(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 532 891**

(51) Int. CI.:

C12Q 1/68 (2006.01)
C12Q 1/70 (2006.01)
C40B 40/06 (2006.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(96) Fecha de presentación y número de la solicitud europea: **01.06.2010 E 10783906 (0)**

(97) Fecha y número de publicación de la concesión europea: **17.12.2014 EP 2438195**

(54) Título: **Descubrimiento de virus mediante secuenciación y ensamblaje de ARNip, miARN, ARNpi derivados de virus**

(30) Prioridad:

02.06.2009 US 183377 P
15.12.2009 US 286742 P

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
01.04.2015

(73) Titular/es:

THE REGENTS OF THE UNIVERSITY OF CALIFORNIA (100.0%)
1111 Franklin Street, 12th Floor
Oakland, CA 94607, US

(72) Inventor/es:

DING, SHOU-WEI y
WU, QINGFA

(74) Agente/Representante:

CARVAJAL Y URQUIJO, Isabel

DESCRIPCIÓN

Descubrimiento de virus mediante secuenciación y ensamblaje de ARNip, miARN, ARNpi derivados de virus

Campo técnico

En una realización, la divulgación proporciona un método para descubrir un genoma microbiano. En otra realización, la invención proporciona un método para identificar un virus. Los métodos usan ARN que interaccionan con PIWI (ARNpi), incluyendo ARNpi aislados o secuenciados a partir de organismos invertebrados tales como insectos *(Anthropoda),* nematodos *(Nemapoda), Mollusca, Porifera* y otros invertebrados, y/o plantas, hongos o algas, cianobacterias y similares.

Antecedentes

El descubrimiento de nