajo la señal de recombinación detectada por el detector Chimaera. Los parámetros para la corrida fueron un valor de $p < 0.001$ y considerando como válidos sólo aquellos eventos detectados por al menos tres detectores de recombinación.

3.3.2. Análisis evolutivo del género *Curtovirus*

Para hacer un replanteamiento de la historia evolutiva de este subgrupo de los geminivirus se obtuvieron las secuencias de genomas completos de todas las especies, cepas y aislados de curtovirus depositadas en la base de datos GenBank, y en primer lugar se igualaron para que todas tuvieran el mismo sitio

de inicio, el cual por consenso es el sitio del nonúmero conservado en el que la proteína Rep introduce el corte para iniciar la replicación por CR del virus. Para organizar los datos también fue necesario obtener la secuencia de las proteínas codificadas en los genomas mediante la herramienta EditSeq del programa Lasergene (DNASTAR, Madison, WI), ya que se encontraron casos en los que la secuencia de esas proteínas en la base de datos no era la correcta e incluía aminoácidos extra en el extremo N-terminal. Además se usaron secuencias genómicas de otras especies de geminivirus, las cuales igualmente se uniformaron para que tuviesen el mismo sitio de inicio.

La estrategia que en general se sigue hoy día (en la era post-genómica) para proponer una hipótesis sobre la historia evolutiva de un linaje consiste en primero detectar los rasgos comunes que hay en los genomas de los miembros del grupo, así como las peculiaridades de cada uno de ellos (etapa de genómica comparativa). Luego hay que hacer una representación jerárquica del grupo (reconstrucción de la filogenia), ya sea basada en los detalles genómicos identificados o en una porción del genoma que pueda representarlos. Finalmente, se usa la información sobre el comportamiento biológico y los factores externos/ambientales involucrados en éste para interpretar el sistema completo. Los tres pasos se siguieron en este trabajo, con las herramientas y los enfoques que se describen en las secciones a continuación.

3.3.2. 1. Identificación de las peculiaridades genómicas

Además de los marcos de los genes que se codifican, hay una serie de características que se pueden buscar en un genoma geminiviral, entre ellos elementos estructurales y reguladores de la transcripción y la replicación, cuya presencia o ausencia, al igual que el orden o la frecuencia con la que aparecen pueden reconocerse como rasgos derivadas de un ancestro, o ser utilizados como marcadores de una jerarquía.

En la figura 3.2 se muestra una forma de comparar la organización genómica de los curtovirus reconocidos por ICTV entre sí, y con respecto a los

otros géneros de la familia *Geminiviridae*. La comparación considera tres aspectos: la presencia de genes en determinada posición (homólogos posicionales), el origen de dichos genes (detección de homología verdadera), y la similitud que hay entre homólogos reales.

Para determinar si el gen tiene homólogos posicionales en los otros géneros se hizo una búsqueda simple de los marcos de lectura en los genomas, marcando su posición con respecto al sitio de corte en la estructura tallo-asa. En los casos donde no se encontró un homólogo posicional se hizo Blastp con los marcos pequeños (menores a 50 aa) para asegurar que no se trataba de que el gen estuviera interrumpido. Todos los genes tuvieron un equivalente posicional en alguno de los otros géneros.

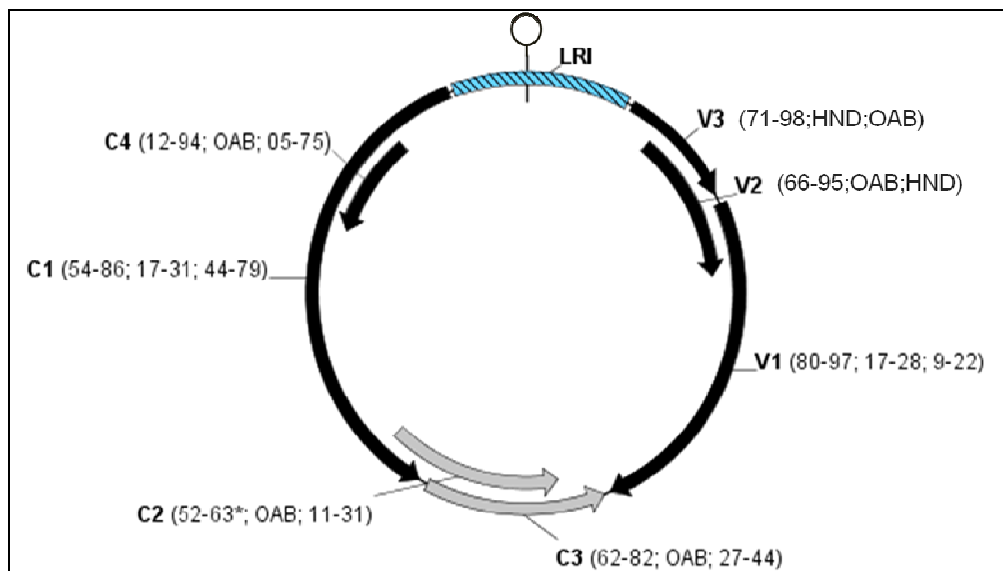


Figura 3.2. Comparación de la organización genómica de los curtovirus reconocidos por el ICTV con la de los otros géneros de la familia *Geminiviridae*.

Se indican los siete marcos de lectura de los curtovirus (C1-C4 y V1-V3). Frente a cada marco, entre paréntesis, están resumidos los datos que indican la homología con otros géneros, separados por punto y coma. Los datos en el paréntesis están organizados de la siguiente manera: el primero indica el intervalo en que oscila el porcentaje de identidad del ORF entre los curtovirus; el segundo indica la similitud del ORF con el de los mastrevirus, para lo cual puede presentarse uno de tres casos, ya sea que el gen si es homólogo y se da un intervalo del porcentaje de identidad, que el

ORF esté ausente (OAB) o que no sea un homólogo real (HND= Homología No Detectada); el tercer dato indica la similitud con los begomovirus y se pueden dar tres casos como con los mastrevirus. LRI= Región Intergénica Larga. *En este caso se excluyó el marco C2 de HrCTV porque de antemano se sabía que no tiene homólogo en los curtovirus, ni en los otros géneros.

Una vez conocido cuales genes de los curtovirus ocupan una posición equivalente a la de genes contenidos en los genomas de virus de otros géneros se determinó si era homología posicional o se trataba de homología verdadera, esto es si los genes tenían el mismo origen ancestral. Para esto las secuencias de proteína se sometieron a Blastp y de los datos de salida se obtuvieron los porcentajes de identidad de cada una de ellas con sus homólogos posicionales, como ya ha sido realizado por otros investigadores (Varsani et al. 2009, Baliqi et al. 2004, Padidam et al. 1995). El producto de los genes V2 y V3 no arrojó datos de similitud con otras proteínas de los geminivirus. Para éstos genes se hizo entonces una corrida para búsqueda de homología remota mediante iteraciones de PSI-Blast (Bhadra et al. 2006), las cuales llevaron a la conclusión de que ambos genes son exclusivos del género *Curtovirus* ya que los valores de identidad que mostraban con otras proteínas de la base de datos no superaba el umbral considerado un efecto del azar, que en este caso fue cuando los valores E eran mayores de 0.001.

En el caso de genes que sí resultaron tener homólogos verdaderos se usó el porcentaje de identidad como indicio de cuanto han divergido. Los datos de Blastp obtenidos durante la búsqueda de los equivalentes posicionales se colectaron en términos de la identidad de la secuencia de aminoácidos, y se resumen en cuanto al porcentaje de identidad máximo y mínimo detectado.

Los datos principales arrojados por la comparación de genomas que se deben tener en cuenta para fases posteriores del análisis evolutivo son que de las dos regiones en las que se puede dividir la parte codificante del genoma de los curtovirus la que contiene los genes en sentido del virión sólo tiene un gen homólogo verdadero al de los otros géneros, que es el que codifica para la

proteína CP, y que todos los genes codificados en el sentido complementario son homólogos a los de los begomovirus.

En la figura 3.3 se muestra cómo la organización de elementos *cis*-reguladores sirve para establecer jerarquías. En ella se observa que la mayoría de los curtovirus (BCTV, BMCTV, BSCTV, SCTV, y también PepCTV, aunque no aparece en la figura) se caracterizan por poseer dos cajas G y elementos de unión a factores de transcripción tipo Dof (Moreno-Risueno et al. 2007) en el lado derecho de la potencial estructura tallo-asa del origen de replicación por CR, y al lado izquierdo tienen los elementos iterados precedidos de una caja TATA.

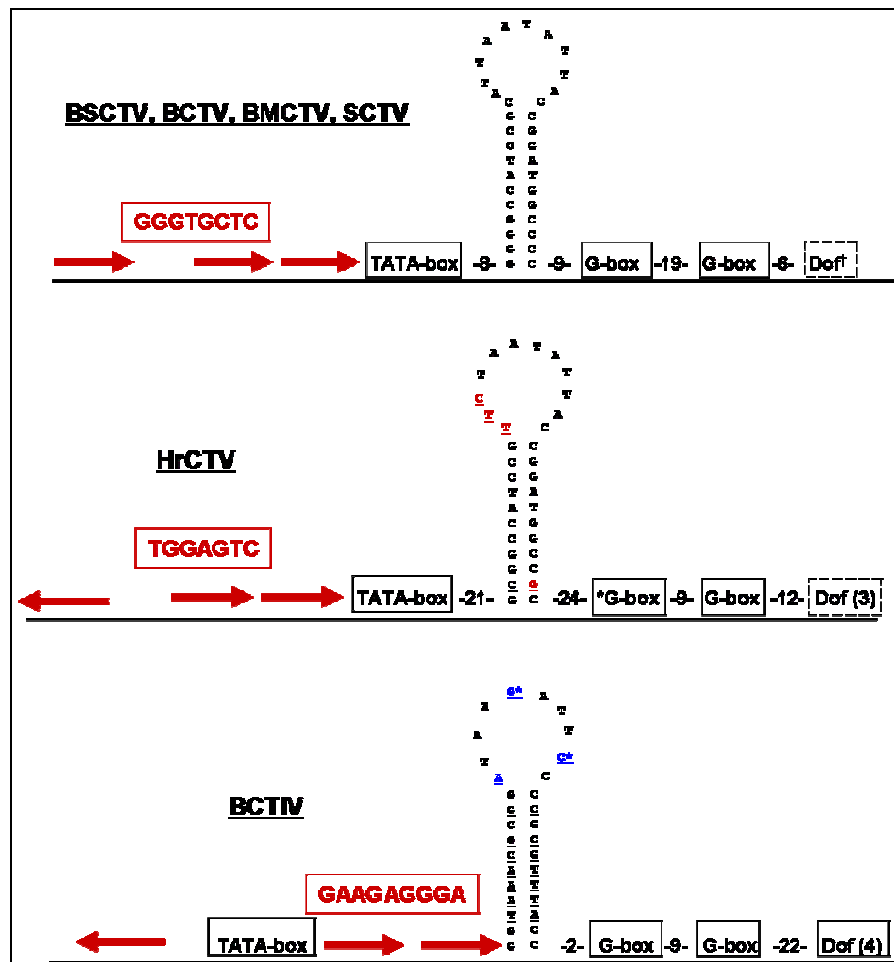


Figura 3.3. Linajes de curtovirus según su organización de elementos reguladores en *cis*. Se muestra la organización de elementos reguladores transcripcionales y replicativos contenidos en la región intergénica de los curtovirus.

Los números entre guiones indican el espaciamiento entre elementos y aquellos entre paréntesis indican el número de elementos Dof presentes, especificando solo la localización del primero de ellos.

Los curtovirus HrCTV y BCTIV tienen los elementos en *cis* organizados de manera distinta y se clasificaron como dos jerarquías diferentes. Las secuencias iteradas y la estructura tallo-asa se identificaron por inspección visual, usando los criterios que se describen en la sección 2.2.1, y los sitios de unión a factores de transcripción se buscaron usando las bases de datos TRANSFAC (BIOBASE Biological Databases) (Matys et al. 2003) y PLACE (Higo et al. 1999).

3.3.2.2. Búsqueda de huellas bio-geográficas

Como se dijo en la introducción, hay evidencia de que los curtovirus americanos tienen un origen recombinante. Con el fin de profundizar en la naturaleza del o los begomovirus involucrado(s) en este evento de recombinación se hizo una búsqueda de huellas filogenéticas en la región de los genes complementarios de los curtovirus. Para esto se hicieron alineamientos de la secuencia de las proteínas homólogas entre begomovirus y curtovirus (C1-C4 y V1), en los que se incluyeron todos los begomovirus de la base de datos de Genbank que tenían reportada la secuencia completa de dicha proteína (alrededor de 150 especies, Fauquet et al. 2008). Los alineamientos se hicieron con varios parámetros, modificando principalmente el tamaño de palabra, y una vez obtenidos se buscó en ellos, de forma visual, los motivos conservados que permitieran colocar a los begomovirus del Viejo y del Nuevo Mundo en grupos separados, para después determinar si la secuencia identificada como característica de uno de estos grupos se presentaba en los curtovirus.

Además de hacer posible el agrupamiento de los begomovirus en los dos clados principales del género, un criterio adicional que se usó para considerar a una secuencia de aminoácidos como marca bio-geográfica fue su tamaño; estos tienen que ser bloques diferenciales (indels y/o sustituciones de aminoácido) de dos o más residuos continuos.

La figura 3.4 muestra un par de huellas en la proteína Rep que relacionan a todos los begomovirus del clado del Nuevo Mundo con las especies de curtovirus aceptadas por el ICTV (de origen americano); se trata de secuencias adyacentes a regiones funcionales de la proteína que son muy conservadas, y probablemente estén sujetas a una fuerte presión selectiva, lo que podría explicar porqué estas “huellas” se siguen manteniendo.

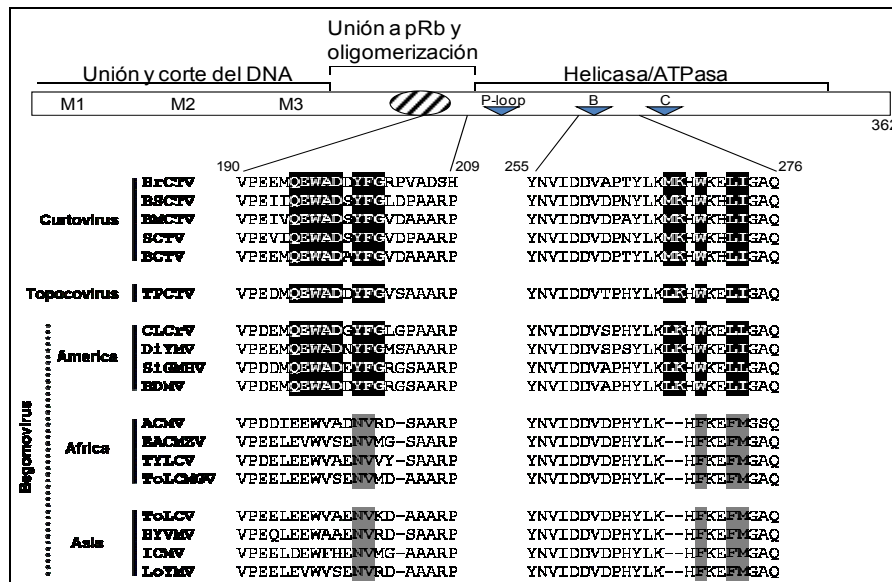


Figura 3.4. Huellas biogeográficas identificadas en la proteína Rep. M1-M3 son los motivos conservados del dominio endonucleasa; P-loop, Walker B y Walker C son los motivos característicos de la helicasa. CLCrV = *Cotton leaf crumple virus* – CLCrV, DiYmV = *Dicliptera yellow mottle virus*, SiGMHV = *Sida golden mosaic Honduras virus*, BDMV = *Bean dwarf mosaic virus*, ACMV = *African cassava mosaic virus*, EACMVZ = *East African cassava mosaic Zanzibar virus*, TYLCV = *Tomato yellow leaf curl virus*, ToLCMGV = *Tomato leaf curl Madagascar virus*, ToLCV = *Tomato leaf curl virus*, HYVMV = *Honeysuckle yellow vein mosaic virus*, ICMV = *Indian cassava mosaic virus*, LoYmV = *Loofa yellow mosaic virus*, ToSCTV = *Tomato severe curly top virus*, TYLCV = *Tomato yellows leaf curl virus*.

Por otra parte, la figura 3.5 muestra el caso de la proteína CP. La figura muestra el alineamiento de una porción de la proteína, en el que se indica la única diferencia en bloque (varios residuos diferenciales continuos) entre los curtovirus identificados en América con respecto a los que se identificaron en

Irán. Esta marca descarta a los curtovirus americanos como descendientes recientes del único curtovirus de indiscutible origen en el Viejo Mundo, BCTIV.

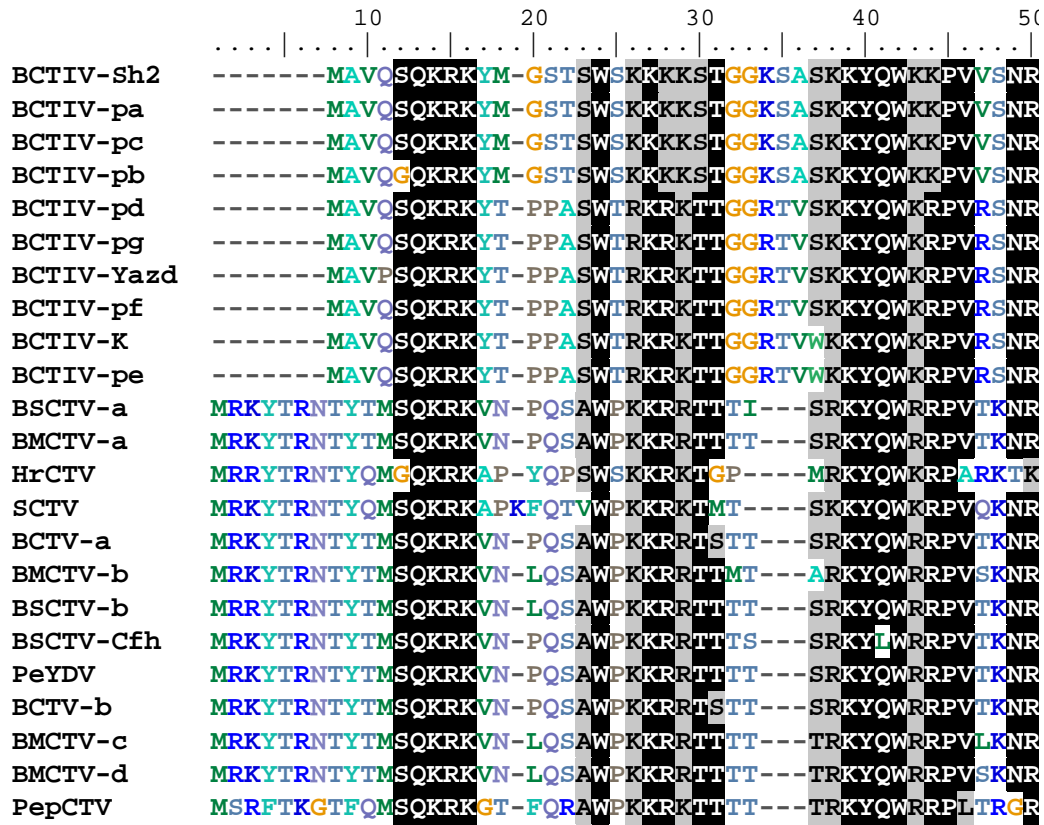


Figura 3.5. Alineamiento de la región N-terminal de todas las proteínas CP de curtovirus disponibles en GenBank (selección no-redundante). Las proteínas BCTIV-pa a BCTIV-pg son secuencias parciales que provienen de aislados en los que no se caracterizó al virus completo.

3.3.2.3. Reconstrucción de una filogenia representativa

Las filogenias conocidas del género *Curtovirus* hasta hace poco carecían de soporte y generaban relaciones “extrañas” cuando se hacían con el genoma completo (Baliji et al. 2004). Ahora que ha aumentado el número de especies en el género, y que se conocen otros geminivirus de tipo “ancestral”, la filogenia del grupo basada en el genoma completo tiene más consistencia, como se puede observar en el árbol de la figura 3.5. Este árbol separa a los curtovirus en tres grupos (los curtovirus típicos, HrCTV y BCTIV), lo cual coincide con los linajes generados por el análisis de las regiones intergénicas (figura 3.3).

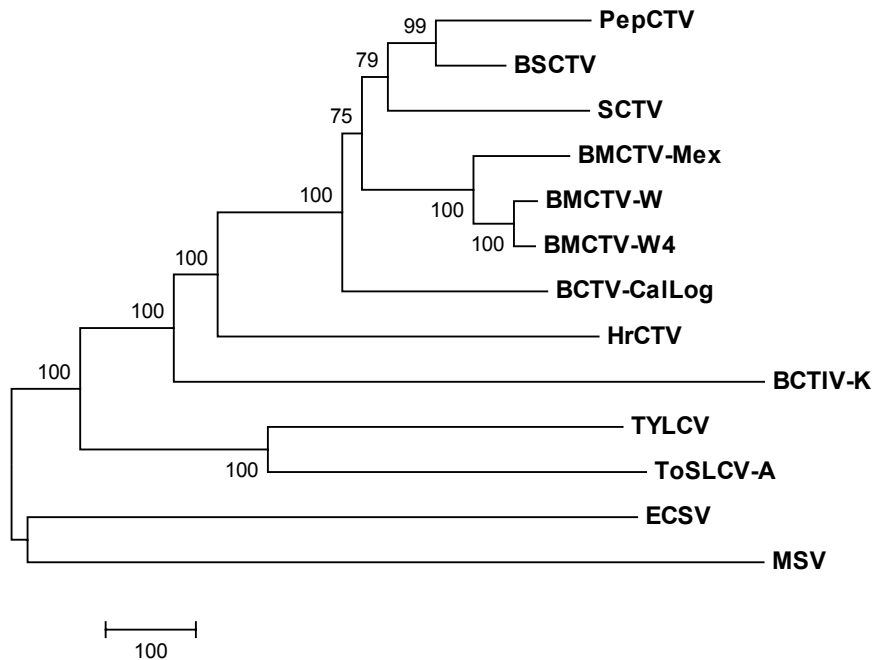


Figura 3.5. Filogenia representativa del camino evolutivo del género curtovirus.

La filogenia se reconstruyó por el método de Neighbor-Joining con la opción Pairwise deletion del programa MEGA (Tamura et al. 2007); el porcentaje de veces (de 100 réplicas) en las que los miembros de cada rama del árbol se agrupan se muestra al inicio de la rama; el árbol es a escala y la longitud de las ramas equivale al número de diferencias nucleotídicas.

3.3.2.4. Conjunción e interpretación de datos

Con la información conocida de la tectónica de los principales bloques continentales que conforman la superficie terrestre, las evidencias moleculares obtenidas mediante el análisis de secuencias, y lo que se conoce respecto a la distribución y evolución de las plantas hospederas y de los insectos vectores, se estableció un escenario plausible sobre la ruta evolutiva que han seguido los curtovirus, la cual se muestra de manera gráfica en la figura 3.7.

Brevemente, se sugiere que los primeros curtovirus ya existían hace unos 100 millones de años, antes de que se formaran los actuales continentes. El curtovirus ancestral posiblemente tenía una organización genómica similar a la del virus BCTIV y estaba en la parte norte de la Pangea (posteriormente

Laurasia). Por cuestiones de la distribución de plantas dicotiledóneas en esa zona, los curtovirus no se dispersaron mucho y más bien permanecieron como relictos en las dos zonas en que se dividió Laurasia: Norteamérica y Europa-Asia.

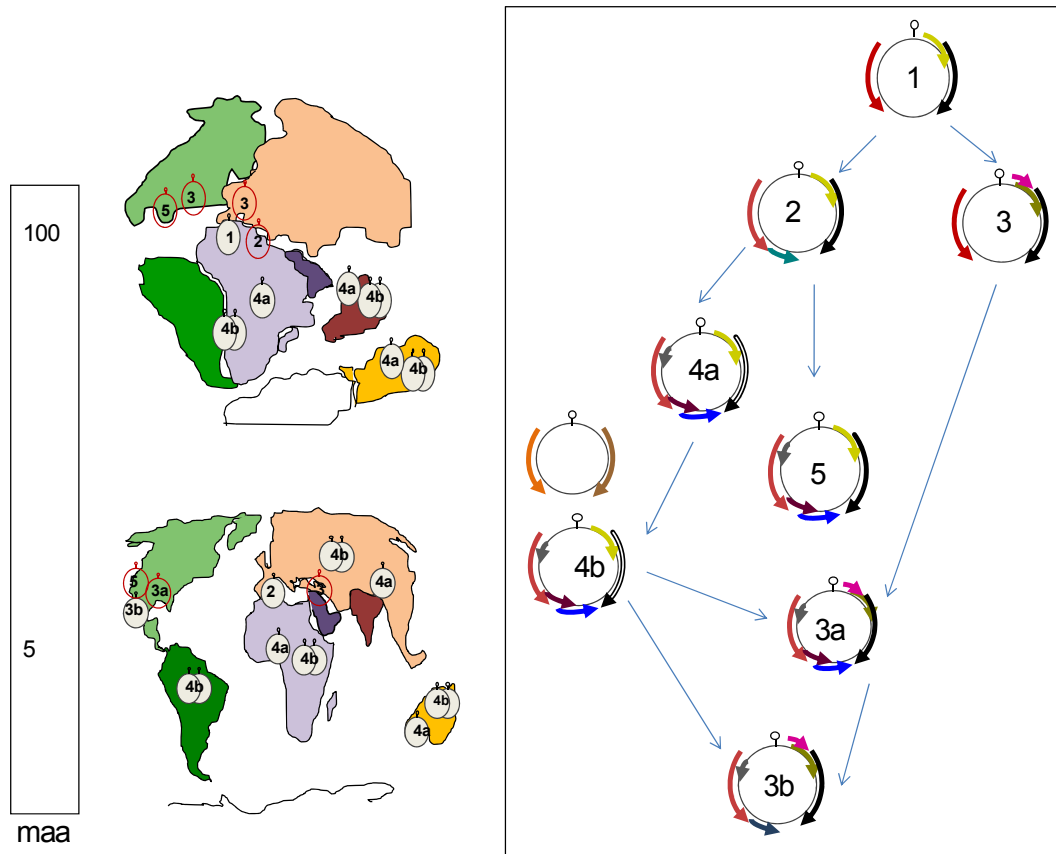


Figura 3.5. Un escenario evolutivo para el género *Curtovirus*. 1) Posible geminivirus ancestral; 2) Un virus con genoma tipo ECSV sería el antecesor de los begomovirus; 3) BCTIV; 3a) Curtovirus americanos típicos; 3b) HrCTV; 4a) Begomovirus del Viejo Mundo; 4b) Begomovirus bipartitas del Nuevo Mundo; 5) TPCTV. Caricaturas en círculos rojos: virus conocidos hoy, pero considerados remanentes de algún linaje por su poca abundancia. maa = millones de años atrás.

Uno de esos relictos fue el curtovirus que dio origen a HrCTV, mediante un evento de recombinación de una porción de la región de los genes en sentido complementario con algún begomovirus americano. Se obtuvieron cuatro evidencias apuntando en esta dirección: 1) Mediante análisis de recombinación se detectó una señal de recombinación en la región de genes en sentido

complementario que no incluye a los genes C2 y C2; 2) dicha señal indica que el fragmento recombinante proviene de un begomovirus americano del subclado del *virus del enrollamiento de la hoja de la calabaza* –SLCV; 3) el virus HrCTV tiene las marcas de los begomovirus del Nuevo Mundo en la secuencia de la proteína Rep, pero 4) su proteína CP es una de las que más ha divergido del grupo de los curtovirus. El hecho de que el programa de recombinación pueda detectar una señal en el genoma de HrCTV indica que este evento no es muy antiguo, y probablemente sucedió en los últimos cinco millones de años, ya que se cree que los begomovirus se introdujeron a la flora neártica a partir de Sudamérica, tras la formación del Istmo de Panamá.

Un evento de recombinación independiente ocurrido entre un curtovirus ancestral con un begomovirus del Nuevo Mundo de un subclado diferente al que produjo a HrCTV daría origen a las demás especies curtovirales de Norteamérica. Estas especies aunque conservan la huella de los begomovirus del Nuevo Mundo en la proteína Rep y comparten con éstos últimos toda la organización genómica del sentido complementario, tienen proteínas REn y Trap que claramente han divergido como linajes distintos. Además no se encontraron huellas biogeográficas en las proteínas REn y Trap compartidas entre los curtovirus y alguno de los dos subgrupos de los begomovirus. En general los datos apuntan a que el curtovirus ancestral que dio origen a los curtovirus típicos ya poseía un gen C2 homólogo al de los begomovirus, y quizá lo adquirió por una recombinación del bloque de genes complementario con un begomovirus hace más de 50 millones de años.

3.4. Resultados

El conjunto de resultados del análisis teórico más los datos experimentales dio origen a un artículo de investigación cuyo manuscrito está en proceso de redacción, y en el que se concluye que un remanente de los curtovirus ancestrales en el Oeste de Laurasia recombinó hace unos pocos millones de años con un begomovirus del Nuevo Mundo, originando a los curtovirus Americanos hoy conocidos.

3.5. Referencias

- Baliji S, Black MC, French R, Stenger D, Sunter G. 2004. Spinach curly top virus: A newly described Curtovirus species from southwest texas with incongruent gene phylogenies. *Phytopathology* 94:772-779.
- Baliji S, Sunter J, Sunter G. 2007. Transcriptional analysis of complementary sense genes in Spinach curly top virus and functional role of C2 in pathogenesis. *MPMI*. 20:194-206.
- Bennett, CW, Tarrisever A. 1958. Curly top disease in Turkey and its relationship to curly top in North America. *J Am Soc Sugar Beet Technol*. 10:189.
- Bennet CW. 1971. The curly top disease of sugarbeet and other plants. *The Am. Phytopathol. Soc. Monogr. No. 7*.
- Bhadra R, Sandhya S, Abhinandan KR, Chakrabarti S, Sowdhamini R, Srinivasan N. 2006. Cascade PSI-BLAST web server: a remote homology search tool for relating protein domains. *Nucleic Acids Res*. 1;34 (Web Server issue):W143-6.
- Briddon RW, Stenger DC, Bedford ID, Stanley J, Izadpanah K, Markham PG. 1998. Comparison of a beet curly top virus isolate originating from the old world with those from the new world. *Europ J Plant Pathol*. 104:77-84.
- Creamer R, Carpenter J, Rascon J. 2003. Incidence of the beet leafhopper, *Circulifer tenellus* (Homoptera: Cicadellidae) in New Mexico chile. *Southwest. Entomol*. 28:177-182.
- Creamer R, Hubble H, Lewis A. 2005. Curtovirus infection of chile plants in New Mexico. *Plant Disease*. 89:480-486.
- Dellaporta S, J Wood, and JB Hicks. 1983. A plant DNA miniprep: version II. *Plant Mol Biol Rept* 1:19-21.
- Duffy S, Holmes EC. 2009. Validation of high rates of nucleotide substitution in geminiviruses: phylogenetic evidence from East African cassava mosaic viruses. *J Gen Virol*. 6:1539-47.
- Fauquet CM, Briddon RW, Brown JK, Moriones E, Stanley J, Zerbini M, Zhou X. 2008. Geminivirus strain demarcation and nomenclature. *Arch Virol*. 153:783-821.

- Fauquet, CM, Stanley, J. 2003. Geminivirus Classification and Nomenclature: progress and problems. *Ann Appl Biol.* 142:165-189.
- Ha C, Coombs S, Revill P, Harding R, Vu M, Dale J. 2006. Corchorus yellow vein virus, a New World geminivirus from the Old World. *J Gen Virol.* 87:997–1003.
- Ha C, Coombs S, Revill P, Harding R, Vu M, Dale J. 2008. Molecular characterization of begomoviruses and DNA satellites from Vietnam: additional evidence that the New World geminiviruses were present in the Old World prior to continental separation. *J Gen Virol.* 89:312-26.
- Harkins GW, Delport W, Duffy S, Wood N, Monjane AL, et al. 2009. Experimental evidence indicating that mastreviruses probably did not co-diverge with their hosts. *Virology.* 6:104.
- Harrison B D, Swanson MM, Fargette D. 2002. Begomovirus coat protein: serology, variation and functions. *Physiol Mol Plant Pathol.* 60:257–271.
- Higo K, Ugawa Y, Iwamoto M, Korenaga T. 1999. Plant cis-acting regulatory DNA elements (PLACE) database. *Nucleic Acids Res.* 27:297-300.
- Hormuzdi SG, Bisaro DM. 1993. Genetic analysis of beet curly top virus: evidence for three virion sense genes involved in movement and regulation of single- and double-stranded DNA levels. *Virology.* 193:900-9.
- Inoue-Nagata AK, Albuquerque LC, Rocha WB, Nagata T. 2004. A simple method for cloning the complete begomovirus genome using the bacteriophage phi29 DNA polymerase. *J Virol Methods.* 116:209-11.
- Klute KA, Nadler SA, Stenger DC. 1996. Horseradish curly top virus is a distinct subgroup II geminivirus species with rep and C4 genes derived from a subgroup III ancestor. *J Gen Virol.* 77:1369-1378.
- Kreuze JF, Perez A, Untiveros M, Quispe D, Fuentes S, Barker I, Simon R. 2009. Complete viral genome sequence and discovery of novel viruses by deep sequencing of small RNAs: a generic method for diagnosis, discovery and sequencing of viruses. *Virology.* 388:1-7.
- Lam N, Creamer R, Rascon J, Belfon R. 2009. Characterization of a new curtovirus, pepper yellow dwarf virus, from chile pepper and distribution in weed hosts in New Mexico. *Arch Virol.* 154:429-36.
- Martin DP, Williamson C, Posada D. 2005. RDP2: Recombination detection and analysis from sequence alignments. *Bioinformatics* 21: 260–262.
- Matys V, Fricke E, Geffers R, Gößling E, Haubrock M, et al. 2003. TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.* 31:374-378.

- Moreno-Risueno MA, Martínez M, Vicente-Carbajosa J, Carbonero P. 2007. The family of DOF transcription factors: from green unicellular algae to vascular plants. *Mol Genet Genomics*. 277:379-90.
- Nawaz-ul-Rehman MS, Fauquet CM. 2009. Evolution of geminiviruses and their satellites. *FEBS Lett*. 583:1825-32.
- Padidam M, Beachy RN, Fauquet CM. 1995. Classification and identification of geminiviruses using sequence comparisons. *J Gen Virol*. 76:249-63.
- Rojas MR, Hagen C, Lucas WJ, Gilbertson RL. 2005. Exploiting the chinks in the plant's armor: Evolution and emergence of Geminiviruses. *Ann Rev Phytopathol*. 43:361-394.
- Rybicki EP. 1994. A phylogenetic and evolutionary justification for three genera of Geminiviridae. *Arch. Virol*. 139: 49-77.
- Stanley J, Bisaro DM, Briddon RW et al. Family Geminiviridae. In: Fauquet CM, Mayo MA, Maniloff J, Desselberger U, Ball LA (eds), *Virus Taxonomy: The eighth report of the international committee on taxonomy of viruses*. Elsevier/Academic Press. London, UK, pp. 301-326.
- Tamura K, Dudley J, Nei M, Kumar S. 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol. Biol. Evol.* 24:1596-1599.
- Varsani A, Shepherd DN, Dent K, Monjane AL, Rybicki EP, Martin DP. 2009. A highly divergent South African geminivirus species illuminates the ancient evolutionary history of this family. *Virol J*. 6:36.
- Velásquez-Valle R, Medina-Aguilar MM, Creamer R. 2008. First report of Beet mild curly top virus infection of chile pepper in North-Central Mexico. *Plant Disease* 92:650.
- Yazdi HR, Heydarnejad J, Massumi H. 2008. Genome characterization and genetic diversity of beet curly top Iran virus: a geminivirus with a novel nonanucleotide. *Virus Genes*. 36:539-45.

4. Estandarización de un sistema experimental para evaluar promotores de begomovirus

4.1. Antecedentes

Los reguladores transcripcionales en *cis* de los geminivirus están mayormente contenidos la región intergénica, la cual dirige la transcripción de genes tanto en sentido complementario como en sentido del virión (Baliji et al. 2007, Hur et al. 2008, Shivaprasad et al. 2005, Velten et al. 2005) y también contiene los elementos en *cis* que controlan la replicación (Eagle & Hanley-Bowdoin 1977). Esta sección del genoma ha sido estudiada por varios investigadores en diferentes contextos, y se sabe que en cuanto a transcripción funciona de manera independiente según la orientación, esto es, contiene promotores divergentes. Así pues, los elementos en *cis* que participan en la regulación del gen de la proteína de la cápside no tienen la misma relevancia en la transcripción del gen Rep (Lacatus & Sunter 2008, Usharani et al. 2006, Frey et al. 2001), aunque no se conoce con detalle la naturaleza y función de todos estos elementos. Se sabe además que los promotores de la región intergénica no son los únicos en el genoma geminiviral, pero los promotores adicionales están poco caracterizados (Shung & Sunter 2009, Tu & Sunter 2007, Shivaprasad et al. 2005).

Trabajos recientes han demostrado la utilidad de los genomas geminivirales en el desarrollo de vectores para uso biotecnológico (Golenberg et al. 2009, Huang et al. 2009, Regnard et al. 2010, Peretz et al. 2007). Los geminivirus tienen ciertas ventajas que los hacen atractivos para la biotecnología, entre ellas que su genoma pequeño facilita las manipulaciones a nivel molecular, que tienen un amplio rango de plantas hospederas, que el método de inoculación puede ser sencillo y que se conoce el funcionamiento y organización del genoma. Por tratarse de entidades virales, se pueden buscar en ellos las características de un promotor ideal para la biotecnología vegetal

(que tenga una actividad fuerte, pueda ser inducido y se pueda controlar en que tejido se expresa), ya que como todos los virus, éstos tienen tropismos de tejidos y sus ciclos infecciosos incluyen etapas en las que hay una producción abundante de las proteínas virales (Nikovics et al. 2001, Dinant et al. 2004, Carter & Saunders 2007, Shimada-Beltran & Rivera-Bustamante 2007).

Un trabajo realizado en el 2003 demostró que el promotor Rep del *virus del enrollamiento de la hoja del algodón* (CLCuMV) tiene una actividad unas cinco veces mayor que la de promotor 35S del *Virus del mosaico del la coliflor*-CaMV (Xie et al. 2003), que es el promotor que se usa como referencia para evaluar la actividad de regiones de regulación transcripcional en sistemas vegetales, y el que con mayor frecuencia se utiliza para dirigir la expresión de transgenes expresados en plantas. En un trabajo más reciente se vio que el promotor de los genes que se transcriben en sentido del virión del *Virus del rizado de las puntas del betabel* –BCTV puede dirigir la expresión del transgén en vectores diseñados para silenciamiento (Golenberg et al. 2009). En ambos casos la fuerza del promotor se debe a elementos contenidos en la región intergénica y tales elementos podrían utilizarse en el diseño de promotores sintéticos, que es la tendencia actual (Cazzonelli & Velten 2008).

En cuanto a los otros linajes de virus de plantas con genomas ssDNA, vale la pena mencionar que en los nanovirus también se han visto promotores con actividad mayor que la del 35S, y de hecho algunos se han incorporado como elementos reguladores de la expresión de proteínas recombinantes (Shirasawa-Seo et al. 2005, Dinant et al. 2004, Dugdale et al. 1998), pero en cambio en los beta-satélites caracterizados hasta ahora no se han reportado promotores fuertes (Eini et al. 2009, Guan & Zhou 2006).

En nuestro laboratorio existen varias líneas de investigación que requieren de la evaluación experimental de elementos reguladores en *cis*. Entre éstas líneas se cuentan los proyectos enfocados a entender la función del gen C2 de los begomovirus, que codifica a la proteína TrAP, cuya función básica es la activación de los genes tardíos (V1 y BV1) (Yang et al. 2007), al parecer a por un efecto potenciador mediado por un elemento en *cis* llamado CLE (Cazzonelli

et al. 2005); C2 es además uno de los genes de los que la evidencia sugiere que están controlados por una región promotora distinta al promotor de Rep (Shung & Sunter 2009).

Otra línea de investigación busca analizar el comportamiento de distintos arreglos modulares conservados identificados en la región intergénica, los cuales se componen de conjuntos de sitios de unión a factores de transcripción distribuidos en un orden característico para los tres principales linajes begomovirales (clados Nuevo Mundo, Viejo Mundo, y del linaje del *Squash leaf curl virus*), y por lo tanto deben determinar propiedades biológicas de los linajes, como el rango de hospederos, el tejido donde se expresan, o su perfil de expresión de proteínas.

Una tercera idea que se ha planteado tiene que ver con el hecho de que el origen de replicación y el promotor de la proteína Rep están sobrelapados y la unión de proteínas Rep a los iterones es responsable de la auto-regulación de la transcripción de este gen (Eagle & Hanley-Bowdoin 1997). Una cuestión intrigante de éste proceso es que dado que la transcripción y la replicación no ocurren al mismo tiempo, no está esclarecido el papel de la unión de la proteína Rep a los elementos iterados durante la transcripción. Es claro que la auto-regulación se da por la presencia de proteínas Rep al alrededor del inicio de transcripción pero hay dos explicaciones alternativas al fenómeno: el nivel de transcripción podría depender de la afinidad de la interacción Rep-Iterón ó el sistema podría regularse por la aglomeración de proteínas Rep en ésta zona, independiente de la afinidad con que la proteína Rep se une a las secuencias repetidas.

Estas líneas de trabajo requieren experimentos en sistemas de expresión transitoria, en los cuales el elemento en *cis* que se va a examinar se analiza en la forma de un promotor híbrido o sintético fusionado a un gen reportero. La construcción se introduce en un sistema celular para que se exprese durante un período de tiempo corto (entre 12 y 72 horas), y luego se procede a coleccionar el material y a cuantificar el nivel de expresión del reportero por el método más

conveniente, que puede ser por actividad enzimática, cuantificación directa de la proteína expresada o cuantificación del transcrito producido.

Hay una cuarta línea de trabajo en el laboratorio, la cual estudia el efecto de mutaciones en el genoma geminiviral sobre la capacidad replicativa. Ésta se ha estado trabajando en un sistema de planta completa, mediante el seguimiento de la sintomatología causada por la infección de los virus mutados, pero éste tiene la desventaja de que los experimentos toman varios días o semanas y por tanto se requiere de un sistema que reporte resultados en menor tiempo, el cual es el ensayo de replicación en células o protoplastos derivados de éstas.

Así pues, el deseo de tener un sistema experimental para analizar la actividad de promotores begomovirales obedece a dos motivos principales: el interés en encontrar elementos reguladores con potencial de ser incorporados en otros sistemas experimentales y biotecnológicos, y la necesidad de aumentar la capacidad de maniobra a la hora de hacer experimentos para ampliar el conocimiento que se tiene de la regulación de la expresión génica y la replicación en el género *Begomovirus*.

4.2. Metodología

Se escogió el gen *uidA*, generalmente conocido como *gus*, que codifica para la enzima β -glucuronidasa (hidroliza enlaces glucosídicos de los glucurónidos), ya que éste es ampliamente usado en los experimentos de actividad transcripcional en sistemas vegetales por tener las siguientes ventajas: 1) Las plantas tienen una actividad de β -glucuronidasa basal baja o nula, por lo que prácticamente no existe un “fondo” que altere las mediciones; 2) su actividad puede ser detectada de forma cualitativa mediante tinción histoquímica o puede ser medida cuantitativamente por un ensayo enzimático de fácil aplicación y costo moderado; 3) la actividad es muy específica y estable, y no se ve opacada por efectos lumínicos como ocurre en el caso del gen reportero de la luciferasa; y 4) en el Instituto se contaba con los equipos requeridos para hacer

las mediciones cuantitativas, evitando así la compra de filtros o aparatos adicionales.

Por otro lado, se optó por un sistema de protoplastos ya que en éste sistema se han realizado la mayoría de los experimentos reportados por otros grupos que son antecedentes para las líneas de investigación del laboratorio.

4.2.1. Fuente constante de material para protoplastos

Para mantener una fuente continua de células útiles para hacer los protoplastos se inició el cultivo de la línea de células de tabaco NT1 (*Nicotiana tabacum-1*), las cuales provienen del mesófilo de las hojas, pero han sido mantenidas en cultivo por diferentes grupos desde hace decenas de años, lo que hace que tengan algunas particularidades producto del cultivo *in vitro*. La línea fue donada por el Dr. Rafael Rivera-Bustamante del Departamento de Biotecnología Vegetal del CINVESTAV-Irapuato. La línea se mantiene el cultivo líquido en medio Murashige-Skoog, como se indica en el Anexo 1, y puede ser usada para generar callos en cantidades abundantes, pero las células han perdido la capacidad de regenerar tejidos y plantas completas (Russell et al, 1992).

4.2.2. Estandarización del proceso digestivo

Con el fin de establecer las condiciones adecuadas para tener un balance entre la calidad y cantidad de protoplastos obtenidos, y la cantidad de enzima utilizada, se hicieron varios ensayos en los que se hizo un seguimiento microscópico de las células sometidas a diferentes concentraciones de la solución enzimática a través del tiempo. Las enzimas digestivas usadas son celulasa de *Trichoderma viride* y pectoliasa de *Aspergillus japonicum* (ambas inicialmente de Sigma-Aldrich Co., y luego de KARLAN Research Productos Co., Cottonwood, Arizona, USA); el modo de preparación de la solución enzimática y el procedimiento que se sigue para digerir las paredes celulares se describen en el Anexo 1, al igual que la forma en que se mantienen en los protoplastos en cultivo.

4.2.3. Construcciones realizadas

Se hicieron tres tipos de construcciones moleculares básicas: los controles positivos, que llevan al gen *uidA* bajo el promotor 35S del virus del mosaico de la coliflor (CaCMV), los controles negativos, que llevan al gen *uidA* sin promotor, y construcciones que llevan al gen *uidA* bajo el control del promotor *Rep* de begomovirus y que sirven para tener los niveles de la expresión de este promotor como punto de referencia a la hora de analizar la actividad de otras regiones del genoma geminiviral.

La fuente del gen *uidA* y del promotor 35S fue el vector binario pBI121, del cual se escindió un fragmento con los dos elementos mencionados más el terminador del gen nopalina-sintetasa (y en otro caso sólo el casete GUS-terminador) y se transfirió a la región de clonación múltiple de los plásmidos pK19 y pBlueScript SK II, como se detalla a continuación.

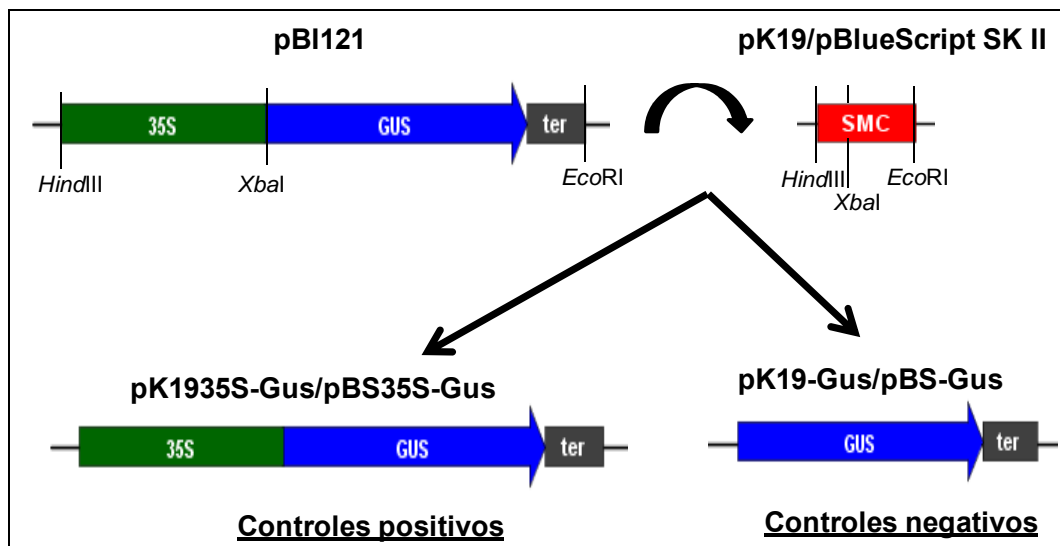


Figura 4.1. Construcciones control. Arriba están las fuentes de los fragmentos y abajo los productos de la ligación de dichos fragmentos. Estos productos son los controles positivos, llamados pK1935S-Gus y pBS35S-Gus, y los controles negativos, llamados pK19-Gus y pBS-Gus. Los nuevos vectores están nombrados indicando primero al plásmido en que se introdujeron los fragmentos, seguido de una indicación del promotor y por último se indica el gen reportero. ter = terminador nopalina-sintetasa, SMC = sitio de clonación múltiple.

La razón por la que se hicieron controles en ambos plásmidos se debe a que se quería tener versiones de las construcciones con dos genes de selección por si se realizaba alguna cotransfección y también porque ya existían algunas construcciones del promotor Rep fusionado a GUS subclonadas en pBlueScript realizadas por Astrid García Moreno-Rubli durante su tesis de maestría, pero ella no contaba con todos los controles correspondientes, además de que el vector pK19 ofrecía un sitio múltiple de clonación más versátil y su factor de selección, el gen de la enzima kanamicina fosfotransferasa, es más estable que el producto del gen de resistencia a ampicilina que portan los vectores pBS.

El plásmido pK19-Gus se usó para generar las construcciones que llevan el promotor del gen de la proteína Rep, y a partir de éstos se realizaron construcciones adicionales con el objetivo de usarlas en experimentos que permitan analizar el papel de los iterones en las propiedades transcripcionales de éste promotor, las cuales se describen en la sección de resultados.

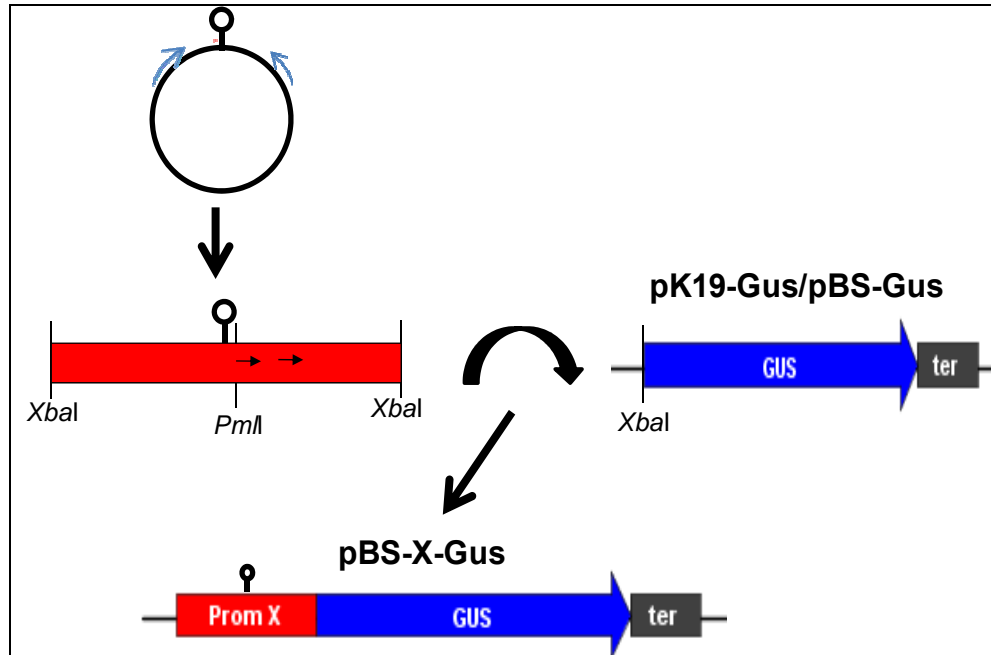


Figura 4.2. Construcciones con el promotor Rep fusionado al reportero GUS (series pBS-X-GUS y pK19-X-GUS, siete vectores). Las flechas delgadas y azules indican el sitio donde se pegan los oligonucleótidos consenso para amplificar regiones intergénicas de begomovirus.

Los promotores Rep se obtuvieron mediante PCR, a partir de clonas de begomovirus conocidos, ó de fragmentos de DNA begomoviral obtenidos de extractos vegetales de los que los genomas virales completos no pudieron ser caracterizados. La regiones intergénicas se amplificaron con la pareja de oligonucleótidos degenerados Rep-Mot-Gus/CP-Mot-Gus (GAGTCTAGATGGATANGTDAGGAAATARTTYTTRGC/GCGTCTAGATCGCC ANGGRGCRTCACGCTTAGGCATT), excepto en el caso de PepGMV, donde se usó el un iniciador directo específico (Rep-Mot-Gus-TPV, GTGGATATGTTAAGAAAATGTTCTTACATTG). Los oligos contienen sitios *Xba*I en los extremos, para facilitar la clonación, y se usaron en mezclas de reacción de 50 µl con la siguiente composición: 50-200 ng de DNA, 75 mmoles de MgCl₂, 23 pmol de cada iniciador, 1.5 µl de DNA polimerasa *Pfu* (Promega, Madison, WI, USA) y 232 µmol de cada dNTP, todo esto en una solución de Tris-HCl pH 8.0 50 mM y NaCl 50 mM. El programa de amplificación consistió en un ciclo inicial de desnaturalización a 94°C por dos minutos, seguido de 35 ciclos de amplificación con 30s desnaturalización a 94°C, 30s de alineamiento a 56°C y 30s de elongación a 72°C, y por último un ciclo de extensión final a 72°C durante 5 minutos.

4.2.4. Estandarización del sistema de transformación

Para este proceso se necesitó una construcción adicional, que consistió en poner el gen de la proteína verde fluorescente (GFP) bajo el control del promotor 35S en el vector pBS (elaborada y amablemente donada por el BQ. Josefát Gregorio Jorge). La construcción se introdujo a los protoplastos mediante varios protocolos de electroporación en el equipo Bio-Rad GenePulser Xcell, después de lo cual se incubaron a 25°C en medio de cultivo para protoplastos por 48 horas, y finalmente se determinó la eficiencia de transformación por conteo de micro-colonias verdes en una gota de 20 ul del cultivo de protoplastos, observada bajo el filtro de luz azul de un estereoscopio LEICA MZ12.5.

4.2.5. Ensayo de actividad β -glucuronidasa

Como sustrato para determinar de manera cuantitativa la actividad del gen GUS se utilizó el 4-metilumbelil- β -D-glucurónido (MUG) (Sigma-Aldrich), ya que de su hidrólisis por la β -glucuronidasa se produce metil-umbeliferona (MU) y ácido glucurónico; el primer producto es un compuesto que fluoresce con una ganancia de fluorescencia que va en función de la concentración, y con un rango de excitación entre 355 y 372 nm, y de emisión entre 440 y 480 nm. Para realizar el ensayo se cosechan los protoplastos, se extrae la proteína total y se cuantifica la misma mediante el método de Bradford.

Una vez cuantificado el contenido de proteína total se lleva a cabo la reacción enzimática. Para esto se mezcla el extracto de proteína total con el buffer de reacción que contiene al sustrato, éste se incuba a 37°C, y se mide la actividad enzimática mediante la cuantificación de la emisión fluorescencia a diferentes tiempos. Los datos se analizan con respecto a una curva estándar de metil-umbeliferona de sodio (NaMU) (Sigma-Aldrich), y los resultados se expresan como la cantidad de MU generada/concentración de proteína/unidad de tiempo, todo como se indica en el Anexo 1.

Los protocolos más conocidos para medir la actividad de esta enzima consisten en la medición de la actividad en un fluorómetro de celdas de vidrio o cuarzo, sin embargo, en el laboratorio se contaba con un fluorómetro de lectura de microplacas (GENios TECAN) (Tecan Group Ltd, Männedorf, Switzerland), capaz de proporcionar las longitudes de onda de emisión y excitación necesarias para la lectura de NaMU, y fue así como se hizo un esfuerzo por adaptar el ensayo a las condiciones de medición en éste aparato, además de que se montó la lectura clásica en el fluorómetro Hoefer Dyna Quant 2000 (Amersham Biosciences), como se describe en el Anexo 1.

4.3. Resultados

4.3.1. Construcciones generadas para el estudio del promotor Rep

4.3.2. Se diseñaron, construyeron y verificaron vectores de expresión con varias versiones del promotor *Rep* de siete begomovirus americanos con diferente especificidad de origen de replicación, fusionados al gen reportero GUS; la organización y secuencia de iterones de éstos virus se indica en la figura 4.3, junto con una representación de la primera versión de éstos promotores, que corresponde a las construcciones que se llamaron pBS-X-Gus y/o pK19-X-Gus donde X se refiere a cada uno de los virus fuente del promotor, y cuyo proceso de construcción se ilustra en la figura 4.2.

La verificación de las construcciones consistió en un chequeo de la orientación en la que se introdujo la región intergénica mediante digestión y/o secuenciación, ya que los iniciadores “-Mot-Gus”, antes mencionados, están diseñados sobre los extremos N-terminal de la región codificante de los genes *Rep* y *CP*, y dependiendo de la orientación se producen fusiones traduccionales *Rep*::Gus ó *CP*::Gus; esto significa que también se generaron fusiones *CP*-X-GUS, las cuales se conservaron.

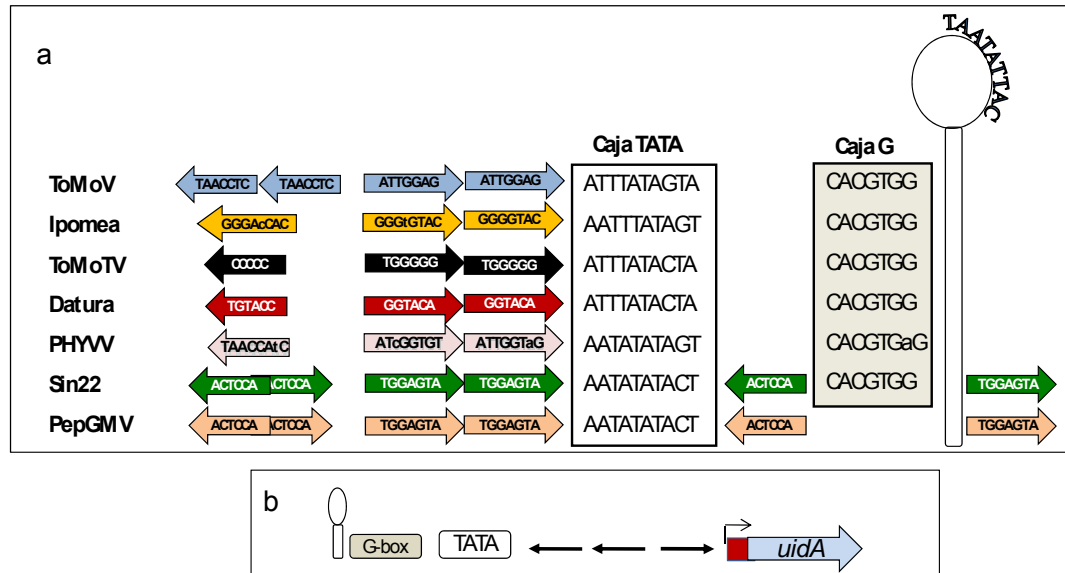


Figura 4.3. Construcciones que permitirían establecer el papel de los iterones en la auto-regulación del promotor *Rep*. a) Secuencia y organización de los iterones de los virus cuyo promotor se fusionó en fase al marco de lectura del gen reportero GUS; b) esquema que ilustra la orientación en la que quedan las construcciones de interés; ToMoV= *Tomato mottle virus*, ToMoTV= *Tomato taino mosaic virus*, PHYVV= *Pepper Huasteco yellow vein virus*; Ipomea, Sin22 (Sinaloa-22) y Datura son regiones

intergénicas obtenidas de muestras de plantas infectadas, de las cuales no se tiene una clona completa del virus.

A partir de las construcciones pBS-X-GUS se hicieron versiones cortas con las que se pretende evaluar solamente el efecto de los iterones, sin presencia de otros elementos en *cis*, esto es, eliminando la estructura tallo-asa, ya que se sabe que las proteínas Rep se acumulan a su alrededor (Singh et al. 2008), y quitando también la caja G, que es el regulador positivo fuerte más común en el promotor *Rep* de los begomovirus del Nuevo Mundo (Eagle & Hanley-Bowdoin 1997, Xie et al. 2003). Estas construcciones se denominaron sp-X-Gus, y previendo que la región del promotor Rep que se conserva puede tener una actividad muy baja debido a la falta de la caja G, se realizó una serie de construcciones que tienen añadido el enhancer del promotor 35S en su extremo 5', las cuales se nombraron e35S-X-Gus.

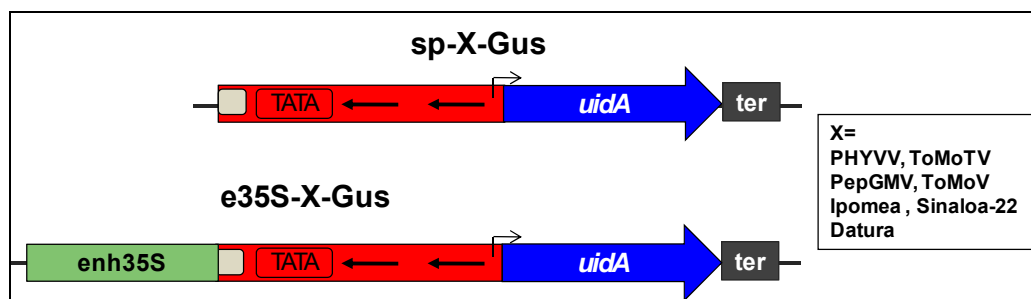


Figura 4.4. Es quema de las series de construcciones que llevan el promotor *Rep* trunco. Para las abreviaturas ver el texto previo. Las flechas curvas indican la fusión traduccional Rep::Gus.

4.3.3. Condiciones óptimas de digestión y de electroporación

Los detalles de las condiciones finales de éstos dos procesos se encuentran en el Anexo 1, pero las figuras a continuación sirven para tener una idea visual de cómo se llegó a éstas y como se deben ver los protoplastos correctamente preparados. En la figura 4.5 se observa como las células sometidas a la solución digestiva van perdiendo su pared sin hincharse y reventar o perder volumen, gracias a que las condiciones osmóticas de cada una de las soluciones por las que pasan las células se mantienen uniformes e isosmóticas respecto a la célula vegetal desnuda.

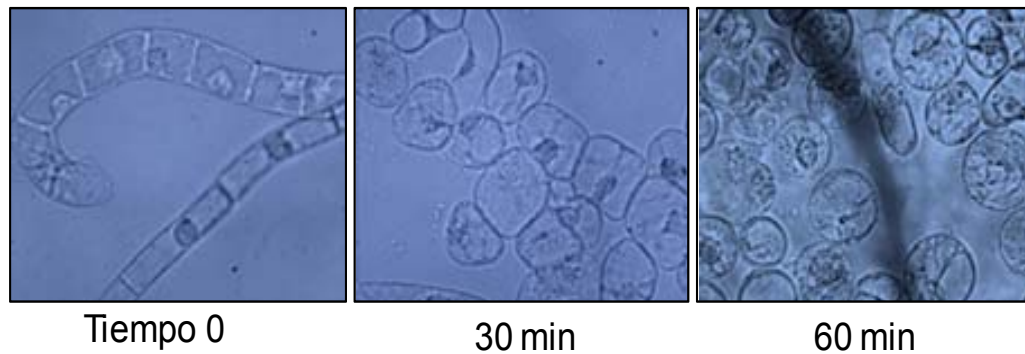


Figura 4.5. Un ciclo de seguimiento del proceso digestivo. Las células NT1 tienen diversas formas pero predominan las células alargadas organizadas a modo de filamento, y se van haciendo esféricas a medida que el proceso digestivo avanza. El momento ideal para detener el proceso digestivo es cuando el 95% de las células ya tienen forma redondeada. (Células vistas a un aumento de 40X).

La intención de coleccionar las células antes de que el 100% de ellas estén digeridas tiene que ver con su susceptibilidad a las sales de la solución de electroporación. Lo que se busca es que no estén completamente desprovistas de pared para que puedan resistir el choque; para conocer esto también se les hizo seguimiento a los protoplastos después de cada uno de los experimentos de electroporación, ya que como parte de la estandarización de este paso, había que conocer qué cantidad de células sobrevivían a las condiciones de electroporación.

La figura 4.6 es para indicar el conteo de puntos verdes fluorescentes, que sirvió para establecer el mejor protocolo de electroporación. En dicha figura se puede notar que el conteo de puntos fluorescentes en el estereoscopio no es el método ideal para analizar la expresión del gen GFP ya que no alcanza la resolución adecuada. Se usó el estereoscopio porque era la fuente de luz azul más cercana disponible y porque para los fines de este trabajo éste es un paso alternativo.

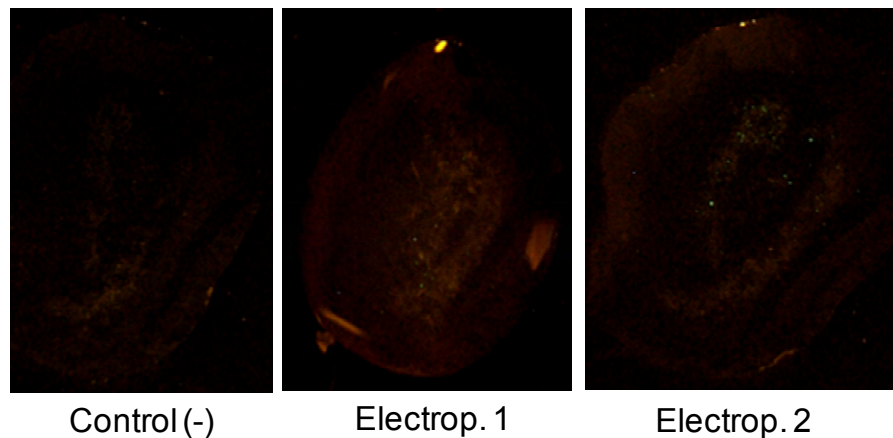


Figura 4.6. Ejemplo de la detección de GFP en gotas de protoplastos. Se observan puntos verdes que son micro-colonias de protoplastos expresando GFP, transformados mediante dos protocolos de electroporación: 1) 130 v y 1000 μ F; 2) 250 v, 500 μ F. El control negativo consiste en un cultivo de protoplastos electroporados con la construcción pBS-Gus.

4.3.4. Lectura de la actividad β -glucuronidasa en el fluorómetro GENios TECAN.

La adaptación se considera un resultado importante dentro de los pasos de estandarización porque el ensayo no se había modificado desde hace más de una veintena de años (Gartland et al. 2000, Jefferson 1987), excepto por una adecuación para hacer la lectura en aparatos de PCR tiempo real publicada en el 2006 (Crow et al. 2006), y que fue la inspiración para hacer la modificación que aquí se describe, que es intermedia entre ambos.

Los pasos del ensayo se indican en el Anexo1, y la figura 4.7 muestra las características de la curva estándar del fluorocromo que se usa como estándar. El parámetro identificado como el más importante a la hora de hacer esta lectura fue la ganancia de fluorescencia, la cual establece en cuanto debe aumentar la cantidad de luz detectada por cada unidad del fluorocromo. La relación concentración-emisión depende tanto del compuesto como del tipo de filtro utilizado para detectar la emisión. En el caso del fluorómetro TECAN solo se pueden hacer mediciones con excitación a 360 nm y emisión a 465 nm y el intervalo de valores que puede tomar la ganancia de fluorescencia va de 1 a 250.

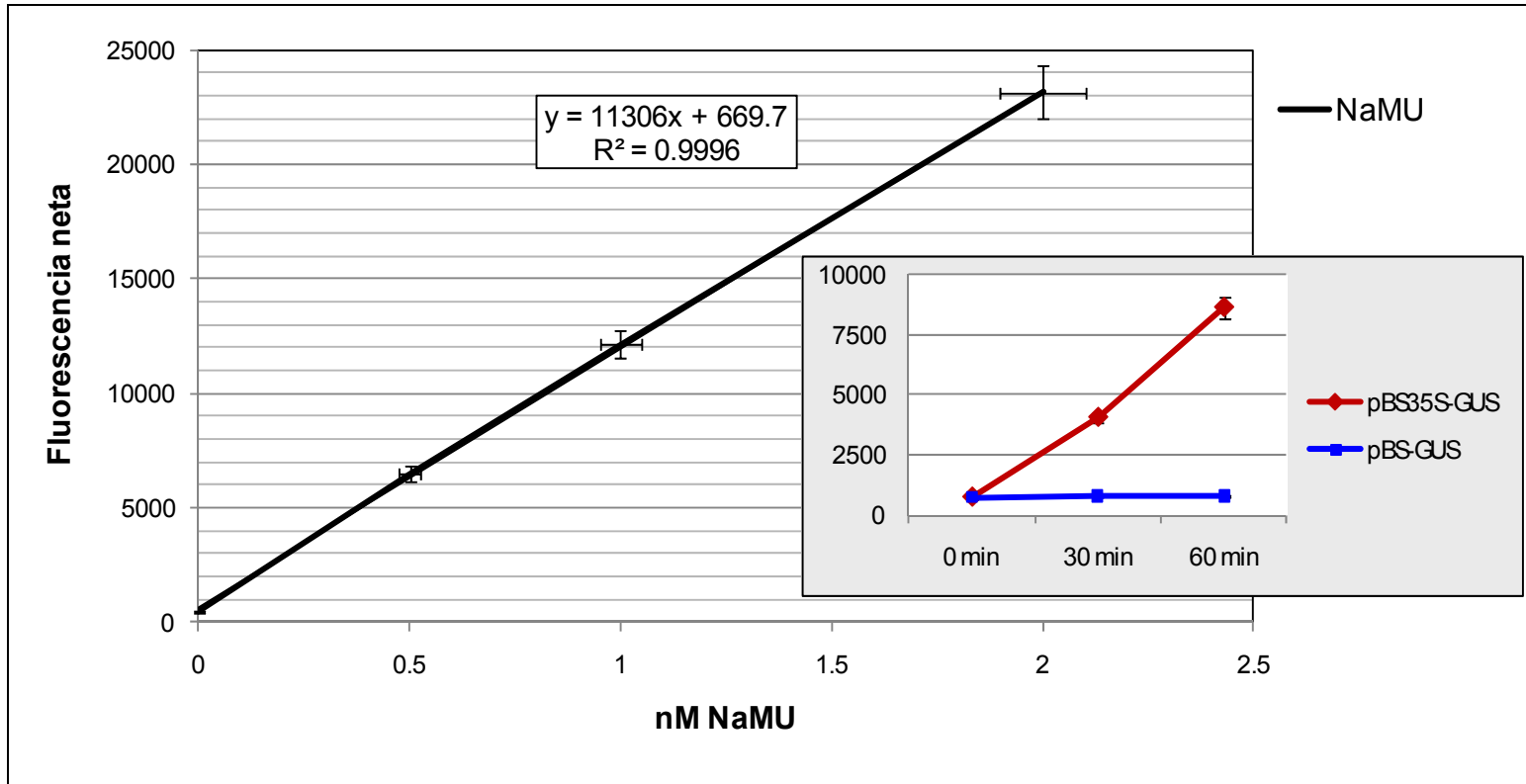


Figura 4.7. Curva estándar de metil-umbeliferona de sodio (NaMU) leída en microplacas en el fluorómetro GENios TECAN. El recuadro pequeño indica los datos de la ecuación de la curva. En el recuadro grande (de fondo gris) se muestra la expresión de GUS a través del tiempo en protoplastos electroporados con dos de las construcciones control, con el eje Y en las mismas unidades que la curva de NaMU.

Tras la estandarización se estableció que 60 es el valor de ganancia de fluorescencia que permite hacer curvas reproducibles en el fluorómetro GENios TECAN. Un recuadro inserto en la figura 4.7 sirve para mostrar que el avance de la reacción enzimática a través del tiempo sí se refleja como un aumento en la cantidad de fluorescencia, que con la curva puede traducirse en nanomoles de metil-umbeliferona liberadas.

4.3.5. Actividad de los promotores

En la figura 4.7 se resumen los datos preliminares que se obtuvieron y que permiten concluir que el sistema experimental se montó adecuadamente. Dicha figura muestra la actividad β -glucuronidasa promedio, y se puede ver que la actividad basal de ésta enzima en las células NT1 es baja y la actividad promedio del promotor 35SCaCMV es de 25.99 nanomoles de metil-umbeliferona/ μ g proteína/hora. La actividad del promotor aislado de un virus de la planta *Datura stramonium* tiene una actividad promedio levemente mayor a la del promotor 35S, mientras que el fragmento Sinaloa-22 (una región intergénica que tiene similitud con los virus del clado del *Squash leaf curl virus*) tiene una actividad tres veces mayor que la del control positivo.

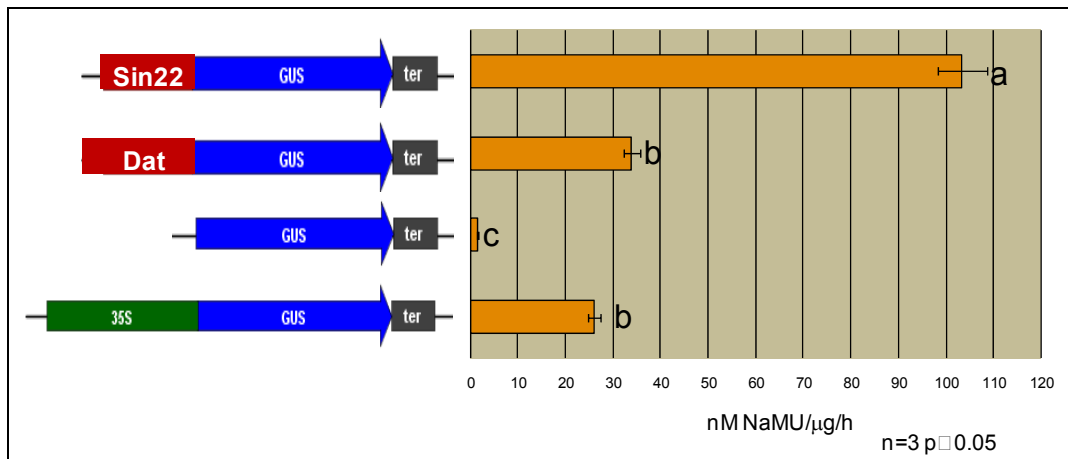


Figura 4.7. Actividad β -glucuronidasa en las construcciones control. Se muestra el promedio de nanomoles de NaMU producidas en tres experimentos independientes con dos repeticiones cada uno y se consideran como diferentes los promedios considerados distintos por una prueba t de student.

Por otro lado, también se realizaron ensayos preliminares con algunas de las construcciones que llevan el promotor *Rep* trunco, y los resultados se muestran en la figura 4.8. Allí se puede ver que la actividad de los promotores aislados de las plantas *Ipomea* y *Datura*, y del begomovirus PHYVV no difiere significativamente de la del promotor 35S. Además se observa que la actividad del promotor *Rep* mínimo de cada uno de estos virus disminuye a más de la mitad; tampoco hay diferencias entre la actividad de éstos promotores mínimos.

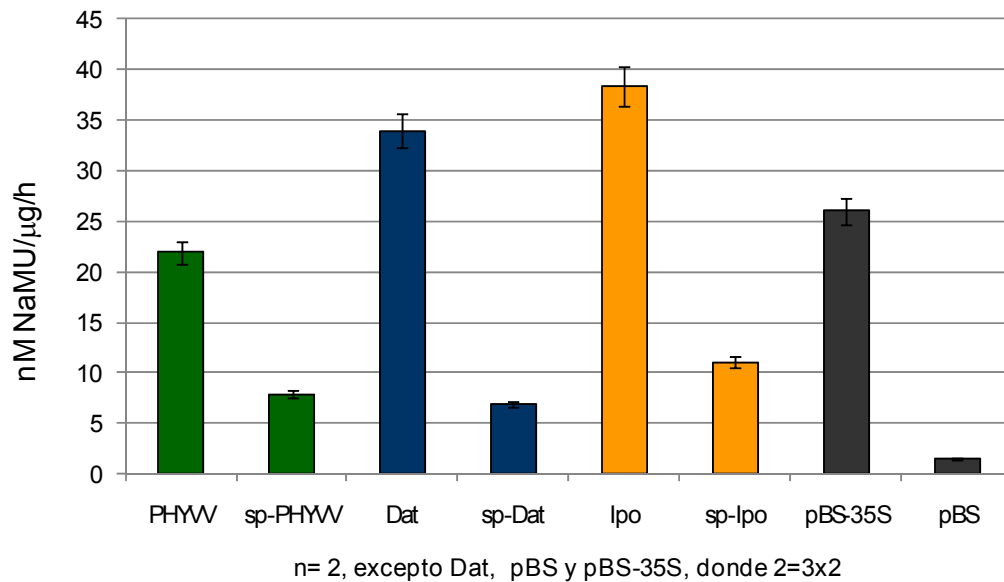


Figura 4.8. Actividad de las versiones cortas del promotor *Rep* en comparación con las versiones más largas. Los datos son preliminares ya que provienen de un solo experimento con dos repeticiones. Los nombres de las construcciones están abreviados con respecto a la nomenclatura asignada en el texto, de manera que las barras a la derecha son de las versiones del promotor largas (pBS-X-Gus/pK19-X-Gus), y a la izquierda, precedido por “sp” está la actividad del promotor trunco correspondiente.

4.4. Discusión

Los puntos críticos del sistema son los siguientes: 1) Es necesario siempre hacer un seguimiento del proceso de digestión, y agregar más enzima en caso de ser necesario, debido a que la concentración o calidad de las enzimas

puede variar de lote en lote; 2) Aunque las células NT1 tienden a ser muy sincrónicas, no siempre están en las mismas condiciones y eso afecta el rendimiento de los protoplastos, haciendo que siempre haya que electroporar la construcción de control positivo, para que la expresión de GUS sirva para normalizar los datos, en función de la concentración de proteína total.

En el sistema montado la actividad de los promotores *Rep* “Datura”, “Ipomea” y “Sinaloa-22” fue mayor a la del promotor CaCMV35S, mientras que la de los begomovirus PHYVV y ToMoV (datos no mostrados) es un poco más baja. El promotor con mayor actividad caracterizado hasta ahora es el “Sinaloa22”, con una actividad tres veces mayor a la del 35S, pero hace falta caracterizar el resto de las construcciones, al igual que hacer las repeticiones restantes de los experimentos que se muestran en la figura 4.8. Por otra parte, los datos preliminares indican que la actividad del promotor mínimo es más o menos similar en todas las construcciones, lo cual es lo que se espera.

Los datos de los promotores “Sinaloa22”, “Datura” e “Ipomea” discrepan de los resultados obtenidos por Astrid García Moreno-Rubli (Tesis de Maestría, 2005), y esto puede deberse a que sus ensayos también fueron experimentos preliminares mediante bombardeo de de hojas de chícharo con micropartículas de tungsteno cubiertas con el DNA a examinar. De éstos ensayos no se hicieron réplicas, y ese sistema tiene además la desventaja de que no es muy uniforme para reportar datos de expresión transitoria, ya que hay mucha variabilidad asociada al tejido de la hoja donde penetren las micropartículas.

4.5. Perspectivas

La preparación de protoplastos y la medición de actividad de β -glucuronidasa se lograron llevar a condiciones de repetitividad con bajas tasas de error, lo cual significa que el sistema queda listo para realizar experimentos con la rigurosidad que exigen los cánones internacionales.

Varias de las construcciones diseñadas quedan a la espera del análisis de actividad, al igual que faltan repeticiones de algunos ensayos. La culminación de estos ensayos, y ensayos adicionales en los que se cotransfecten los protoplastos con un vector que exprese la proteína Rep correspondiente, la de un virus de diferente especificidad, ó proteínas Rep híbridas, generaría resultados suficientes para definir que tan importante es la afinidad por la secuencia de iterones en la auto-regulación del promotor *Rep*.

4.6. Referencias

- Baliji S, Sunter J, Sunter G. 2007. Transcriptional analysis of complementary sense genes in Spinach curly top virus and functional role of C2 in pathogenesis. *MPMI*. 20: 194-206.
- Carter John & Saunders Venetia. *Virology: principles and applications*. John Wiley & Sons Ltd, West Sussex, England, 2007.
- Cazonelli CI, Velten J. 2008. In vivo characterization of plant promoter element interaction using synthetic promoters. *Transgenic Res*. 17:437-57.
- Crow RM, Gartland JS, McHugh AT, Gartland KM. 2006. Real-time GUS analysis using Q-PCR instrumentation. *J Biotechnol*. 126:135-9.
- Dinant S, Ripoll C, Pieper M, David C. 2004. Phloem specific expression driven by wheat dwarf geminivirus V-sense promoter in transgenic dicotyledonous species. *Physiol Plant*. 121:108-116.
- Dugdale B, Beetham PR, Becker DK, Harding RM, Dale JL. 1998. Promoter activity associated with the intergenic regions of banana bunchy top virus DNA-1 to -6 in transgenic tobacco and banana cells. *J Gen Virol*. 79:2301-11.
- Eagle PA, Hanley-Bowdoin L. 1997. cis elements that contribute to geminivirus transcriptional regulation and the efficiency of DNA replication. *J Virol*. 71:6947-55.
- Eini O, Behjatnia SA, Dogra S, Dry IB, Randles JW, Rezaian MA. 2009. Identification of sequence elements regulating promoter activity and replication of a monopartite begomovirus-associated DNA beta satellite. *J Gen Virol*. 90:253-60.
- Frey PM, Schärer-Hernández NG, Fütterer J, Potrykus I, Puonti-Kaerlas J. 2001. Simultaneous analysis of the bidirectional African cassava mosaic

- virus promoter activity using two different luciferase genes. *Virus Genes*. 22:231-42.
- Gartland KM, McHugh AT, Vitha S, Benes K, Irvine RJ, Gartland JS. 2000. Analysis of genetically modified plant gene expression using GUS fluorimetry. *Mol Biotechnol*. 14:235-9.
- Golenberg EM, Sather DN, Hancock LC, Buckley KJ, Villafranco NM, Bisaro DM. 2009. Development of a gene silencing DNA vector derived from a broad host range geminivirus. *Plant Methods*. 5:9.
- Guan C, Zhou X. 2006. Phloem specific promoter from a satellite associated with a DNA virus. *Virus Res*. 115:150-7.
- Huang Z, Chen Q, Hjelm B, Arntzen C, Mason H. 2009. A DNA replicon system for rapid high-level production of virus-like particles in plants. *Biotechnol Bioeng*. 103:706-14.
- Hur J, Choi E, Buckley KJ, Lee S, Davis KR. 2008. Identification of a promoter motif involved in Curtovirus sense-gene expression in transgenic *Arabidopsis*. *Mol Cells*. 26:131-9.
- Jefferson RA. 1987. Assaying chimeric genes in plants: the GUS gene fusion system. *Plant Mol. Biol. Rep*. 5:387-405.
- Lacatus G, Sunter G. 2008. Functional analysis of bipartite begomovirus coat protein promoter sequences. *Virology* 376:79-89.
- Nikovics K, Simidjieva J, Peres A, Ayaydin F, Pasternak T, Davies JW, Boulton MI, Dudits D, Horváth GV. 2001. Cell-cycle, phase-specific activation of Maize streak virus promoters. *Mol Plant Microbe Interact*. 14:609-17.
- Peretz Y, Mozes-Koch R, Akad F, Tanne E, Czosnek H, Sela I. 2007. A universal expression/silencing vector in plants. *Plant Physiol*. 145:1251-63.
- Regnard GL, Halley-Stott RP, Tanzer FL, Hitzeroth II, Rybicki EP. 2010. High level protein expression in plants through the use of a novel autonomously replicating geminivirus shuttle vector. *Plant Biotechnol J*. 8:38-46.
- Russell JA, Roy MK, Sanford JC. 1992. Major Improvements in Biolistic Transformation of Suspension-Cultured Tobacco Cells . *In Vitro Cellular & Developmental Biology*. 28:97-105.
- Shimada-Beltrán H, Rivera-Bustamante RF. 2007. Early and late gene expression in pepper huasteco yellow vein virus. *J Gen Virol*. 88:3145-53.
- Shirasawa-Seo N, Sano Y, Nakamura S, Murakami T, Seo S, Ohashi Y, Hashimoto Y, Matsumoto T. 2005. Characteristics of the promoters

derived from the single-stranded DNA components of Milk vetch dwarf virus in transgenic tobacco. *J Gen Virol.* 86:1851-60.

Shivaprasad PV, Akbergenov R, Trinks D, Rajeswaran R, Veluthambi K, Hohn T, Pooggin MM. 2005. Promoters, transcripts, and regulatory proteins of Mungbean yellow mosaic geminivirus. *J Virol.* 79:8149-63.

Shung CY, Sunter G. 2009. Regulation of Tomato golden mosaic virus AL2 and AL3 gene expression by a conserved upstream open reading frame. *Virology.* 383:310-8.

Singh DK, Malik PS, Choudhury NR, Mukherjee SK. 2008. MYMIV replication initiator protein (Rep): roles at the initiation and elongation steps of MYMIV DNA replication. *Virology.* 380:75-83.

Tu J, Sunter G. 2007. A conserved binding site within the Tomato golden mosaic virus AL-1629 promoter is necessary for expression of viral genes important for pathogenesis. *Virology.* 367:117-25.

Usharani KS, Periasamy M, Malathi VG. 2006. Studies on the activity of a bidirectional promoter of Mungbean yellow mosaic India virus by agroinfiltration. *Virus Res.* 119:154-62.

Velten J, Morey KJ, Cazzonelli CI. 2005. Plant viral intergenic DNA sequence repeats with transcription enhancing activity. *Virol J.* 2:16.

Xie Y, Liu Y, Meng M, Chen L, Zhu Z. 2003. Isolation and identification of a super strong plant promoter from cotton leaf curl Multan virus. *Plant Mol Biol.* 53:1-14.

Yang X, Baliji S, Buchmann RC, Wang H, Lindbo JA, Sunter G, Bisaro DM. 2007. Functional modulation of the geminivirus AL2 transcription factor and silencing suppressor by self-interaction. *J Virol.* 81:11972-81.

VIII. CONCLUSIONES GENERALES

Tras estudiar tres familias virales con genoma de DNA circular de cadena sencilla (*Geminiviridae*, *Nanoviridae* y *Circoviridae*), que se replican por el mecanismo de círculo rodante, los datos obtenidos de mayor relevancia son:

Después de estudiar decenas de replicones de los tres grupos, se establecieron relaciones claras entre las secuencias repetidas asociadas a la estructura tallo-asa que se forman en el origen de replicación por círculo rodante y la proteína Rep que las une. Esto deja abierto el camino hacia la caracterización molecular del dominio endonucleasa de las proteínas Rep.

Se obtuvo evidencia que indica que en todas estas familias la proteína Rep posee dos regiones que forman una interacción tipo lámina β , la cual participa en la unión específica al DNA del origen de replicación. Esto demuestra que estas familias a pesar de tener hospederos distintos comparten una historia evolutiva a través de su proteína iniciadora de la replicación.

La identificación de una constante en el funcionamiento del dominio endonucleasa de estas proteínas representa una ruta a seguir en el diseño de métodos de control para los miembros patógenicos de estos grupos.

Se contribuyó a aumentar el número de secuencias de curtovirus reportadas en la base de datos GeneBank, mediante el depósito de la secuencia de una especie en el género Curtovirus (*Geminiviridae*), y de tres aislados del virus del rizado de la punta del betabel, BMCTV, identificados en un cultivo de Chile en Villa de Arista, San Luis Potosí.

Los aislados de BCTV encontrados en México representan una ampliación del rango de distribución conocido de los curtovirus y confirman la presencia de este tipo de virus en el país, lo cual tiene varias implicaciones fitosanitarias.

Se replanteó la evolución del género curtovirus, proponiendo un escenario evolutivo basado en evidencias contenidas en el genoma de los miembros del grupo. La información obtenida indica que la mayoría de virus de éste género han evolucionado hace poco en Norteamérica, tras adquirir porciones de un begomovirus que permaneció aislado por millones de años en Sudamérica

Adicionalmente, se dejó montado un sistema de preparación de células vegetales para probar, de manera transitoria, la actividad de promotores y otros elementos *in cis* involucrados en la regulación transcripcional de los begomovirus.

El trabajo experimental, consistente en montar un sistema de cultivo de células en suspensión y preparación de protoplastos amplía el abanico de metodologías experimentales del grupo de trabajo.

IX.ANEXOS

Anexo 1. Protocolos experimentales

Precipitación de DNA

1. Agregar 1/10 de volumen de Acetato de sodio 3M (ó >) al DNA
2. Agregar 2 volúmenes de etanol absoluto (al menos > 80%)
3. Poner en frío (-4 ó -20 °C) por unos 15 min
4. Centrifugar 10 min a 13000 rpm
5. Descartar sobrenadante y lavar la pastilla con 500 ul de etanol al 70% (preferiblemente frío)
6. Descartar sobrenadante y secar la pastilla a temperatura ambiente
7. Resuspender en TE ó H₂O, poner a 65°C para mezclar completamente.

Soluciones:

TE pH 8.0 (80 ml)

<i>Reactivo</i>	<i>Cantidad</i>	<i>Conc. final</i>
Tris 1.0 M pH 8.0	0.8 ml	10 mM
EDTA 0.5 M pH 8.0	0.16 ml	1.0 mM

Aforar al volumen planeado, ajustar el pH a 8.0 y esterilizar en autoclave

Tris 1.0 M (80 ml)

<i>Reactivo</i>	<i>Cantidad</i>
Tris base	9.68g

Aforar, ajustar pH a 8.0 con HCl concentrado y esterilizar en autoclave

EDTA 0.5 M (80 ml)

<i>Reactivo</i>	<i>Cantidad</i>
EDTA-Na ₂ -2H ₂ O	14.89g
H ₂ O	60 ml

Ajustar pH a 8.0 con NaOH 5M, aforar a 80 ml y esterilizar en autoclave

Extracción de DNA de muestras vegetales (Método Dellaporta modificado)

1. Pesar 50 mg del tejido vegetal, moler en nitrógeno líquido con ayuda de un pistilo.
2. Agregar 480 ul de buffer de extracción
3. Adicionar 37.5 ul de SDS al 20%, mezclar por inversión
4. Calentar a 65°C por 10 minutos
5. Enfriar a temperatura ambiente por 5 min
6. Agregar 94 ul de Acetato de potasio 5M, mezclar por inversión
7. Colocar a 4°C por 5 min
8. Centrifugar a 13000 rpm por 5 min
9. Transferir el sobrenadante a tubo nuevo
10. Adicionar un volumen de fenol-cloroformo 1:1, mezclar por vortex
11. Centrifugar a 13000 rpm por 3 min
12. Transferir cuidadosamente la fase acuosa (capa superior) a tubos nuevos
13. Adicionar un volumen de fenol-cloroformo-álcool isoamílico 25:25:1. Mezclar por vortex
14. Centrifugar a 13000 rpm por 3 min
15. Recuperar la fase acuosa en tubos nuevos
16. Agregar 10 ul de RNAsa e incubar a temperatura ambiente por 30 min
17. Agregar 600 ul (1 vol) de isopropanol frío, mezclar por inversión
18. Incubar en hielo por 5 min
19. Centrifugar a 13000 rpm por 5 min
20. Descartar sobrenadante y lavar pastilla con 500 ul de etanol al 70%
21. Dejar secar a temperatura ambiente, o max. A 37°C
22. Resuspender en 50 ul de TE pH 8.0

Digestión

-Reacción:

Agua	26.0
Buffer	4.0
DNA	7.0
Enzima	<u>3.0</u>
	40.0

Incubar 1:1/2 horas a 37°C

-Detalles: La cantidad de enzima(s) no debe superar el 10% de la reacción porque el exceso de glicerol genera efecto estrella. Fijarse siempre en el Buffer recomendado, y para digestiones dobles usar la tabla de ajuste del Buffer, si no hay ajuste, hacer las digestiones por pasos. En caso de digestión parcial se puede precipitar la primera digestión e iniciar una nueva reacción desde el precipitado, o se puede agregar más enzima a la misma reacción, aumentando el volumen de reacción y ajustando la cantidad de buffer.

Ligación

-Reacción:

Agua	5.0
Buffer ligasa 5X	2.0
Vector	1.0 (siempre en menor proporción que el inserto, ej. diluido 1/10)
Inserto	1.0
T4 DNA Ligasa	<u>1.0</u>
	10.0

Incubar 1-2 horas a 25°C, usar 4.0 ul para transformar *E. coli*, conservar el resto para uso posterior.

Defosforilación con fosfatasa alcalina de camarón (SAP)

-Reacción:

Agua	3.5
Buffer SAP	5.0
DNA	40.0
Enzima	<u>1.5</u>
	50.0

Incubar 1 hora a 37°C, inactivar incubando 15 min a 65°C, purificar por columna ó precipitar.

Relleno de extremos cohesivos con fragmento Klenow de la DNA polimerasa

-Reacción:

Agua	12.2
Mezcla dNTPs 2mM	1.0
Buffer 10X	4.0
DNA digerido	22.0
Klenow	<u>0.8</u>
	40.0

Incubar 10 min a 37°C, inactivar incubando 10 min a 70°C, purificar por columna ó precipitar.

Fosforilación de fragmentos con polinucleótido cinasa

-Reacción:

Agua	11.5
Buffer 5Xfor	5.0
ATP 10mM	2.5
DNA	5.0
T4 cinasa	<u>1.0</u>
	25.0

Incubar 10 min a 37°C, inactivar incubando 10 min a 65°C, purificar con fenol-cloroformo y lavar con etanol.

Reacción en cadena de la polimerasa

-Reacción:

Agua	32.1
Buffer 10X	5.0
MgCl ₂ 25 mM	3.0
dNTPs 10 mM	1.16
Oligo for 10 pM	2.3
Oligo rev 10 pM	2.3
DNA	5.0
Polimerasa	<u>2.54</u>
	50

-Ciclo de amplificación:

Desnaturalización inicial	94°C 2 min
Amplificación (35X)	94°C 30 seg
	56°C 30 seg
	72°C 30 seg
Extensión final	72°C 5 min

-Detalles: Este protocolo es sólo para una pareja de oligos determinada. Las concentraciones de oligos, dNTPs, magnesio, DNA y polimerasa pueden variar según el tamaño del producto esperado, abundancia del fragmento a amplificar, etc; por las mismas razones varían los protocolos del ciclo de amplificación

Transformación de *E. coli* TOP 10 por choque térmico

1. Descongelar las células competentes en hielo durante aprox. 5 min.
2. Agregar 1.0 ul de DNA si se trata de un vector ó 5 ul si es producto de ligación (un aproximado de 100 ng).
3. Colocar el tubo en hielo durante 20-30 min.
4. Colocar el tubo en baño María a 42°C por 1.5 min (funciona bien con 1 min).
5. Reposar los tubos en hielo durante 10 min.
6. Adicionar 250 ul de LB (sin antibiótico), en campana.
7. Incubar a 37°C durante 45 min, con agitación constante.
8. Sembrar 100 ul de las células sobre cajas con antibiótico, y según la construcción, agregar previamente X-gal (15 ul) e IPTG (40 ul).
9. Incubar a 37°C durante toda la noche.

Minipreparación de DNA

Soluciones:

Birnboim I (150 ml)

Reactivo	Cantidad	Conc. final
Glucosa	1.35g	50mM
Tris 1.0 M pH 8.0	3.75ml	25 mM
EDTA 0.5 M pH 8.0	3 ml	10 mM

Esterilizar en autoclave

Birnboim II (10 ml) SE PREPARA AL MOMENTO

Reactivo	Cantidad
NaOH 5M	400 ul
SDS 20%	500 ul

Aforar con agua destilada

Birnboim III (100 ml)

Reactivo	Cantidad	Conc. final
Acetato de potasio	29.5g	3M
Ácido acético glacial	11.5ml	

Esterilizar en autoclave

Procedimiento:

1. Picar las colonias de interés y ponerlas a crecer en 3.0 ml de LB con antibiótico desde la noche anterior.
2. Centrifugar 1-1.5 ml del cultivo durante 3 min.
3. Descartar medio sobrenadante.
4. Agregar 100 ul de Birnboim I al pellet de células, mezclar bien usando Vortex
5. Agregar 200 ul de Birnboim II (esta sln no se debe conservar más de una semana)
6. Mezclar por inversión, sin brusquedad
7. Agregar ½ del vol anterior de Birnboim III, mezclar sin Vortex
8. Poner en hielo durante 5 min
9. Centrifugar 3-5 min
10. Transferir el sobrenadante a un Eppendorf nuevo
11. Agregar 900 ul de etanol absoluto
12. Poner en hielo por 5-10 min
13. Centrifugar 3-5 min
14. Lavar con 500 ul de etanol al 70%
15. Resuspender en TE ó H₂O

Maxipreps

1. Incubar preinóculo en 200 ml de LB con antibiótico de 15 a 18 h
2. Cosechar las células centrifugando a 6900 rpm/10 min
3. Eliminar el sobrenadante y resuspender en 5 ml de solución Birnboim I.
4. Adicionar 10 ml de Birnboim II, mezclar por inversión
5. Incubar 5 min a temperatura ambiente
6. Agregar 7.5 ul de Birnboim II, mezclar por inversión
7. Incubar en hielo durante 10 min
8. Centrifugar a 100000 rpm (10K)/10 min
9. Transferir el sobrenadante a tubo limpio
10. Adicionar 18 ml de isopropanol frío (0.8 a 1 vol aprox.), mezclar por inversión
11. Reposar 20 min en hielo
12. Centrifugar a 10K por 20 min y eliminar sobrenadante
13. Lavar la pastilla con 5 ml de etanol al 70%
14. Centrifugar a 10K/5 min
15. Eliminar el sobrenadante y secar la pastilla a temperatura ambiente

16. Resuspender en 500 ul de TE pH 8.0 y trasferir a tubos Eppendorf.
17. Adicionar 1 vol de fenol-cloroformo 1:1 y mezclar por vortex
18. Centrifugar a 13K/3 min
19. Transferir la fase superior a un tubo nuevo
20. Adicionar 1 vol de fenol-cloroformo-alcohol isoamílico 25:25:1 y mezclar por vortex
21. Centrifugar a 13K/3 min
22. Transferir sobrenadante a tubo nuevo
23. Adicionar 3 ul de RNasa e incubar 30 min a temperatura ambiente
24. Agregar 1/10 de acetato de sodio 3M y 2 vol de etanol absoluto
25. Reposar en hielo 15 minutos
26. Centrifugar a 13K/10 min y descartar sobrenadante
27. Lavar la pastilla con 500 ul de etanol 70%
28. Centrifugar a 13K/3 min, descartar sobrenadante y secar
29. Resuspender en 250 ul de TE pH 8.0

Cultivo de células de tabaco (Línea NT1)

a) Cultivo en suspensión

Este se hace con el fin de mantener la línea celular. Las células se crecen en volúmenes de unos 100 ml de medio NT1 líquido, en oscuridad y con agitación constante a 125 rpm. El cambio de medio de cultivo se hace cada 7-10 días, vaciando una alícuota de unos 20 ml de cultivo al nuevo frasco recién esterilizado con el medio NT1 (a T° ambiente).

Medio líquido NT1

1. Preparar las siguientes soluciones stock

Solución	Contenido	Cantidad (g)/ 250 mL (stock 100X)
I (Nitratos)	Nitrato de amonio Nitrato de potasio	41.25 47.5
II (Sulfatos)	Sulfato de magnesio 7H ₂ O Sulfato de manganeso H ₂ O Sulfato de Zinc 7H ₂ O Sulfato de cobre 5H ₂ O	8.57 0.4225 0.215 0.000625
III (Halógenos)	Cloruro de calcio 2H ₂ O Yoduro de potasio Cloruro de cobalto 6H ₂ O	11 0.021 0.000625
IV (Fosfatos)	KH ₂ PO ₄ Ácido bórico Na ₂ MoO ₄	4.25 0.155 0.000625
V (Quelatos y vitaminas)	FeSO ₄ 7H ₂ O EDTA 2H ₂ O Myo-inositol Tiamina HCL	0.695 0.9325 2.5 0.025

Almacenar la solución I a temperatura ambiente y las soluciones II-V a -4°C. Preparar por separado un stock de ácido 2-4, Diclorofenoxiacético (2-4, D) a una concentración de 1.0 mg/ml. Almacenar a -20°C.

2. Preparar el Medio NT1 de la siguiente manera:

Para un litro:

- Agregar 10 ml de cada una de las soluciones stock en 500 ml de agua destilada
- Agregar 30 g de sacarosa
- Agregar 2 ml de 2-4,D 1.0 mg/ml
- Aforar a 1.0 L
- Ajustar a un pH entre 5.2 – 5.7 con KOH
- Esterilizar al momento de la preparación y almacenar a -4°C
- Esterilizar nuevamente cada alícuota que se vaya a usar para cambio de medio

b) Cultivos en medio sólido

Se usan para bombardeo de células.

Medio NT1 sólido

A 1.0 L de NT1 líquido, agregar 2.5g de Gelrite

Medio NT1 osmótico

A 1.0 L de NT1 líquido, agregar 2.5g de Gelrite y 45.5g de Manitol

Preparación de protoplastos de células NT1 para ensayos de expresión transitoria

<u>Reactivos</u>	<u>Materiales</u>
Manitol	Agitador orbital
MES (morpholineethanesulfonic acid)	Cajas Petri 100 x 25 mm
Celulasa de <i>T. viride</i>	Filtros 0.22 um
Pectoliasa de <i>A. japonicum</i>	Centrifuga
Solución enzimática	Cubetas de electroproración de 0.4 cm
Medio de cultivo para protoplastos	Electroporador
Buffer de electroporación	Cajas Petri 30 x 15 mm

Detalles de soluciones:

a. Solución enzimática

50 ml deben alcanzar para generar material para 50 electroporaciones.

<i>Reactivo</i>	<i>Cantidad/50 ml</i>	<i>Conc. Final</i>
Manitol	3.64g	0.4M
MES	0.213g	20 mM
Celulasa	0.5g	1%
Pectoliasa	0.05g	0.1%

Disolver todo en agua destilada estéril durante toda la noche a 4°C. Esterilizar la solución pasándola a través de un filtro de 0.22 um. Poner a temperatura ambiente antes de usar.

b. Medio de cultivo de protoplastos

Medio NT1 líquido + Manitol a concentración final 0.4 M (72.86 g para 1.0 L). Ajustar el pH entre 5.5- 5.7 y autoclavar. Almacenar protegido de la luz.

c. Buffer de electroporación

<i>Reactivo</i>	<i>Cantidad/ 500 ml</i>	<i>Conc. Final</i>
NaCl	4g	0.8%
KCl	0.1g	0.02%
KH ₂ PO ₄	0.1g	0.02%
Na ₂ HPO ₄	5.5g	0.11%
Manitol	36.43g	0.4M

Ajustar pH a 6.5 con HCl y esterilizar

Procedimiento:

1. Mantener las células NT1 en fase logarítmica subcultivando cada cuatro días, al menos dos pases antes.
2. En campana de flujo laminar, transferir 15 ml del cultivo de células NT1 (3-4 días después del subcultivo, al final se obtiene un paquete +/- de 1.0 ml de células, suficiente para seis electroporaciones) a un tubo cónico y centrifugar a 1000 rpm por 2 minutos a temperatura ambiente.
3. Retirar el sobrenadante
4. Lavar con Manitol 0.4M
5. Centrifugar a 1000 rpm por 2 min y retirar sobrenadante
6. Agregar 1.5 volúmenes de solución enzimática por cada volumen de células compactadas
7. Mezclar lentamente por inversión hasta que todo el pellet esté disuelto.
8. Incubar en el mismo tubo, a 25°C con agitación a 65 rpm por unos 45 minutos
9. Para evaluar la eficiencia de la preparación, observarla en microscopio de luz a los 30 minutos y mientras se estandariza el protocolo, hacerlo cada media hora hasta establecer el momento en el que más del 95% de las células adquieren una forma redondeada.

Nota: No se recomienda dejarlos en la solución enzimática más de una hora. En caso de no irse a usar inmediatamente, se centrifugan a 1000 rpm por 2 min y se retira cuidadosamente el líquido con pipetas Pasteur (los protoplastos pueden quedar flotando); se lavan con manitol 0.4 M y se ponen en frascos con medio de cultivo para protoplastos. Pueden ser cultivados en oscuridad por 1-2 días, sin agitación, pero al tercer día ya tendrán la pared completamente regenerada.

Electroporación de protoplastos de células NT1

Preparación del DNA:

1. Cuantificar el DNA plásmidico a transfectar y tener disponibles 15 ug de DNA purificado por cada transfección a realizar.
2. Precipitar el DNA a electroporar y eluirlo en buffer de electroporación a una concentración de 1ug/ul.

Electroporación:

1. Lavar muy bien y esterilizar las celdillas, y tenerlas en hielo.

2. Colectar protoplastos a la hora de incubación a 25°C (o el tiempo al que un 95% de las células se ven redondas)
 3. Centrifugar a 1000 rpm por 2 min
 4. Retirar cuidadosamente el líquido con pipeta Pasteur
 5. Agregar 15 ml de Manitol 0.4 M
 6. Centrifugar a 1000 rpm 2 min
 7. Repetir pasos 3-5
 8. Repetir pasos 3-6, lavando ahora dos veces con 15 ml de Buffer de electroporación
 9. Resuspender las células en unos 10 ml de Buffer de electroporación
- Nota útil:** Una buena preparación de protoplastos se puede resuspender en 2 volúmenes de éste buffer y tendrá en promedio la concentración final requerida, evitando el conteo de células.
10. Contar la cantidad de células por ml poniendo 100 ul en la cámara de Neubauer, contar en los cuatro cuadros de las esquinas y en el del centro.
 11. Diluir las células en Buffer de electroporación hasta una concentración de $3-5 \times 10^6$ células/ml, poner en hielo y electroporar cuanto antes, ya que este buffer no es muy favorable para las células.
-
12. Mezclar cuidadosamente en un tubo las células con el DNA. Se ponen 400 ul de células resuspendidas por cada 15 ug de DNA a electroporar.
 13. Pasar los 400 ul de células + DNA a las celdas de electroporación.
 14. Dejar 5 min a temperatura ambiente
 15. Electroporar a 500 uF y 250 volts, en el protocolo de voltaje constante.
 16. Reposar las células en hielo durante 15 minutos
 17. Colectar las células electroporadas lavando cuidadosamente la celdilla con medio NT1+ manitol 0.4M y ponerlas en cajas Petri pequeñas (Un volumen final de 7 ml de cultivo de protoplastos en NT1+manitol).
 18. Incubar a 25 °C por 48 horas, sin agitación.

Cuantificación de proteínas por el método de Bradford

<u>Soluciones</u>	<u>Materiales</u>
Agua destilada estéril	
Etanol absoluto	Espectrofotómetro
Ácido fosfórico	Puntas, pipetas
Azul de Coomasie	
Albúmina sérica de bovino (BSA)	

Detalles de soluciones

Solución A

<i>Reactivo</i>	<i>Cantidad</i>
Etanol 95%	25 ml
Ácido fosfórico 85%	50 ml
Azul de Coomasie	87.5 mg

Mezclar y agitar hasta disolver completamente. Filtrar con filtro de 0.22 um y almacenar a 4°C.

Solución B (Reactivo de Bradford)

<i>Reactivo</i>	<i>Cantidad</i>
Etanol 95%	7.5 ml
Ácido fosfórico 85%	15 ml
Solución A	15 ml

Aforar a 250 ml con agua destilada estéril y almacenar a 4°C.

Curva estándar de BSA

Hacer un stock de BSA a 100mg/ml en el mismo buffer de fosfatos en que se tiene la muestra de proteínas a cuantificar. La curva recomendable para extractos vegetales: 0, 1, 5, 10 y 20 ug/ul.

Procedimiento

1. Preparar varias celdillas, para el estándar y las muestras. Mantener en frío.
2. Poner 20 ul de la solución problema (y/o dilución de ésta en caso necesario) en 1 ml de Solución B. Hacer igual para las diluciones de la curva.
3. Mezclar bien e incubar 5 minutos a temperatura ambiente.
4. Leer la absorbancia a 595 nm
5. Graficar valores de la curva estándar y extrapolar lectura de la muestra problema.

Ensayo de actividad de GUS (*uidA*), adaptado para Fluorómetro lector de microplacas

Soluciones:

a) Buffer de extracción de proteínas (Buffer de extracción de GUS)

<i>Reactivo</i>	<i>Cantidad/500 ml</i>	<i>Conc. Final</i>
Fosfato de sodio 0.5 M pH 7.0	50 ml	50 mM
Ditiotreitol	0.771g	10 mM
Na ₂ EDTA	0.093g	10 mM
Lauril-sarcosina	0.5g	0.1% w/v
Triton X-100 10% v/v	50 ml	0.1% v/v

b) Buffer de ensayo de GUS

<i>Reactivo</i>	<i>Cantidad/10 ml</i>	<i>Conc. final</i>
Buffer de extracción de GUS	10 ml	
4-Metil-umberil-β-glucuronido (MUG)	3.52mg	1 mM

Alicuotar en volúmenes de 1.0 ml y almacenar a -80°C.

c) Buffer de parada de la reacción

<i>Reactivo</i>	<i>Cantidad/200 ml</i>	<i>Conc. final</i>
Carbonato de sodio (Na ₂ CO ₃)	4.24g	0.2 M

d) Curva estándar de metil umbeliferona

<i>Reactivo</i>	<i>Cantidad/10 ml</i>	<i>Conc. final</i>
-----------------	-----------------------	--------------------

Metil umbeliferona de sodio (NaMU) 1.982 mg 1 uM
Buffer de parada de reacción 10 ml

Hacer las diluciones necesarias para una curva entre 0 y 5 nM de Na-metil umbeliferona.

Procedimiento:

1. Pesar 100 mg de tejido o de células en cultivo, incluyendo controles negativos. Almacenar a -80°C mientras se usan.
2. Triturar el tejido con pistilo estéril, usando nitrógeno líquido en los casos necesarios.
3. Agregar 500 ul de buffer de extracción de proteína
4. Centrifugar a 13000 rpm por 10 min
5. Transferir el sobrenadante a un tubo Eppendorf nuevo. Almacenar a -80°C (NUNCA A -20) hasta su uso.
6. Cuantificar el contenido de proteína total (Ver método de Bradford)
7. ---
8. En una serie de tubos nueva servir 60 ul de buffer de ensayo de GUS. Agregar a los tubos anteriores 6.0 ul de cada extracto de proteínas, mezclar bien.
9. Poner 18 ul de la dilución anterior en 182 ul de Buffer de parada previamente servido en la microplaca (Tiempo 0). Mezclar bien
10. Incubar el volumen restante en los tubos a 37°C durante media o una hora, tomando alícuotas de 18 ul a cada intervalo que se quiera hacer una medición (Tiempos 1 y 2)*, y poniéndolas en el buffer de parada en el pozo respectivo.
11. Para la curva estándar de Metil Umbeliferona servir 200 ul de cada dilución de la curva en el pozo correspondiente (por triplicado).
12. Leer en fluorómetro a longitud de onda de excitación de 365 nm y emisión de 450 nm. (el rango de excitación de la metil umbeliferona esta entre 360 y 372 nm, y el de emisión entre 440 y 470 nm).
13. *Los intervalos de tiempo se pueden modificar según convenga.

Reportar los datos como:

$$\text{Actividad GUS} = \frac{\text{nmoles MUG hidrolizados}}{\text{Hora}}$$

Nota. Si se tienen los datos de cuantificación de proteínas totales reportar como:

$$\text{Actividad GUS} = \frac{\text{nmoles MUG hidrolizados}/\mu\text{g proteína total}}{\text{Hora}}$$

Lectura clásica de actividad β -glucuronidasa en fluorómetro de celdas de vidrio:

14. En una serie de tubos nueva servir 250 ul de buffer de ensayo de GUS. Agregar a los tubos anteriores 25 ul de cada extracto de proteínas, mezclar bien.
15. Poner 100 ul de la mezcla anterior en 1900 ul de Buffer de parada previamente servido en la microplaca en tubos limpios (ésta será la lectura del tiempo 0).

16. Incubar el volumen restante en los tubos a 37°C durante media o una hora, tomando alícuotas de 100 ul a cada intervalo que se quiera hacer una medición (Tiempos 1 y 2)*, y poniéndolas en el buffer de parada en tubos limpios.
17. Leer en fluorómetro a longitud de onda de excitación de 365 nm y emisión de 450 nm. (el rango de excitación de la metil umbeliferona esta entre 360 y 372 nm, y el de emisión entre 440 y 470 nm). Varios fluorómetros necesitan un valor de referencia de emisión de fluorescencia, más que un dato de la longitud de onda para poder empezar la lectura, usar 5000 para 1nM de NaMU en el fluorómetro Hoefer.
18. Para leer una curva estándar de Metil Umbeliferona se sirven 100 ul de cada dilución de la curva en 1900 ul de buffer de parada y se leen por triplicado.

Si se tienen los datos de cuantificación de proteínas totales reportar como:

$$\text{Actividad GUS} = \frac{\text{nmoles MUG hidrolizados}/\mu\text{g proteína total}}{\text{Hora}}$$

Anexo 2. Artículo aceptado en *Archives of Virology*

From: "ArchVirol Editorial Office" <jacquiline.dy@springer.com>
To: grarguel@ipicyt.edu.mx
Sent: 22 Feb 2010 19:06:53 -0500
Subject: Submission Confirmation for AVIROL-D-09-00552R1

Ref.: Ms. No. AVIROL-D-09-00552R1
DNA-binding specificity determinants of replication proteins encoded by eukaryotic ssDNA viruses are adjacent to widely separated RCR conserved motifs

Dear Dr. Arguello-Astorga,

Archives of Virology has received your revised submission. You may check the status of your manuscript by logging onto Editorial Manager at (<http://avirol.edmgr.com/>).

Kind regards,

Edward Rybicki, PhD
Editor Archives of Virology

From: "ArchVirol Editorial Office" <jacquiline.dy@springer.com>
To: grarguel@ipicyt.edu.mx
Sent: 8 Feb 2010 08:23:50 -0500
Subject: RE: Your Submission

Ref.: Ms. No. AVIROL-D-09-00552
DNA-binding specificity determinants of replication proteins encoded by eukaryotic ssDNA viruses are adjacent to widely separated RCR conserved motifs
Archives of Virology

Dear Gerardo,

Reviewers have now commented on your paper. You will see that they are advising that you revise your manuscript. If you are prepared to undertake the work required, I would be pleased to accept it.

For your guidance, reviewers' comments are appended below.

If you decide to revise the work, please submit a list of changes or a rebuttal against each point which is being raised when you submit the revised manuscript.

Your revision is due by 09 Apr 2010.

To submit a revision, go to <http://avirol.edmgr.com/> and log in as an Author. You will see a menu item call Submission Needing Revision. You will find your submission record there.

Yours sincerely

Edward Rybicki, PhD
Editor Archives of Virology

1 **DNA-binding specificity determinants of replication proteins encoded by eukaryotic**
2 **ssDNA viruses are adjacent to widely separated RCR conserved motifs**

3

4 Aurora Londoño, Lina Riego-Ruiz, Gerardo R. Argüello-Astorga*

5

6 División de Biología Molecular, Instituto Potosino de Investigación Científica y Tecnológica, San
7 Luis Potosí, México.

8

9 *Author for correspondence. Instituto Potosino de Investigación Científica y Tecnológica
10 (IPICYT), Camino a la Presa San José 2055, San Luis Potosí, México. C.P 78216. Phone: +52
11 (444) 8342000 Ext. 2079. Fax: +52 (444) 8342010. E-mail: grarguel@ipicyt.edu.mx

12

13

14 **Key words:** rolling-circle replication, Rep protein, iteron, nanovirus, circovirus, nanovirus-like
15 satellites, geminivirus.

16

17 **Running title:** Determinants for DNA-binding specificity of virus replication proteins

1 **Abstract**

2

3 Eukaryotic ssDNA viruses encode a rolling-circle replication (RCR) initiation protein, Rep, which
4 binds to iterated DNA elements functioning as essential elements for virus-specific replication.
5 By using the iterons of all known circoviruses, nanoviruses and nanovirus-like satellites as
6 heuristic devices, we have identified certain amino acid residues that presumably determine the
7 DNA-binding specificity of their Rep proteins. These putative “Specificity Determinants” (SPDs)
8 cluster in two discrete protein regions which are adjacent to distinct conserved motifs. A
9 comparable distribution of SPDs was uncovered in the Rep protein of geminiviruses. Modeling
10 of the tertiary structure of diverse Rep proteins showed that SPD regions interact to form a
11 small β -sheet element, that has been proposed to be critical for high affinity DNA-binding of
12 Rep. Our findings indicate that eukaryotic circular ssDNA viruses have a common ancestor, and
13 suggest that SPDs present in replication initiators from a huge variety of viral and plasmid RCR
14 systems are associated with the same conserved motifs.

1 Introduction

2 A vast diversity of genetic systems spanning the three primary domains of life, Bacteria,
3 Archaea and Eukarya, multiply their genomes by the mechanism of rolling circle replication
4 (RCR), an asymmetric process in which synthesis of both leading and lagging DNA strands are
5 uncoupled [29]. The RCR mechanism has been well studied in a number of systems including
6 the ssDNA coliphage ϕ X174 [9], plasmids from Gram-positive bacteria like pMV158 and pT181
7 [28, 40], and plant viruses from the *Geminiviridae* family [15, 25]. All these genetic entities
8 encode a replication initiation protein (Rep) that binds DNA in a sequence-specific fashion and
9 possesses DNA nicking-closing activity. Initiation of RCR involves the binding of Rep to
10 particular sequence elements associated to the replication origin, where the protein introduces a
11 site- and strand-specific nick in a conserved nucleotide sequence generally located at the apex
12 of a potential stem-loop element [34]. The nick leaves a 3'-OH end that is used as a primer for
13 leading-strand synthesis by host DNA polymerases, while Rep stays covalently attached to the
14 5' end of the original plus strand. After one round of polymerization new binding and nick sites
15 for Rep are generated, and termination takes place by cleavage of the newly synthesized strand
16 and simultaneous ligation of the 5'- and 3'-ends of the parental plus strand linked to Rep [21,
17 34]. The diversity of proteins mediating initiation and termination of RCR is extraordinary, but a
18 broad class of them share certain sequence motifs which are arranged in a characteristic way,
19 thus defining a large superfamily of Rep proteins encoded by a variety of bacterial, archaeal and
20 eukaryotic replicons [24]. These RCR initiators have in common three sequence signatures:
21 motif 1 (Fu(t/u)(l/y)t/p), motif 2 (HuHuuu), and motif 3 (YxxKE/D), where "u" is a hydrophobic
22 amino acid residue. When these motifs were first described, no hypothetical function could be
23 assigned to motif 1, but it was postulated that motif 2 participates in divalent metal coordination
24 by binding Mg^{2+} or Mn^{2+} ions that are required for its catalytic activity, whereas motif 3 contains
25 the site-active tyrosine that attaches covalently to DNA [24, 31].

26
27 Three families of eukaryotic circular single-stranded DNA viruses are currently well-
28 characterized: *Nanoviridae*, *Circoviridae*, and *Geminiviridae*. All of them have very small
29 genomes and replicate through an RCR mechanism. Members of the family *Nanoviridae*,
30 classified into two genera, *Nanovirus* and *Babuvirus*, are plant pathogens that have a genome
31 composed of six to eight circular molecules of ssDNA ranging in sizes from 0.95 to 1.1kb
32 encapsidated in individual virions of 17-20 nm in diameter [14]. Each genomic component
33 encodes a single protein and includes a common region containing the origin of replication. The
34 so-called master Rep protein supports the replication of the multiple genomic components of a
35 nanovirus [59, 60]. In addition to authentic nanoviruses, in the last years a great number of
36 nanovirus-like satellites, previously called "DNA1", which are associated with whitefly-
37 transmitted geminiviruses (i.e., begomoviruses) have been described. These satellites are self-
38 replicating, circular ssDNA molecules that depend on the helper begomovirus for encapsidation,
39 movement and vector transmission, and do not seem to play an essential role in the
40 maintenance of the disease associated with the helper virus [4, 5]. The nanovirus-like satellites

1 have a replication origin exhibiting all the sequence signatures characteristic of *Nanoviridae* and
2 encode a single protein that is significantly similar (~45% of sequence identity) to Rep proteins
3 of *bona fide* nanoviruses [4].
4
5 The family *Circoviridae* is divided into two genera apparently unrelated: *Gyrovirus* and
6 *Circovirus*. The first genus includes *Chicken anemia virus* (CAV) that does not encode a protein
7 homologous to the RCR N-123-C initiators [24]. The genus *Circovirus* comprises mammal- and
8 bird-infecting viruses containing a monopartite ssDNA genome (1.7 to 2.1 kb in size), packed
9 into an icosahedral capsid of ~20 nm in diameter [11, 42, 61]. The genome of circoviruses
10 contains two major ORFs in an ambisense organization, one encoding the Rep protein, and the
11 other the capsid protein. The intergenic region contains the origin of replication that includes a
12 conserved sequence (5'-TAGTATTAC-3') flanked by inverted repeats, where Rep introduces a
13 nick to initiate virus RCR [11, 37]. Although the endonuclease domain of the circovirus Rep
14 (residues 1-110) is significantly similar to the equivalent domain of RCR initiators of
15 nanoviruses, these viral proteins are greatly divergent in their C-terminal domain [12, 13].
16
17 The largest group of eukaryotic circular ssDNA viruses is the family *Geminiviridae*, that includes
18 more than 200 species of plant-infecting viruses causing economically important diseases in a
19 variety of cereal and vegetable crops worldwide [10, 17, 45]. They have small genomes
20 consisting of one or two single-stranded circular DNA molecules (2.5- 3 kb in length) that are
21 encapsidated into geminated virions. The replication of geminiviruses initiate with the sequential
22 binding of Rep to a set of iterative sequences or "iterons" located at variable distances from a
23 potential stem-loop containing the conserved nonanucleotide 5'-TAATATTAC-3', where Rep
24 cleaves the positive strand of viral DNA to initiate the RCR process [15, 17, 34, 55]. The iterons
25 generally differ in nucleotide sequence among viral species, and are the major (but not the only)
26 *cis*-acting determinants of virus-specific replication [1, 17]. In an attempt to identify the *trans*-
27 acting Specificity Determinants (SPDs) of geminivirus replication, Arguello-Astorga and Ruiz-
28 Medrano [2] analyzed the predicted Rep proteins from more than 120 geminiviruses by a
29 comparative method that uses the iterons as heuristic devices. A hypervariable domain of Rep
30 whose aa sequence is similar among far-related viruses exhibiting identical iterons was
31 identified and termed "Iteron-Related Domain" (IRD). It was postulated that certain residues
32 within the IRD function as determinants of the specific-DNA binding properties of geminivirus
33 Rep proteins. The IRD is adjacent to the conserved RCR motif 1, and it was hypothesized that
34 this motif is, in fact, the core structural element of a novel DNA-binding domain possessing a β -
35 sheet as recognition subdomain [2]. This hypothesis was later supported by experimental data
36 from the three-dimensional structure of the endonuclease domain of *Tomato yellow leaf curl*
37 *Sardinia virus* Rep. The TYLCSV Rep structure was compared with known 3-D structures of
38 bovine papillomavirus E1 and SV40 Large-Tag, and it was found that the structural element of
39 geminivirus Rep which is equivalent to the dsDNA binding surface of the former viral replication

1 proteins, is a mini β -sheet composed by the β 1 and β 5 strands [6]. The TYLCSV Rep β 1-strand
2 (i.e., SIKa) is the IRD core sequence adjacent to motif 1 [2].

3
4 Recent studies of the replication of *Porcine circovirus* type 2 (PCV-2) and *Banana bunchy top*
5 *virus* (BBTV) demonstrated that Rep proteins of circoviruses and nanoviruses also recognize
6 and bind short iterated elements that are closely associated to the Rep nick-site [20, 35, 56].
7 This suggests that RCR initiators of these viral groups may well be analyzed by the comparative
8 method utilized to identify SPDs in geminivirus Rep proteins, which uses the iterons as heuristic
9 devices. Here we present the results of an extensive analysis of replication associated proteins
10 and DNA elements of all known circoviruses, nanoviruses and nanovirus-like satellites.
11 Additionally, the DNA-binding domain of geminivirus Rep proteins was re-examined to search
12 for extra, undetected SPDs located out of the IRD region. This comparative study revealed a
13 striking similarity in the relative position of putative SPDs in Rep proteins from all examined viral
14 systems, hence indicating an unequivocal evolutionary relationship among these groups of
15 ssDNA viruses.

16

17 **Materials and methods**

18 ***Virus sequences***

19 The genomic and protein sequences of geminiviruses, circoviruses, nanoviruses and nanovirus-
20 like satellites were downloaded from the NCBI-GenBank database. Viruses and satellites
21 names, acronyms and GenBank accession numbers are given in Online Resource 1.

22

23 ***Comparative approach***

24 The strategy to map the SPDs of RCR initiators encoded by eukaryotic ssDNA viruses was
25 implemented as follows: 1) Identification of putative DNA-binding sites (iterons) in all examined
26 replicons. 2) Classification of the proteins encoded by members from a viral or satellital lineage
27 into several “Iso-specific Protein Groups” (IsoPG), namely, clusters of Rep proteins with
28 equivalent DNA-binding specificity. 3) Comparative analysis of selected pairs of Rep proteins
29 belonging to different IsoPG, to define a minimal set of differential residues which are potentially
30 responsible for their differences in DNA-binding preferences. 4) The number of potential SPDs
31 is further minimized by sequential rounds of comparative analysis of differential residues with
32 their counterparts in members from the same IsoPG; thus, aa residues which are not conserved
33 within a given IsoPG are discarded as putative SPD. 5) If the IsoPG are diverse enough (e.g.,
34 with more than four divergent members), it is feasible to predict the actual residues that are
35 responsible for differences of DNA-binding specificity between the compared proteins. These
36 residues should be conserved in proteins of a particular IsoPG, and differ from the equivalent
37 residues of proteins of at least one distinct IsoPG. Supplementary Figure 1 (Online Resources
38 2) illustrates our comparative approach by showing three specific examples that include
39 members from five different IsoPG.

40

1 ***Definition of iso-specific groups of proteins***

2 The conserved nonanucleotide sequence in each viral genome (i.e., the Rep nick-site) which is
3 in the apex of a potential stem-loop element, was used as a point of reference to detect the
4 repeated DNA sequences that are bound by the cognate Rep protein. “Iterons” were considered
5 as short DNA repeats of five to eight nucleotides and, like in other RCR N-123-C systems
6 described [2, 52], they could be located close to one or both arms of the stem-loop structure.
7 The repeats were therefore visually searched between the nucleotides comprising the borders
8 of the stem loop element and the starting codon of the nearest ORF in both arms of the hairpin-
9 like structure. Each genome was analyzed without taking into account previous reports of
10 iterons. Once the iterated sequences were identified, the members of the different virus
11 lineages were clustered in groups exhibiting the same iteron sequence (i.e., IsoPG).

12
13 ***Alignments and phylogenetic reconstruction***

14 Paired alignments were obtained by the ClustalW method in the MegAlign application of the
15 Lasergene package (DNASTAR Inc., Madison, WI), using the default parameters. In some
16 cases the alignment was further improved by visual examination and manual adjustment.
17 Multiple alignments of protein sequences were performed using the ClustalW module in Mega
18 4.0 [58] using the PFAM matrix. Unless otherwise indicated, the same alignment method was
19 used to reconstruct phylogenies, which were done by Neighbour-joining within the Lasergene
20 package.

21
22 ***Theoretical models of the three-dimensional structures of Rep proteins***

23 The tertiary structure of the endonuclease domain of several Rep proteins was modeled using
24 the SwissModel server [53]. Prior to modeling, pGenTHREADER from the PsiPred server [36]
25 was used to determine the most suitable template for structural modeling. The validation for the
26 structural models obtained was performed with PROCHECK [33] and the overall stereochemical
27 quality of the protein was assessed by Ramachandran plot analysis at the MolProbity host [8].
28 The 3-D protein images were produced using the UCSF Chimera package from the Resource
29 for Biocomputing, Visualization, and Informatics at the University of California, San Francisco
30 [46].

31
32 **Results**

33 ***General Approach***

34 Three heuristic hypotheses were used in the present analysis: 1) the iterative sequences that
35 are closely associated to the Rep “nick-site” in the virus replication origin (*Ori*), constitute the
36 specific-binding sites for the RCR initiator; 2) certain aa residues within the DNA-binding domain
37 of Rep determine its preference for a specific iteron, hence acting as SPDs of this protein; and
38 3) homologous Rep proteins from viruses displaying dissimilar iterons should differ in one or
39 more DNA-binding SPDs and, conversely, proteins from viruses harboring identical iterons
40 should have similar aa residues in equivalent positions, regardless of their host range,

1 geographic origin or phylogenetic distance. Based on these assumptions, a strategy to map the
2 SPDs of the replication associated proteins of nanoviruses, nanovirus-like satellites, and
3 circoviruses was implemented (see Materials and methods). This approach is properly
4 described as a Comparative Analysis of Groups of Homologous Iso-specific Proteins (CAGHIP)
5 method.

6

7 ***Analysis of nanovirus-like satellites***

8 The usefulness of the CAGHIP approach to identify potential SPDs in DNA-binding proteins is
9 highly dependent on the number and sequence diversity of the members of the distinct IsoPG of
10 a given lineage [2]. Consequently, small viral taxa like the family *Nanoviridae* and the genus
11 *Circovirus* with only 6 and 12 known species, respectively, are not suitable by themselves for
12 this type of analysis. Nonetheless, a large number of subviral agents associated with
13 begomoviruses have been described in the last years, including more than 90 nanovirus-like
14 satellites. In view of its remarkable diversity, the collection of nanovirus-like replicons or
15 “alphasatellites”, as they have been recently renamed [41], is clearly fitted for analysis by the
16 CAGHIP method. Consequently, we started the search for Rep SPDs by examining the proteins
17 of those subviral entities.

18

19 The first phase of the analysis entailed the identification of the putative Rep-binding sites from
20 all the alphasatellites whose sequence was available at the NCBI-GenBank databases by
21 August 15, 2009. Ten different IsoPG were recognized; five of them contain at least seven
22 members, but four IsoPG include a single known component. The identified iterons exhibited
23 variations in the number of copies and position relative to the putative TATA box and the stem-
24 loop element (Fig.1a). Preliminary comparisons between IsoPG showed that differential aa
25 residues are mainly located in the 1-100 region of the protein, where the endonuclease domain
26 of *Faba bean necrotic yellows virus* (FBNYV) Rep protein has been delimited [64]. For example,
27 two alphasatellites associated to *Tomato yellow leaf curl virus* (Accession no. AJ579356 and
28 AJ888449) exhibiting different iterons (i.e., GGIICCC and GGAACCC, respectively) encode
29 RCR initiators 315 aa long that differ between them in only 13 residues, 11 of which are located
30 within the 1-68 protein region. Accordingly, subsequent comparative analyses were restricted to
31 the Rep N-terminal domain encompassing aa residues 1 to 120. After several cycles of cross
32 comparative analyses, four putative SPDs were identified in alphasatellite Reps. The first two
33 SPDs correspond to residues 5 and 7, whereas the second pair is located at either positions 59
34 and 61 (in seven IsoPG) or positions 53 and 55 (in three IsoPG). Interestingly, these putative
35 SPDs are contiguous to conserved aa sequences, namely, motif $\square 1$ (consensus: WCFTuFF)
36 encompassing residues 9 to 15, and motif $\square 2$ (consensus: HLQGuuQuKG) comprising either
37 residues 49 to 58 (in seven IsoPG) or 43 to 52 (in three IsoPG) (Fig. 1b). The predicted SPDs
38 are conserved in all members of a particular IsoPG, but none of the ten different IsoPG exhibit
39 the same combination of SPDs (see Fig. 1b). The chemical nature of the identified specificity

1 determinants is rather heterogeneous, including basic (R, K), acidic (E), strongly polar (Q),
2 weakly polar (S, T) and non-polar (A, V, L) amino acid residues.

3

4 ***Analysis of nanovirus Rep proteins***

5 The members of the family *Nanoviridae* have multipartite genomes (with the exception of
6 *Coconut foliar decay virus*, CFDV), and several of these genomic components encode a RCR
7 initiation protein, although only one seems to be essential for the infective process, the so-called
8 “master” Rep protein [22, 59, 60]. Four non-essential Rep-encoding DNAs have been described
9 in FBNYV and *Milk vetch dwarf virus* (MVDV), two in *Subterranean clover stunt virus* (SCSV),
10 and five in the babuvirus BBTV (Fig. 2a). Fifteen different iterons were identified in the Rep-
11 encoding nanovirus components. We were unable to identify in FBNYV-C2 (encoding the
12 master Rep) and its closest relatives SCSV-C8 and MVDV-C11, the typical iterative sequences
13 five to eight nt in length that display a tandem arrangement, common in other nanoviruses.
14 Therefore, the iteron-like motifs indicated in Figure 2b for FBNYV C-2 and its relatives
15 correspond to the sequences defined by Timchenko et al. [59, 60] as putative M-Rep binding
16 sites. Nine of the 15 distinct IsoPG included only one known member, a fact that hampers the
17 application of the CAGHIP method. For this reason, we firstly examined the few cases of highly
18 similar Rep proteins differing in their cognate iterons in order to find potential SPDs, and
19 subsequently all IsoPG were compared in the equivalent domains. The potential SPDs were
20 mapped in two discrete regions adjacent to conserved motifs n1 and n2, which display the
21 consensus WCFTuNn/f and HuQGy/fuXuK, respectively (Fig. 2b). The regions where the
22 putative SPDs congregate (i.e., SPD-r1 and SPD-r2) are identical or very similar between
23 proteins with identical cognate iterons, but different between proteins with distinct DNA-binding
24 sites (Fig.2b).

25

26 ***Analysis of circovirus Rep proteins***

27 Seven different types of *Ori*-associated iterons, organized in four distinctive arrangements, were
28 identified among the 12 recognized species of circoviruses (Fig. 3a). The distinct IsoPG were
29 very heterogeneous in terms of the number and sequence diversity of their members. Thus,
30 whereas three IsoPG included only one member (i.e., *Gull circovirus* (GuCV) [61], *Swan*
31 *circovirus* (SwCV) [16], and *Finch circovirus* (FiCV) [61]), the other four iso-specific groups
32 included at least 15 non-identical members each. For instance, the IsoPG including both the
33 non-pathogenic PCV type 1 and the pathogenic PCV type 2, is represented by more than 300
34 complete genomic sequences. The *Beak and feather disease virus* (BFDV) [19, 42] IsoPG
35 includes 45 isolates from four continents; the group containing to both *Goose circovirus* (GoCV)
36 and *Duck circovirus* (DuCV) [18] encompasses more than 40 completely sequenced DNAs, and
37 the highly diversified IsoPG that includes to *Columbid circovirus* (CoCV) [38], *Canary circovirus*
38 (CaCV) [47], *Raven circovirus*, (RaCV) [57], and *Starling circovirus* (StCV) [26] contains 16
39 members. A phylogeny of circoviruses derived from their predicted Rep proteins revealed the
40 existence of four major clades that match with the four different iteron arrangements revealed

1 by our analysis (Fig. 3a). Owing to the considerable divergence between circovirus Rep
2 proteins, no particular comparison between a pair of hetero-specific proteins allowed the
3 unambiguous identification of potential SPDs. Consequently, we used an alternative approach,
4 looking for “convergent” protein domains between RCR initiators from distantly related viruses
5 with identical iterons. For example, we aligned the predicted Rep proteins of the 13 known
6 isolates of CoCV with the homologous proteins of CaCV, RaCV, and StCV that belong to the
7 same IsoPG, and looked for segments exhibiting sequence conservation. Notwithstanding the
8 significant divergence of the endonuclease domain of those proteins (i.e., 73-65% aa identity), a
9 “convergent” segment (A/sAAKR) was identified adjacent to the conserved Rep motif c1 (Fig.
10 3b). The hypothesis that some residues in that pentapeptide stretch are SPDs is supported by
11 the fact that FiCV and GuCV, two close relatives of StCV and CaCV, exhibit divergent
12 sequences in the equivalent Rep segment (i.e., SPCKR and SGARR, respectively), in
13 accordance with their different DNA-binding affinities (Fig.3b). Likewise, DuCV and GoCV Rep
14 proteins exhibit a GNYSYKR sequence adjacent to motif c1, that is different to the equivalent
15 segment (i.e., SDYGYKR) of the protein encoded by SwCV, a close relative of GoCV with
16 distinct iterons (Fig. 3b). On the other hand, residues homologous to the pair of SPDs located
17 near to motif \square 2 of alphasatellite proteins are also conserved in the different circovirus IsoPG.
18 (Fig. 3b). Together, these observations suggest that aa residues adjacent to the conserved
19 motifs c1 and c2 of circovirus Rep proteins are, plausibly, determinants of their DNA binding
20 properties.

21

22 ***Analysis of geminivirus Rep proteins***

23 A previous study of geminivirus Rep proteins identified a short domain (i.e., the “IRD”) adjacent
24 to the RCR Motif 1 where all discernible SPDs were mapped [2], and subsequent analysis of
25 Rep proteins encoded by bipartite geminiviruses forming infectious reassortants discovered one
26 additional SPD out of the IRD region [49]. This putative SPD is located within a structural
27 element termed \square 5-strand, identified in a study of the 3-D structure of TYLCSV Rep [6].
28 Because the homologous residues of this distal SPD do not consistently vary among
29 geminivirus proteins differing in DNA-binding specificity, we systematically re-examine the
30 sequence variations in the endonuclease domain (residues 1-120) of geminivirus Rep proteins,
31 looking for potential SPDs not located in the IRD region. After an extensive analysis
32 encompassing 170 of the ~200 described geminivirus species [10], two SPDs not associated to
33 the IRD were identified. One of them is the same residue (at position 69) mapped by Ramos et
34 al. [49] in the protein of *Tomato mottle Taino virus* (ToMoTV), while the second one is located
35 two positions ahead.

36

37 Figure 4 illustrates the three general cases found in this new analysis: 1) proteins differing in
38 IRD residues, but identical in the \square 5 element; 2) proteins diverging in only one \square 5-strand
39 residue, and in one or more IRD residues; and 3) proteins differing in IRD sequence and in two
40 residues of the \square 5 region. Additional cases from geminivirus proteins with different cognate

1 iterons are presented in Online Resources 3. This new analysis of geminivirus Reps revealed
2 that potential SPDs cluster in two discrete regions, separated by ~60 intermediate aa residues.
3 This distribution of SPDs is reminiscent of the one observed in RCR initiators which are
4 encoded by alphasatellites and related ssDNA viruses, that apparently are not evolutionarily
5 related to geminiviruses [12, 32, 41, 50].

6

7 **Comparative analysis of SPD positions in viral Rep proteins.**

8 With the purpose of comparing the relative position of predicted SPDs in the RCR initiators
9 considered in this study, an alignment of the endonuclease domain sequences from Rep
10 proteins encoded by several ssDNA viruses was carried out. The alignments exposed two
11 relevant features of those proteins: 1) the canonical RCR motifs 1 and 2 from geminivirus Rep
12 are apparently homologous to the first two conserved motifs from RCR initiators encoded by
13 alphasatellites, nanoviruses and circoviruses, in spite of their low sequence identities; and 2)
14 the position of the SPDs with respect to the RCR motifs 1 and 2 is analogous in all compared
15 viral proteins. For instance, the predicted SPDs proximal to the Rep N-terminus cluster in a
16 small amino acid stretch consistently separated by 3-4 residues from the F(t/l)(t/l)(y/n) core
17 sequence of the first conserved motif. Likewise, the SPDs located near to the second conserved
18 motif, are separated from it by a constant number of aa residues. Thus, in proteins of
19 nanoviruses, alphasatellites and circoviruses, the predicted SPDs are invariably situated at
20 positions 8 and 10 ahead of the HuQ core of motifs n2, □2, and c2, respectively, whereas the
21 SPDs of geminivirus proteins are consistently located at residues 10 and 12 in front of the HuH
22 core of motif 2 (Fig.5).

23

24 **Modeling of the tertiary structure of Rep endonuclease domain.**

25 The clustering of viral Rep SPDs in two discrete proteins regions separated by 50-60 aa
26 residues might be explained by the folding of the endonuclease domain in a three-dimensional
27 structure. This structure bring together the residues adjacent to the N-end of motif 1 (or its
28 equivalents) and those ~10 positions ahead of the HuH/ HuQ core of motifs 2/□2, as observed
29 in the solution NMR 3-D structure of the catalytic domain of TYLCSV [6] and PCV-2 Rep
30 proteins [63]. In these cases a double stranded mini □-sheet (i.e., □1/□5) was identified as the
31 structural element most probably involved in dsDNA recognition. However, the 3-D structure of
32 the FBYNV master Rep did not reveal a mini □-sheet equivalent to the □1/□5 element of
33 TYLCSV and PCV-2 proteins [64]. The absence of the latter structural element in FBYNV M-
34 Rep is significant because this is the only viral protein included in our analysis that does not
35 have an easily recognizable cognate iteron, as previously mentioned. Considering that FBYNV
36 M-Rep might not be the most appropriate model for all Rep proteins of nanovirus-like systems,
37 a theoretical modeling of the tertiary structure of the catalytic domain of nanovirus and
38 alphasatellite RCR initiators was performed, using as template the homologous domain of PCV-
39 2 Rep (see Materials and methods). The modeled 3-D structures of Rep proteins from a
40 babuvirus, an alphasatellite and a bird-infecting circovirus, are shown in Figure 6a, where the

1 predicted tertiary structure of a geminivirus Rep is also illustrated. In Fig. 6a it is evident that in
2 all the 3-D Rep structures the regions containing the predicted SPDs (SPD-r, in red) form a
3 structural element equivalent to the mini α 1/ α 5 sheet of TYLCSV and PCV-2 Reps, thus
4 indicating that this structural element presumably involved in dsDNA recognition [6, 63, 64] is
5 conserved in RCR initiators of circular ssDNA viral systems.

6
7 To obtain further insight into how amino acid residues in the α 1- and α 5-strands influence the
8 DNA binding specificity of geminivirus Rep, we carried out a comparison of the mini α -sheet in
9 the predicted tertiary structure of Rep proteins encoded by two strains of TYLCSV (i.e.,
10 “Sardinia” and “Sicily”) that exhibit different iterons. As can be observed in Figure 6b, two of the
11 residues located on the α 1-strand (SIKA in TYLCSV-Sar, and QINA in TYLCSV-Sic), and one
12 residue on the α 5-strand (N69 in both cases) point their side chains towards the exposed
13 surface of the α -sheet, hence providing a different hydrogen bonding pattern for interactions
14 with the major groove of DNA. The potential combinations of three, four or even more variable
15 amino acid residues in strands α 1 and α 5 (or their structural equivalents) may easily explain
16 the great diversity of iterons found among the known RCR viral systems.

17 18 **Discussion**

19 By using a comparative approach based on several heuristic hypotheses, we have identified in
20 the RCR initiators from four groups of ssDNA viral systems the amino acid residues that
21 probably determine their high-affinity DNA-binding specificity. These predicted SPDs cluster in
22 two discrete protein segments closely associated to distinct conserved amino acid motifs. The
23 group of SPDs adjacent to the RCR motif 1 was previously identified in geminivirus Rep
24 proteins [2], but the existence of a comparable domain in nanovirus and circovirus replication
25 initiators was doubtful given that the real presence of the first two RCR motifs in those proteins
26 was debatable [12]. In this new, more comprehensive comparative analysis of viral RCR
27 initiators, it was demonstrated that motifs 1 and 2 from geminivirus Rep are truly homologous to
28 conserved motifs of replication proteins encoded by other eukaryotic ssDNA viral systems.
29 Furthermore, two previously unnoticed SPDs associated to motif 2 were identified in geminivirus
30 Rep proteins, both of which have evolutionary counterparts in the replication initiators of
31 circoviruses, nanoviruses, and alphasatellites. Our results are in close agreement with the scant
32 experimental data currently published on replication specificity determinants of those systems.
33 In particular, we point out the following data. 1) The *trans*-acting replication factors of the non-
34 pathogenic PCV type 1 (PCV-1) and the pathogenic PCV-2, are functionally exchangeable [39],
35 a fact that is in accordance with the identity of their Rep SPD-r1 and SPD-r2 segments (Fig.3b).
36 2) The master Rep proteins from FBYNV, MVDV and SCSV, all of which display identical SPD-
37 r1 and SPD-r2 elements (Fig.2b), are able to support replication of heterologous nanovirus
38 DNAs harboring similar iterons [60]. 3) The only *trans*-acting replication SPD of a geminivirus
39 that has been experimentally identified, namely, the residue 10 of *Tomato leaf curl New Delhi*
40 *virus* Rep [7], corresponds to a predicted SPD identified by the CAGHIP method [2]. 4) The

1 SPD-r1 and SPD-r2 segments of circovirus and geminivirus Rep proteins are a part of the small
2 two-stranded β -sheet extension identified as the structural element mediating dsDNA binding of
3 PCV-2 and TYLCSV Rep proteins [6, 63].
4

5 **A heuristic “code of SPDs” for Rep DNA cognate elements**

6 Despite the limited experimental evidence currently available, several lines of indirect evidence
7 support the hypothesis that the predicted SPDs are, actually, aa residues controlling the Rep
8 affinity for specific DNA sequences. That evidence is particularly sound in the case of the
9 alphasatellite RCR initiators. Indeed, the potential SPDs of the ~90 analyzed Rep proteins of
10 this subviral group were consistently identified by our approach into four invariant positions: 5,
11 7, 59, and 61 (or 53 and 55 in three proteins; see Fig.1). The aa residues in those Rep positions
12 are conserved in all members of a given IsoPG, while the specific combination of the four SPDs
13 is exclusive of each particular iso-specific group. These facts suggest the existence of a kind of
14 “code of SPDs” determining the Rep DNA-binding preferences. A representation of that
15 hypothetical code for the 10 distinct iterons of alphasatellites is shown in Figure 1c. For
16 simplicity, the SPD code is depicted as two sets of three letters (separated by a period)
17 corresponding to Rep aa residues 5 to 7, and 59 to 61, respectively. The heuristic usefulness of
18 that hypothetical code of SPDs is well illustrated in the case of the *Nanoviridae* RCR initiators.
19

20 Importantly, all nanovirus Rep proteins with identical cognate iterons display similar aa residues
21 at positions homologous to those of the four alphasatellite Rep SPDs, as shown in Fig. 2. In
22 contrast, all proteins differing in their cognate DNA sequences also differ in one or more of
23 those four Rep residues, independently of their phylogenetic distance. The case of the master-
24 Rep proteins of three nanoviruses (i.e., FBYNV, MVDV and SCSV) and two babuviruses (i.e.,
25 BBTV and ABTV) is exemplar of related proteins with different cognate iterons. These two
26 groups exhibit different aa residues 4 and 6, homologous to alphasatellite Rep SPDs at
27 positions 5 and 7, and accordingly recognize distinct DNA sequences (see Fig.2, Clade C).
28 Additional remarkable examples are the following: 1) The Rep proteins of the BBTV C1.2 and
29 BBTV C1.4 replicons display very high aa sequence identity (i.e., 94%) but recognize iterons
30 differing in four nucleotides, which leads to a completely different combination of predicted
31 SPDs, namely, [PsL, RiR] and [SsF, SiK], respectively (see Fig.2, Clade B, B1 and B3); 2) Rep
32 proteins of the BBTV C1.4 and BBTV C2.1b replicons, which exhibit divergent aa sequences
33 (i.e., 62% identity) but belong to the same IsoPG, display identical aa residues in the equivalent
34 positions, i.e., [SsF, SiK] (see Fig.2, Clade B, B3 and B4). In the case of circovirus Rep
35 proteins, the analysis suggests that aa residues homologous to alphasatellite Rep SPDs are
36 also important for their DNA-binding specificity. For instance, the 16 members from the
37 “GGAGCCAC” IsoPG, which are classified into four circovirus species, encode Rep proteins
38 exhibiting an identical pattern of putative SPDs (i.e., [AaK, KxR]) in spite of their considerable
39 aa sequence divergence (see Fig.3). On the contrary, circoviruses closely related to members
40 of the former IsoPG, like GuCV, and FiCV, which exhibit distinct iterons, display a different

1 pattern of putative SPDs, i.e., [GaR, KqR] and [PcK, KqR], respectively. A comparable case is
2 that of the circoviruses infecting geese and swans, that present distinct iterons and encodes
3 Rep proteins displaying several different aa residues within the amino acid stretch preceding the
4 conserved motif c1, including the homologous residue to alphasatellite Rep SPD in position 5,
5 that is S8 in GoCV Rep, and G8 in SwCV Rep (see Fig. 3b). It is important to notice in this case
6 that one aa residue that is not homologous to an alphasatellite Rep SPD also could be a
7 specificity determinant, namely, N6 of GoCV and DuCV, and D6 of SwCV. This observation
8 point out a limitation of the simplified representation of the Rep SPD code in alphasatellites, that
9 is restricted to only four positions. In geminivirus Rep proteins three or four potential SPDs have
10 been identified in the protein region adjacent to motif 1 (see Fig.4b for a specific example),
11 hence suggesting the existence of more complex SPD codes in certain families of RCR
12 initiators [2, 49]. Amongst the circovirus Reps, besides the instance of the GoCV and SwCV
13 proteins, the case of the BFDV and GuCV replication initiators suggests the existence of
14 additional Rep SPDs, because their simplified SPD code is similar (i.e., [GxR, KxR]) although
15 their Rep cognate DNA elements are distinct (Fig. 3). This apparent exception to the rule
16 observed in alphasatellite Reps could be explained if another IsoPG-specific aa residue (i.e., G6
17 of BFDV Rep and D6 of GuCV Rep) is included as putative SPD. In this case, the pattern of
18 SDPs would be different: [GsGxR, KxR] and [DsGxR, KxR], respectively.

19

20 Notably, the main geminivirus Rep SPDs are also homologous to the ones found in
21 alphasatellite Reps (Fig.5). For example, the X1 and X3 residues of geminivirus Rep IRD, that
22 have been postulated to play a central role in the control of Rep DNA-binding specificity [2], are
23 the evolutionary counterparts of the alphasatellite Rep SPDs at positions 5 and 7 (Fig. 5).
24 Interestingly, the substitution of these IRD residues in the Rep protein of *Tomato mottle virus*
25 (ToMoV) by the homologous residues of proteins encoded by three begomoviruses with distinct
26 iterons, conferred to the mutant ToMoV Rep proteins the capability to trans-replicate the
27 genomic component B of those three viruses (Bañuelos-Hernandez and Arguello-Astorga, in
28 preparation). Taken together, the reported experimental data, the results of the theoretical
29 modeling of the 3-D structure of diverse Reps, and the coherent set of DNA/protein correlations
30 observed in the examined viral systems, lead to the conclusion that the potential SPDs
31 identified in the RCR initiators of eukaryotic ssDNA viruses are, almost certainly, amino acid
32 residues determining the Rep preference for specific iterative sequences.

33

34 ***Are SPDs of RCR N-123-C proteins invariably associated to Motifs 1 and 2?***

35 The diversity of entities encoding RCR initiators N-123-C is remarkable, comprising several
36 lineages of prokaryotic and eukaryotic replicons. Examples of these lineages include phages
37 and plasmids of Bacteria, like microviruses, plectroviruses, the large plasmid families
38 represented by pMV158, pBI101, and pC194, and plasmids of cyanobacteria and phytoplasmas
39 [24, 30, 31, 32, 43, 44, 52]; viruses and extrachromosomal replicons of Eukarya like some
40 plasmids of red algae, the vertebrate-infecting circoviruses, a number of plant-infecting ssDNA

1 viruses, and a set of new circovirus-like genomes reconstructed from marine metagenomic
2 sequences [12, 13, 14, 24, 31, 42, 51]; and several Archea systems like the recently described
3 *Halorubrum* pleomorphic virus 1 [48] and the plasmids pH5B of *Halobacterium* [27], pGS5 of
4 *Archaeoglobus profundus* (FJ707368), pZMX201 of *Natrinema* sp and other plasmids of
5 Haloarchaea [65]. Despite the extreme divergence among members of the Superfamily N-123-C
6 of RCR initiation proteins, Ilyina and Koonin [24] proposed that all of them are evolutionarily
7 related because it is unlikely that a similar arrangement of the three conserved RCR motifs
8 could have evolved independently in several lineages. This notion is supported by the fact that
9 motifs 1, 2 and 3 are not universally required, and are absent in RCR initiators of several
10 plasmids and viral systems, like pT181 and phage M13 [28, 62]. From the data assembled in
11 this study, and considering the common ancestry of the aforementioned Rep proteins, a natural
12 conclusion is that SPDs of all N-123-C RCR initiators could be located in analogous positions.
13 This prediction is consistent with a recent 3D crystal structure reported for the Rep protein
14 (RepB) of plasmid pMV158 of *Streptococcus agalactiae* [3]. In this last study it was found that
15 the aa residues of RepB that are apparently involved in specific dsDNA binding are K3, K5 and
16 K7, adjacent to motif 1 (FLLYP, residues 11-15), and R72, K73 and K74, located ahead of motif
17 2 (HYHVLY, residues 55-60) [3]. These data are in remarkable agreement with our predictions,
18 based on the concept of a close association between the clusters of determinants of DNA
19 recognition and the conserved motifs 1 and 2 of Rep proteins.

1 **Acknowledgements**

2

3 We thank to Drs. Roberto Ruiz-Medrano (CINVESTAV, IPN), Trinidad Ascencio-Ibañez (North
4 Carolina State University) and Braulio Gutiérrez-Medina (IPICYT) for critical reading of the
5 manuscript and many helpful suggestions.

6 A.L. was supported by a fellowship from the Instituto Potosino de Investigación Científica y
7 Tecnológica, A.C., and a PhD fellowship (211758) from CONACYT, Mexico. This research was
8 supported by the Consejo Nacional de Ciencia y Tecnología, Mexico (grant no. 42639-Q to
9 G.R.A.-A. and grant no. 49039 to L.R.-R).

1 **References**

2

3 1.Arguello-Astorga GR, Guevara-Gonzalez RG, Herrera-Estrella LR, Rivera-Bustamante RF
4 (1994) Geminivirus replication origins have a group-specific organization of iterative elements: a
5 model for replication. *Virology* 203:90-100

6

7 2.Arguello-Astorga GR, Ruiz-Medrano R (2001) An iteron-related domain is associated to Motif
8 1 in the replication proteins of geminiviruses: identification of potential interacting amino acid-
9 base pairs by a comparative approach. *Arch Virol* 146:465-485

10

11 3.Boer DR, Ruíz-Masó JA, López-Blanco JR, Blanco AG, Vives-Llàcer M, Chacón P, Usón I,
12 Gomis-Rüth FX, Espinosa M, Llorca O, del Solar G, Coll M (2009) Plasmid replication initiator
13 RepB forms a hexamer reminiscent of ring helicases and has mobile nuclease domains. *EMBO*
14 *J* 28:1666-1678

15

16 4.Briddon RW, Bull SE, Amin I, Mansoor S, Bedford ID, Rishi N, Siwach SS, Zafar Y, Abdel-
17 Salam AM, Markham PG (2004) Diversity of DNA 1: a satellite-like molecule associated with
18 monopartite begomovirus-DNA beta complexes. *Virology* 324:462-474

19

20 5.Briddon RW, Stanley J (2006) Subviral agents associated with plant single-stranded DNA
21 viruses. *Virology* 344:198-210

22

23 6.Campos-Olivas R, Louis JM, Clerot D, Gronenborn B, Gronenborn AM (2002) The structure of
24 a replication initiator unites diverse aspects of nucleic acid metabolism. *Proc Natl Acad Sci USA*
25 99:10310-10315.

26

27 7.Chatterji A, Padidam M, Beachy RN, Fauquet CM (1999) Identification of replication specificity
28 determinants in two strains of tomato leaf curl virus from New Delhi. *J Virol.* 73: 5481–5489

29

30 8.Davis IW, Leaver-Fay A, Chen VB, Block JN, Kapral GJ, Wang X, Murray LW, Arendall (2007)
31 MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic*
32 *Acids Res (Web Server issue):*W375-83

33

34 9.Eisenberg S, Griffith J, Kornberg A (1977) ϕ X174 **cistron A** protein is a multifunctional
35 enzyme in DNA replication. *Proc Natl Acad Sci USA* 74:3198–3202

36

37 10.Fauquet CM, Briddon RW, Brown JK, Moriones E, Stanley J, Zerbini M, Zhou X (2008)
38 Geminivirus strain demarcation and nomenclature. *Arch Virol* 153:783-821.

39

- 1 11.Faurez F, Dory D, Grasland B, Jestin A (2009) Replication of porcine circoviruses. *Virology* 416:60
- 2
- 3
- 4 12.Gibbs MJ, Smeianov VV, Steele JL, Upcroft P, Efimov BA (2006) Two families of rep-like
- 5 genes that probably originated by interspecies recombination are represented in viral, plasmid,
- 6 bacterial, and parasitic protozoan genomes. *Mol Biol Evol* 23:1097-1100
- 7
- 8 13.Gibbs MJ, Weiller GF (1999) Evidence that a plant virus switched hosts to infect a vertebrate
- 9 and then recombined with a vertebrate-infecting virus. *Proc Natl Acad Sci USA* 96:8022-8027
- 10
- 11 14.Gronenborn B (2004) Nanoviruses: genome organisation and protein function. *Vet Microbiol*
- 12 98:103-109
- 13
- 14 15.Gutierrez C (1999) Geminivirus DNA replication. *Cell Mol Life Sci* 56(3-4):313-329
- 15
- 16 16.Halami MY, Nieper H, Müller H, Johne R (2008) Detection of a novel circovirus in mute
- 17 swans (*Cygnus olor*) by using nested broad-spectrum PCR. *Virus Res* 132:208-212
- 18
- 19 17.Hanley-Bowdoin L, Settlage SB, Orozco BM, Nagar S, Robertson D (1999) Geminiviruses:
- 20 Models for plant DNA replication, transcription, and cell cycle regulation. *Crit Rev Plant Sci*
- 21 18:71-106
- 22
- 23 18.Hattermann K, Schmitt C, Soike D, Mankertz A (2003) Cloning and sequencing of Duck
- 24 circovirus (DuCV). *Arch Virol* 148:2471-2480
- 25
- 26 19.Heath L, Martin DP, Warburton L, Perrin M, Horsfield W, Kingsley C, Rybicki EP, Williamson
- 27 AL (2004) Evidence of unique genotypes of beak and feather disease virus in southern Africa. *J*
- 28 *Virology* 78:9277-9284
- 29
- 30 20.Herrera-Valencia VA, Dugdale B, Harding RM, Dale JL (2006) An iterated sequence in the
- 31 genome of Banana bunchy top virus is essential for efficient replication *J Gen Virol* 87:3409-
- 32 3412
- 33
- 34 21.Heyraud-Nitschke F, Schumacher S, Laufs J, Schaefer S, Schell J, Gronenborn B (1995)
- 35 Determination of the origin cleavage and joining domain of geminivirus Rep proteins. *Nucleic*
- 36 *Acids Res* 23:910-916
- 37
- 38 22.Horser CL, Karan M, Harding RM, Dale JL (2001) Additional rep-encoding DNAs associated
- 39 with banana bunchy top virus. *Arch Virol* 146:71-86
- 40

- 1 23.Hughes AL (2004) Birth-and-death evolution of protein-coding regions and concerted
2 evolution of non-coding regions in the multi-component genomes of nanoviruses. *Mol*
3 *Phylogenet Evol* 30:287-294
4
- 5 24.Ilyina TV, Koonin EV (1992) Conserved sequence motifs in the initiator proteins for rolling
6 circle DNA replication encoded by diverse replicons from eubacteria, eucaryotes and
7 archaeobacteria. *Nucleic Acids Res* 20:3279-3285
8
- 9 25.Jeske H, Lütgemeier M, Preiss W (2001) DNA forms indicate rolling circle and
10 recombination-dependent replication of Abutilon mosaic virus. *EMBO J* 20:6158-6167
11
- 12 26.Johne R, Fernandez-de-Luco D, Hofle U, Muller H (2006) Genome of a novel circovirus of
13 starlings, amplified by multiply primed rolling-circle amplification. *J Gen Virol* 87:1189-1195
14
- 15 27.Kagramanova VK, Derckacheva NI, Mankin AS (1988) The complete nucleotide sequence of
16 the archaeobacterial plasmid pHSB from Halobacterium, strain SB3. *Nucleic Acids Res* 16:4158
17
- 18 28.Khan SA (1997) Rolling-circle replication of bacterial plasmids. *Microbiol Mol Biol Rev*
19 61:442-455
20
- 21 29.Khan SA (2005) Plasmid rolling-circle replication: highlights of two decades of research.
22 *Plasmid* 53:126-136
23
- 24 30.Koonin EV, Ilyina TV (1992) Geminivirus replication proteins are related to prokaryotic
25 plasmid rolling circle DNA replication initiator proteins. *J Gen Virol* 73:2763-2766
26
- 27 31.Koonin EV, Ilyina TV (1993) Computer-assisted dissection of rolling circle DNA replication.
28 *Biosystems* 30:241-268
29
- 30 32.Krupovic M, Ravantti JJ, Bamford DH (2009) Geminiviruses: a tale of a plasmid becoming a
31 virus. *BMC Evol Biol* 9:112
32
- 33 33.Laskowski RA, MacArthur MW, Moss DS, Thornton JM (1993) PROCHECK: a program to
34 check the stereochemical quality of protein structures. *J Appl Cryst* 26:283-291
35
- 36 34.Laufs J, Traut W, Heyraud F, Matzeit V, Rogers SG, Schell J, Gronenborn B (1995) In vitro
37 cleavage and joining at the viral origin of replication by the replication initiator protein of tomato
38 yellow leaf curl virus. *Proc Natl Acad Sci USA* 92:3879-3883
39

- 1 35.Lin WL, Chien MS, Du YW, Wu PC, Huang C (2009) The N-terminus of porcine circovirus
2 type 2 replication protein is required for nuclear localization and ori binding activities. *Biochem*
3 *Biophys Res Commun* 379:1066-1071
4
- 5 36.Lobley A, Sadowski MI, Jones DT (2009) pGenTHREADER and pDomTHREADER: New
6 methods for improved protein fold recognition and superfamily discrimination. *Bioinformatics*
7 25:1761-1767
8
- 9 37.Mankertz A, Caliskan R, Hattermann K, Hillenbrand B, Kurzendoerfer P, Mueller B, Schmitt
10 C, Steinfeldt T, Finsterbusch T (2004) Molecular biology of Porcine circovirus: analyses of gene
11 expression and viral replication. *Vet Microbiol* 98:81-88
12
- 13 38.Mankertz A, Hattermann K, Ehlers B, Soike D (2001) Cloning and sequencing of columbid
14 circovirus (CoCV), a new circovirus from pigeons. *Arch Virol* 145:2469-2479
15
- 16 39.Mankertz A, Mueller B, Steinfeldt T, Schmitt C, Finsterbusch T (2003) New reporter gene-
17 based replication assay reveals exchangeability of replication factors of porcine circovirus types
18 1 and 2. *J Virol* 77:9885-9893
19
- 20 40.Moscoso M, del Solar G, Espinosa M (1995) Specific nicking-closing activity of the initiator
21 of replication protein RepB of plasmid pMV158 on supercoiled or single-stranded DNA. *J Biol*
22 *Chem* 270:3772-3779
23
- 24 41.Nawaz-ul-Rehman MS, Fauquet CM (2009) Evolution of geminiviruses and their satellites.
25 *FEBS Lett* 583:1825-1832
26
- 27 42.Niagro FD, Forsthoefel AN, Lawther RP, Kamalanathan L, Ritchie BW, Latimer KS, Lukert
28 PD (1998) Beak and feather disease virus and porcine circovirus genomes: intermediates
29 between the geminiviruses and plant circoviruses. *Arch Virol* 143:1723-1744
30
- 31 43.Nishigawa H, Miyata S, Oshima K, Sawayanagi T, Komoto A, Kuboyama T, Matsuda I,
32 Tsuchizaki T, Namba S (2001) In planta expression of a protein encoded by the
33 extrachromosomal DNA of a phytoplasma and related to geminivirus replication proteins.
34 *Microbiology* 147:507-513
35
- 36 44.Oshima K, Kakizawa S, Nishigawa H, Kuboyama T, Miyata S, Ugaki M, Namba S (2001) A
37 plasmid of phytoplasma encodes a unique replication protein having both plasmid- and virus-
38 like domains: clue to viral ancestry or result of virus/plasmid recombination? *Virology* 285:270-
39 277
40

- 1 45.Palmer KE, Rybicki EP (1998) The molecular biology of mastreviruses. *Adv Virus Res*
2 50:183–234
3
- 4 46.Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE
5 (2004) UCSF Chimera - A Visualization System for Exploratory Research and Analysis. *J*
6 *Comput Chem* 25:1605-1612
7
- 8 47.Phenix KV, Weston JH, Ypelaar I, Lavazza A, Smyth JA, Todd D, Wilcox GE, Raidal SR
9 (2001) Nucleotide sequence analysis of a novel circovirus of canaries and its relationship to
10 other members of the genus *Circovirus* of the family *Circoviridae*. *J Gen Virol* 82:2805-2809
11
- 12 48.Pietila MK, Roine E, Paulin L, Kalkkinen N, Bamford DH (2009) An ssDNA virus infecting
13 archaea: a new lineage of viruses with a membrane envelope *Mol Microbiol* 72:307-319
14
- 15 49.Ramos PL, Guevara-González RG, Peral R, Ascencio-Ibañez JT, Polston JE, Argüello-
16 Astorga GR, Vega-Arreguín JC, Rivera-Bustamante RF (2003) Tomato mottle Taino virus
17 pseudorecombines with PYMV but not with ToMoV: implications for the delimitation of cis- and
18 trans-acting replication specificity determinants. *Arch Virol* 148:1697-1712
19
- 20 50.Rojas MR, Hagen C, Lucas WJ, Gilbertson RL (2005) Exploiting chinks in the plant's armor:
21 evolution and emergence of geminiviruses. *Annu Rev Phytopathol* 43:361-394
22
- 23 51.Rosario K, Duffy S, Breitbart M (2009) Diverse circovirus-like genome architectures revealed
24 by environmental metagenomics. *J Gen Virol.* 90: 2418-2424
25
- 26 52.Ruiz-Masó JA, Lurz R, Espinosa M, del Solar G (2007) Interactions between the RepB
27 initiator protein of plasmid pMV158 and two distant DNA regions within the origin of replication.
28 *Nucleic Acids Res* 35:1230-1244
29
- 30 53.Schwede T, Kopp J, Guex N, Peitsch MC (2003) SWISS-MODEL: An automated protein
31 homology-modeling server. *Nucleic Acids Res* 31:3381-3385
32
- 33 54.Sharman M, Thomas JE, Skabo S, Holton TA (2008) Abacá bunchy top virus, a new
34 member of the genus *Babuvirus* (family *Nanoviridae*). *Arch Virol* 153:135-147
35
- 36 55.Singh DK, Malik PS, Choudhury NR, Mukherjee SK (2008) MYMIV replication initiator
37 protein (Rep): roles at the initiation and elongation steps of MYMIV DNA replication. *Virology*
38 380:75-83
39

- 1 56.Steinfeldt T, Finsterbusch T, Mankertz A (2001) Rep and Rep' protein of porcine circovirus
2 type 1 bind to the origin of replication in vitro. *Virology* 291:152-160
3
- 4 57.Stewart ME, Perry R, Raidal SR (2006) Identification of a novel circovirus in Australian
5 ravens (*Corvus coronoides*) with feather disease. *Avian Pathol* 35:86-92
6
- 7 58.Tamura K, Dudley J, Nei M, Kumar S (2007) *MEGA4*: Molecular Evolutionary Genetics
8 Analysis (MEGA) software version 4.0. *Mol Biol Evol* 24:1596-1599
9
- 10 59.Timchenko T, de Kouchkovsky F, Katul L, David C, Vetten HJ, Gronenborn B (1999) A
11 single rep protein initiates replication of multiple genome components of faba bean necrotic
12 yellows virus, a single-stranded DNA virus of plants. *J Virol* 73:10173-10182
13
- 14 60.Timchenko T, Katul L, Sano Y, de Kouchkovsky F, Vetten HJ, Gronenborn B. (2000) The
15 master rep concept in nanovirus replication: identification of missing genome components and
16 potential for natural genetic reassortment. *Virology* 274:189-195
17
- 18 61.Todd D, Scott AN, Fringuelli E, Shivraprasad HL, Gavier-Widen D, Smyth JA (2007)
19 Molecular characterization of novel circoviruses from finch and gull. *Avian Pathol* 36:75-81
20
- 21 62. van Wezenbeek PM, HulsebosTJ, Schoenmakers JG (1980) Nucleotide sequence of the
22 filamentous bacteriophage M13 DNA genome: comparison with phage fd. *Gene* 11:129-148
23
- 24 63.Vega-Rocha S, Byeon IJ, Gronenborn B, Gronenborn AM, Campos-Olivas R. (2007a)
25 Solution structure, divalent metal and DNA binding of the endonuclease domain from the
26 replication initiation protein from porcine circovirus 2. *J Mol Biol* 367:473-487
27
- 28 64.Vega-Rocha S, Gronenborn B, Gronenborn AM, Campos-Olivas R (2007b) Solution
29 structure of the endonuclease domain from the master replication initiator protein of the
30 nanovirus faba bean necrotic yellows virus and comparison with the corresponding geminivirus
31 and circovirus structures. *Biochemistry* 46:6201-6212
32
- 33 65.Zhou L, Zhou M, SunC, HanJ, Lu Q, Zhou J, Xiang H (2008) Precise determination, cross-
34 recognition, and functional analysis of the double-strand origins of the rolling-circle replication
35 plasmids in haloarchaea. *J Bacteriol* 190:5710-5719
36

1 **Figure Legends**

2

3 **Fig. 1. Iterons and SPDs of alphasatellites.** A) Organization of Ori-associated iterative
4 sequences. The arrangement of iterons exhibited by SiLCV-DNA1 is representative of six
5 alphasatellite IsoPG (i.e., GAGACCC, GGMACCC, GGWTCCC, CGACCCT, CCTCGGN, and
6 ACCTCT groups). Filled arrows show the orientation of the iterons with respect to the stem-loop
7 element (SLE); numbers denote the nucleotides spanned between each drawn element (the
8 SLE, the putative TATA box, and the start codon of *rep* gene). Lower case letters in an iterated
9 element indicate a nucleotide that does not match with the iteron consensus. B) Summary of
10 potential DNA-binding SPDs of alphasatellite RCR initiators. Amino acid residues identified as
11 putative SPDs are shadowed. These residues cluster in two discrete regions that are labeled as
12 SPD-region 1 (SPD-r1) and SPD-region 2 (SPD-r2). Representative aa sequences of a few
13 members of each IsoPG are showed to illustrate natural variations in residues flanking the
14 putative SPDs. The conserved motifs $\alpha 1$ and $\alpha 2$ are indicated at the top of the alignments.
15 Numbers in front of the iteron sequence indicate the number of members of that particular
16 IsoPG. Numbers at the end of each partial Rep sequence indicate the alphasatellite to which
17 that specific sequence correspond, as follows: (1) AM236764; (2) NC_007640; 3) AJ512959; (4)
18 NC_010620; (5) AJ888451; (6) AJ512956; (7) FJ218493; (8) EU384644; (9) AJ888453; (10)
19 AJ888448; (11) NC_009563; (12) NC_009564; (13) NC_012789; (14) FJ218494; (15)
20 FM164740; (16) FM164739; (17) NC_003414. C) Simplified representation of the four SPDs of
21 proteins recognizing a specific iteron, that constitute the heuristic “code of SPDs” of
22 alphasatellites.

23

24 **Figure 2. Putative SPDs in the DNA-binding domain of nanovirus Rep proteins.** A)
25 Neighbor-joining tree showing phylogenetic relationships between Rep proteins encoded by
26 essential and satellite-like genomic components of nanoviruses. TYLCV is the outgroup. The
27 tree was constructed using MEGA 4 software, based on the Poisson-corrected distance
28 estimates. The optimal tree with the sum of branch length = 5.41661483 is shown. The number
29 at each node indicates the bootstrap score over 1000 replicates for that node. The bootstrap
30 values less than 50% are not shown. All positions containing gaps and missing data were
31 eliminated from the dataset (Complete deletion option). There were a total of 257 positions in
32 the final dataset. The scale at the bottom is in units of amino acid substitutions per site. B) The
33 N-terminal domain of Rep proteins of nanovirus components are grouped in three major clades,
34 as shown in panel A. These lineages are roughly equivalent to the four nanovirus clades
35 defined by Hughes, 2004 [23]. The protein regions where the putative SPDs were identified by
36 our theoretical approach are indicated by a light-coloured box; brackets indicate viral genomes
37 having the same iterated sequence. Segments n1 and n2 are conserved motifs identified in
38 nanoviral Reps; numbers between dashes indicate the length of the omitted protein region.
39 Amino acid residues homologous to the alphasatellite Rep SPDs are shadowed. GenBank
40 accession numbers of the nanovirus genomic components are as follows: [A1] AJ005964; [A2]

1 NC_003647; [A3] NC_003638; [A4] AB000922; [A5] U16735; [A6] AJ005966; [B1] L32166; [B2]
2 L32167; [B3] FJ389724; [B4] AF216222; [B5] AF416471; [B6] NC_003558; [B7] NC_003639;
3 [B8] U16731; [C1] NC_003479; [C2] NC_010319; [C3] NC_003560; [C4] NC_003648; [C5]
4 NC_003812.

5

6 **Fig. 3. Potential SPDs in Rep proteins of Circoviruses** A) Neighbour-joining phylogeny of the
7 *Circovirus* genus members based on the amino acid sequence of the Rep endonuclease
8 domain. Branches are proportional to the number of changes by each 100 positions. A colour-
9 coded vignette illustrating the iterons arrangement characteristic for each clade is depicted on
10 their respective branches. Sequences of the corresponding iterons are shown colour coded and
11 boxed at the right of the figure. B) Identification of the “convergent” protein region that
12 presumably contain residues functioning as DNA binding SPDs. Partial sequences of Rep
13 proteins from the seven IsoPG defined in panel A are shown. Two boxes indicate the locations
14 of aa residues that probably determine the specificity of Rep. Differences between the aligned
15 sequences are marked with asterisks. Regions c1 and c2 are conserved motifs identified in
16 circovirus Reps. Amino acid residues homologous to the alphasatellite Rep SPDs are
17 shadowed. For clarity, only the complete N-terminal sequence of proteins belonging to the
18 GGAGCACC IsoPG is showed; for the other IsoPG most of the amino acid residues were
19 omitted. GenBank accession numbers of the circovirus genomes are as follows: [1] AJ298229;
20 [2] DQ146997; [3] NC_003410; [4] DQ172906; [5] NC_008522; [6] NC_008521; [7] AF311299;
21 [8] AF311296; [9] DQ166838; [10] AF536935; [11] EU056310; [12] AY184287; [13]
22 AY321983.

23

24 **Fig. 4. Geminivirus Rep proteins have a second SPD region close to Motif 2** In each chart
25 the amino acid sequences of the N-terminal domain (1-75) of two begomovirus Rep proteins
26 from distinct IsoPG are aligned. The differential residues between each pair are marked with an
27 asterisk (*), and SPDs are highlighted. Amino acid residues homologous to the alphasatellite
28 Rep SPDs are shadowed. Filled arrows indicate the position of the predicted beta-sheets one
29 and five. Boxes indicate the conserved motifs 1 and 2, respectively. The underlined region
30 corresponds to the core sequence of the Iteron Related Domain (IRD) described by Arguello-
31 Astorga and Ruiz Medrano, 2001 [2]. a) An example in which SPDs are identified in the IRD
32 region, but differences in the predicted beta-strand 5 are not observed. b) A case in that one
33 potential SPD is identified in the beta-strand 5 element, in addition to SPDs in the IRD region. c)
34 An example in that two putative SPDs are identified in the beta-strand 5 region. GenBank
35 accession numbers of the viral sequences are as follows: (1) AB330079, (2) AF448058, (3)
36 NC_008492, (4) AY965900, (5) NC_009548, (6) NC_003357.

37

38 **Fig. 5. Conservation of SPD-regions in RCR initiators of eukaryotic circular ssDNA**
39 **viruses.** Structure-based alignment of the Rep endonuclease domain sequences from selected
40 circoviruses, nanoviruses (including nanovirus-like satellites) and geminiviruses. The member of

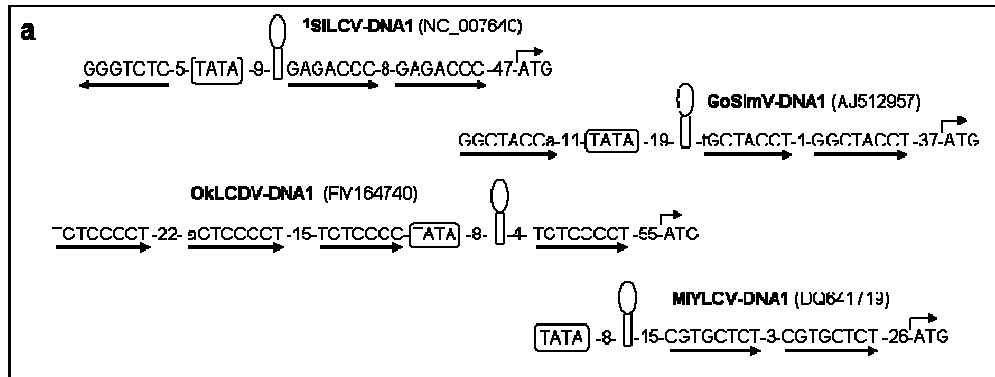
1 each viral lineage with a previously reported 3D structure of the Rep N-terminal domain is
2 shown at the bottom of its corresponding alignment. The beta strands and alpha helices that
3 compose the secondary structure, represented by rectangles and ellipses, respectively, are
4 depicted below each group. Shadowed in red are the residues or sections that were identified
5 as DNA-binding specificity determinants. Residues shadowed in blue indicate the conserved
6 motifs characteristics of each lineage, homologous to RCR motifs 1, 2 and 3 described by Ilyina
7 and Koonin, 1992 [30]. Brackets at the top of each group indicate the distance between the
8 second conserved motif of each lineage and the corresponding SPD-r2. GenBank accession
9 numbers of the virus genomes are given in Online Resource 1.

10

11 **Fig. 6. Distant SPD regions physically interact in the three-dimensional structure of RCR**

12 **initiators** A) A model of the tertiary structure for the endonuclease domain of Rep proteins of a
13 begomovirus (AYYV, EF527823), a nanovirus-like satellite (SiLCV-DNA1, NC_007640), a
14 nanovirus component (BBTV-C2.1a, AF216221) and a circovirus (FiCV, NC_008522) are
15 shown. AYYV was modeled with the initiator protein of *Tomato yellow leaf curl Sardinia virus*
16 (PDB ID=1L5I) as template, while FiCV, SiLCV-DNA1 and BBTV-C2.1a models were performed
17 with the Rep protein of *Porcine circovirus-2* (PDB ID=2HW0). In each model, the regions
18 containing the mapped SPDs and the β -strands that form the β -sheet element are indicated in
19 red. B) Enlarged view of the β 1- β 5 scaffold in models of two highly similar proteins that bind
20 different iterons. The conformer PDB ID=1L2M of *Tomato yellow leaf curl Sardinia virus*
21 (TYLCSV-Sar) is compared with a model of the Sicily strain of TYLCSV (TYLCSV-Sic) made on
22 the PDB 1L5I template. In both images the amino acids backbone equivalent to the SPD-r1 and
23 SPD-r2 regions is shown.

1 Figure 1



b

Iteron	SPD-r1		SPD-r2		[]
	SPD-r1	SPD-r2	SPD-r1	SPD-r2	
CGACCCT (4)	MPSL [KsT]	FRCFTVFF -33-	HLQGYLQLKG [ErT]	LNQVK	[1]
GAGACCC (12)	MPAL [KsQ]	FRCFTVFF -33-	HLQGYLQLKG [QrT]	LNQVK	[2]
	MAAL [KsQ]	FRCFTI FF -33-	HLQGYLQLKG [QrT]	LNQVK	[3]
	MPSI [KsQ]	FRCFTVFF -33-	HLQGYLQCKG [QrT]	LSQIK	[4]
GGWTCCC (15)	MPCV [QsQ]	FRCFTVFF -33-	HLQGYLQLKG [KrsS]	FNQVK	[5]
	MPTI [QsQ]	FRCFTVFF -33-	HLQGYLQLKG [KrsS]	LAQVK	[6]
GGCTACC (35)	MPAV [QsL]	FRCFTI FF -33-	HLQGYLQLKT [KksS]	LSAVK	[7]
	MPSA [QsV]	FRCFTI FF -33-	HLQGYLQLKT [KksS]	LSAVK	[8]
GGMACCC (19)	MPSI [Tsv]	FRCFTI FF -33-	HLQGYLQLKG [KrtT]	LNQVK	[9]
	MPSV [Tsv]	FRCFTVFF -33-	HLQGYLQLKG [KrtT]	LNQVK	[10]
	MPAL [Tsv]	FRCFTVFF -33-	HLQGYLQLKG [KrtT]	LNQVK	[11]
CGTGCTC (1)	MPSV [AsV]	FRCFTVFF -33-	HLQGYLQLKG [RrT]	LNQVK	[12]
TAGACCC (7)	MAAV [KsV]	FRCFTI FF -33-	HLQGYLQCKG [QrT]	FKQVK	[13]
	MAAI [KsV]	FRCFTI FF -33-	HLQGYLQCKG [QrT]	LKQVK	[14]
TGTCCCCT (1)	MPSI [KsV]	CWCFTLNF -27-	HLQGFQFKG [RrsS]	LLQAK	[15]
TGGCCCCCT (1)	MPSI [RsT]	CWCFTLNF -27-	HLQGFQFKG [RrsS]	LLQAK	[16]
TCCACAC (1)	MAP [QgK]	RWCFTSFD -28-	HWQGFITFVG [Qkr]	LNTVK	[17]

c

	GAGACCC	[Ks*Q, QrT]	
GGWTCCC	[QsQ, KrsS]	CGACCCT	[KsT, ErT]
		TGGCCCC	[RrsT, RrsS]
GGCTACC	[QsLV, KksS]	TAGACCC	[KsV, QrT]
		GGMACCC	[Tsv, KrtT]
TCCACAC	[QgK, Qkr]	TGTCCCCT	[KsV, RrsS]
		CGTGCTC	[AsV, RrtT]

2

Figure 4

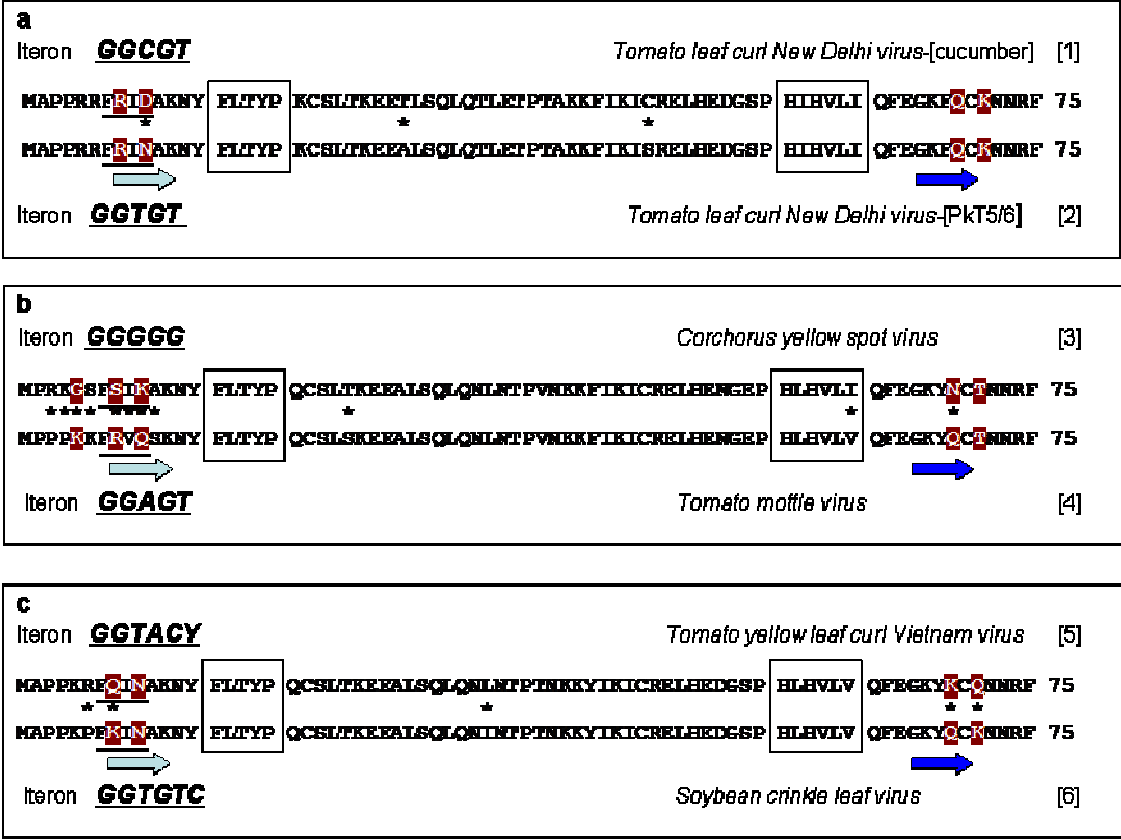


Figure 5

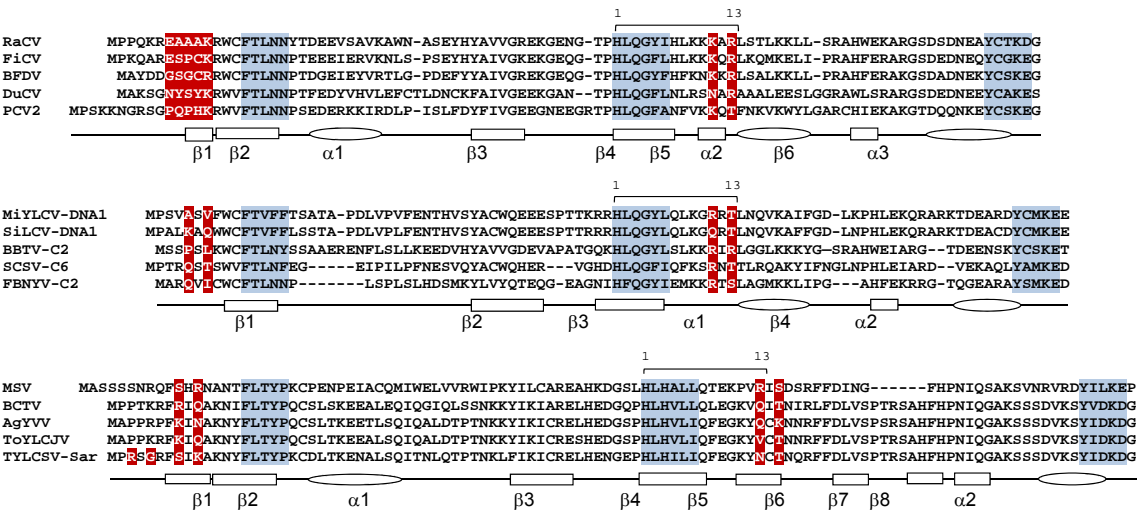
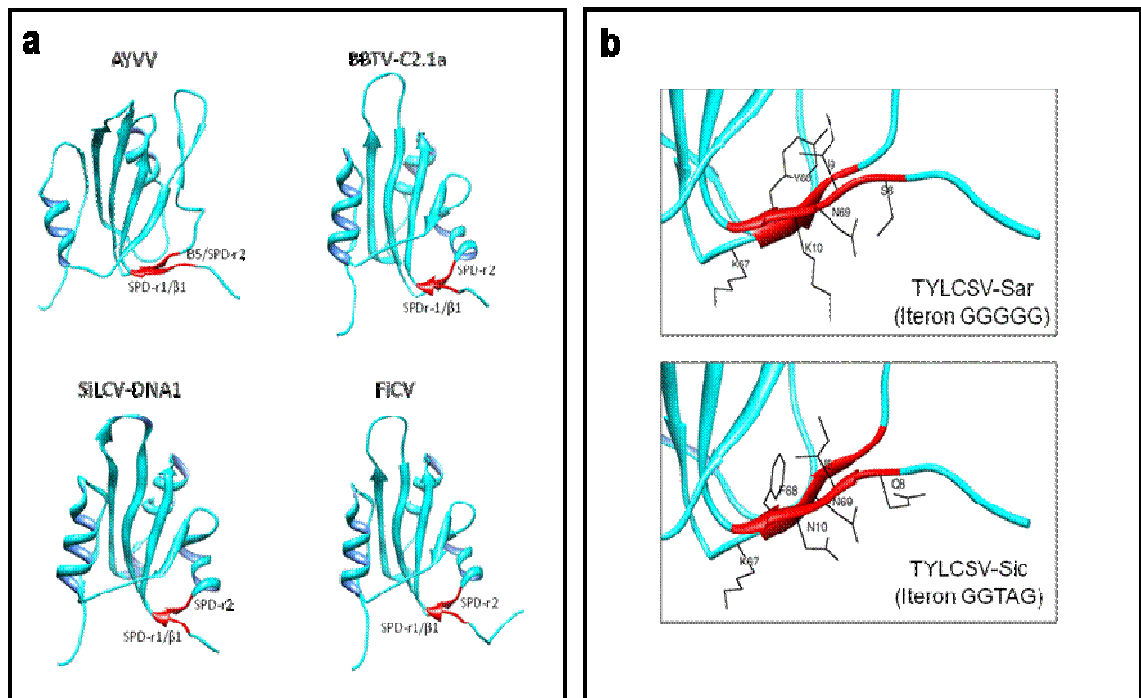


Figure 6



Supplementary Table 1. List of viruses whose Rep sequence was included in this study.

Lineage	Virus	Host	Accession #
Alphasatellites	AYVV-DNA1 ^a	<i>Ageratum sp.</i>	AJ512958
	AYVV-DNA1 ^a		AJ512950
	AYVV-DNA1 ^a		AJ512959
	AYVV-DNA1 ^a		AJ512957
	AYVV-DNA1 ^a		AJ512947
	AYVV-DNA1 ^a		AJ512948
	AYVV-DNA1 ^a		AJ512956
	AYVV-DNA1 ^a		AJ512960
	AYVV-DNA1	<i>Ageratum sp.</i>	AJ238493
	AYVV-DNA1	<i>Ageratum sp.</i>	AJ416153
	CLCuMV-DNA1	Cotton	AJ132344
	CLCuMV-DNA1	Cotton	AJ132345
	GoSimV-DNA1	Cotton	AJ512957
	MaYMV-DNA1	<i>Malvastrum sp.</i>	NC_008561
	MaYMV-DNA1		AM236764
	MaYMV-DNA1		AM236767
	MaYMV-DNA1		AM236765
	MiYLCV-DNA1	<i>Mimosa sp.</i>	DQ641719
	OkLCV-DNA1	<i>Okra sp.</i>	NC_005954
	SiLCV-DNA1	<i>Sida sp.</i>	AM050735
	SiLCV-DNA1		NC_007640
	TbCSV-DNA1	Tobacco	AJ579351
	TbCSV-DNA1		NC_005057
	TbCSV-DNA1		AJ579349
	TbCSV-DNA1		AJ579346
	TbCSV-DNA1		AJ579348
	TbCSV-DNA1		AJ579352
	TbCSV-DNA1		AJ579347
TbLCYV-DNA1	AJ888455		
TbLCYV-DNA1	NC_005060		
ToYLCCV-DNA1	Tomato		AJ579356
ToYLCCV-DNA1		AJ579347	
ToYLCCV-DNA1		AJ888446	
ToYLCCV-DNA1		AJ888451	
ToYLCCV-DNA1		AJ579358	

	ToYLCCV-DNA1		AJ579357	
	ToYLCCV-DNA1		AJ579354	
	ToYLCCV-DNA1		AJ579355	
	ToYLCCV-DNA1		AJ888449	
	ToYLCCV-DNA1		AJ888447	
	ToYLCCV-DNA1		AJ579360	
	ToYLCCV-DNA1		AJ888445	
	ToYLCCV-DNA1		AJ888448	
<i>Nanoviridae</i>	ABTV-C1	Abacá	NC_010319	
	BBTV-C1.1	Banana	NC_003479	
	BBTV-C1.1a		AF416477	
	BBTV-C1.1b		AB108458	
	BBTV-C1.2		L32166	
	BBTV-C1.3		L32167	
	BBTV-C2.1a		AF216221	
	BBTV-C2.1b		AF216222	
	BBTV-C3		AF416471	
	CFDV		Coco nut	NC_001465
	FBNYV-C1.1	<i>Vicia faba</i>	X80879	
	FBNYV-C1.2		NC_003558	
	FBNYV-C2		NC_003560	
	FBNYV-C7		AJ005964	
	FBNYV-C9		AJ005966	
	MVDV-C1.1	<i>Astragalus</i> <i>sp.</i>	NC_003638	
	MVDV-C1.2		AB027511	
	MVDV-C1.3		AB000920	
	MVDV-C2		NC_003639	
	MVDV-C3		AB000922	
	MVDV-C4		NC_003641	
	MVDV-C5		NC_003642	
	MVDV-C7		NC_003644	
	MVDV-C8		NC_003645	
	MVDV-C9		NC_003646	
	MVDV-C10		NC_003647	
	MVDV-C11	NC_003648		
	SCSV-C2	<i>Trifolium</i> <i>sp.</i>	U16731	
	SCSV-C6		U16735	
	SCSV-C8		NC_003812	
	<i>Circoviridae</i>	BFDV	<i>Agapomis</i> <i>roseicollis</i>	AF311296

	BFDVa	<i>Trichoglossus</i> <i>sp.</i>	AF311299
	CaCV	Canary	NC_003410
	CoCV	Columbids	NC_002361
	DuCV	Muscovy duck	DQ166838
	DuCVa	Mulard duck	AY228555
	FiCV	Finch	NC_008522
	GoCV	goose	AF536935
	GuCV	gull	NC_008521
	PCV1	swine	DQ472016
	PCV1a		AY184287
	PCV1b		AY699796
	PCV2		AY321983
	PCV2a		AY484410
	RaCV	Corves	DQ146997
	StCV	Starling	DQ172906
<i>Geminiviridae</i>	ToLCVNDV- [cucumber]	Tomato	AB330079
	ToLCVNDV- [Pkt5/6]	Tomato	AF448058
	CoYSV	<i>Corchorus</i> <i>sp.</i>	NC_008492
	ToMoTV	Tomato	AY965900
	TYLCVNV	Tomato	NC_009548
	SbCLV	Soybean	NC_003357
	AYVV	<i>Ageratum</i>	EF527823
	ToLCJV	<i>Ageratum</i>	AB162141
	ToLCCBV	Tomato	EU487048
	CYVMV	Croton	EU682401
	PepLCBDV	Pepper	DQ116881
TYLCSV-Sic	Tomato	DQ845787	

Abbreviations and details for nomenclature.

Nanovirus-like satellites. The name of the viral entity corresponds to the helper begomovirus plus the suffix DNA1. ^aAYVV-DNA1: These genomes were characterized using *African cassava mosaic virus* as the helper begomovirus but were isolates from different species of plants with symptoms similar to ageratum yellow vein disease. AYVV-DNA1 *Ageratum yellow vein virus-associated DNA1*, CLCuMV-DNA1 *Cotton leaf curl mosaic virus-associated DNA1*, MaYMV-DNA *Malvastrum yellow mosaic virus-associated DNA1*, MiYLCV-DNA1, *Mimosa yellow leaf curl virus-associated DNA1*, OkLCV-DNA1 *Okra leaf curl virus-associated DNA1*, SiLCV-DNA1 *Sida leaf curl virus-associated DNA1*, TbCSV-DNA1 *Tobacco curly shoot virus-associated*

DNA1, TbLCYV-DNA1 *Tobacco leaf curl Yunnan virus associated DNA1*, ToYLCCV *Tomato leaf curl China virus-associated DNA1*.

Nanoviruses. It is used the abbreviation for the name of the species followed by the suffix –CX to indicate the number of the component. In some of the previous works all the nanoviral genomes encoding a Rep were named component C1. Here we put an additional number to these components to clarify that they are not variants of the same C1 rather different replicons, and the same for components C2. An additional lower case letter is used to indicate strains of the same species that are non-redundant in the N-terminal end of the Rep protein. Examples: ABTV-C1 *Abaca bunchy top virus component 1*, BBTV-C1.1 *Banana bunchy top virus component 1.1*, BBTV-C1.1a *Banana bunchy top virus component 1.1-isolate a*, BBTV-C3 *Banana bunchy top virus component 3*, CFDV *Coconut foliar decay virus*, FBNYV-C2 *Faba bean necrotic yellow virus component 2*, MVDV-C1.2 *Milk vetch disease virus component 1.2*, MVDV-C11 *Milk vetch disease virus component 11*, SCSV-C2 *Subterranean clover stunt virus component 2*.

Circoviruses. It is used the abbreviation for the name of the species. An additional lower case letter is used to indicate strains of the same species that are non-redundant in the N-terminal end of the Rep protein. BFDV *Beak and feather disease virus*, CaCV *Canary circovirus*, CoCV *Columbidae circovirus*, DuCV *Duck circovirus*, FiCV *Finch circovirus*, GoCV *Goose circovirus*, GuCV *Gull circovirus*, PCV1 *Porcine circovirus 1*, PCV2 *Porcine circovirus 2*, RaCV *Raven circovirus*, StCV *Starling circovirus*.

Geminiviruses. Names, acronyms and GenBank accession numbers according to Fauquet CM. et al., 2008. Geminivirus strain demarcation and nomenclature. Arch Virol. 153(4):783-821.

Supplementary Figure 1.

Identification of DNA-binding specificity determinants (SPDs) by the CAGHIP approach. The endonuclease domain of three pairs of alphasatellite Rep proteins belonging to different IsoGP are compared. The differential amino acid residues between the proteins (indicated with an asterisk and boxed) are further compared with their equivalents in other proteins from the same IsoPG. For clarity, the aa sequence of those additional proteins is omitted, and only residues homologous to the differential amino acids are shown. Residues that are identical between proteins that recognize similar iterons, but different between proteins with distinct cognate DNA elements, are identified as potential SPDs (denoted with a # symbol). A two points (:) character indicates a residue that was discarded as potential SPD after comparisons with all members of its own IsoPG. Acronyms: AYVV-DNA1, Ageratum yellow vein virus-associated DNA1 (two different isolates); MiYLCV-DNA 1, Mimosa yellow leaf curl virus-associated DNA1; SiLCV-DNA1, Sida leaf curl virus-associated DNA1 (two isolates); TbCSV-DNA1, Tobacco curly shot virus-associated DNA1 (two isolates); ToYLCCV-DNA1, Tomato yellow leaf curl China virus-associated DNA1 (three different isolates).

A**Iteron GGMACCC vs CGTGCTCT**

1) TbCSV-DNA1 (AJ579346)

3) MiYLCV-DNA1 (DQ641719)

2) ToYLCCV-DNA1 (AJ888449)

	#								#
1)	T								K
2) MPSV	T	SVFWCFTVFFTSATAPDLVPVFENTHVS	YACWQEEESPTTKRRHLQGYLQLKG					K	RTL NQ 65
	*							*	
3) MPSV	A	SVFWCFTVFFTSATAPDLVPVFENTHVS	YACWQEEESPTTKRRHLQGYLQLKG					R	RTL NQ 65
	AI							R	
VK	SL	FGDLKPHLEKQRARKTDEA		C	DYCMKEETRVSGPFEFGDYCPSGSHKRRQRES				120
	**			*				*	
VK	AI	FGDLKPHLEKQRARKTDEA		R	DYCMKEETRVSGPFEFGDYCPSGSHKRRQRES				120

B**Iteron GGMACCC vs CGACCC**

1) ToYLCCV-DNA1 (AJ888447)

3) SiLCV-DNA1 (AM050735)

2) TbCSV-DNA1 (AJ888453)

	#	#		:					#
1)	VTSV		V	AA					K
2) MPS	ITSV	FWCFT	I	FFT	SA	SAPDLVPLFENTHVS	YACWQEEESPTTRRRHLQGYLQLKG		K
	**	*	*	**	**				*
3) MPS	LKST	FWCFT	V	FFT	AS	SAPDLVPLFENTHVS	YACWQEEESPTTRRRHLQGYLQLKG		E
	S								
NQVK	A	IFGDLKPHLEKQRARKTDEAC	DYCMKEETRVSGPFEFGDYCPSGSHKRRQRES						120
	*								
NQVK	S	IFGDLKPHLEKQRARKTDEAC	DYCMKEETRVSGPFEFGDYCPSGSHKRRQRES						120

C**Iteron GAGACCY vs GGWTCCC**

1) SiLCV-DNA1 (AM050734)

3) AYVV-DNA1 (AJ512948)

2) AYVV-DNA1 (AJ512959)

4) ToYLCC-DNA1 (AJ579358)

	:	#	:					#	#
1)	PALKA		V					Q	T
2) M	AALKG	QWWCFT	I	FFLSATAPDLVPLFENTHVS	YACWQEEESPTTRRRHLQGYLQLKG			Q	R
	*****		*					*	T
3) M	PTIQS	QWWCFT	V	FFLSATAPDLVPLFENTHVS	YACWQEEESPTTRRRHLQGYLQLKG			K	R
4)	PCVQS		V					K	S
	N		F						
L	N	QVKA	I	FGDL	K	PHLEKQRARKTDEAC	DYCMKEETRVSGPFEFGDYCPSGSHKRRQRES		120
	*		*		*				
L	A	QVKA	L	FGDL	N	PHLEKQRARKTDEAC	DYCMKEETRVSGPFEFGDYCPSGSHKRRQRES		120
	N		I		N				

Supplementary Figure 2

Additional examples of IRD-β5 combinations in begomoviruses. In each chart the amino acid sequences of the endonuclease domain of two highly similar proteins from begomovirus species with different iterons are displayed. The differential residues are marked with an asterisk (*). Arrows indicate the position of the predicted beta-sheets 1 and 5. Boxes indicate the conserved motifs 1 and 2, respectively. The shadowed region corresponds to the Iteron Related Domain (IRD) core sequence. GenBank accession numbers of the illustrated begomoviruses are the following: AYVV- FJ495183; ToLCJVAB162141; TYLCVV- NC_009548; ToLCCBV- EU487048; CYVMV- EU682401; PepLCBDV- DQ116881.

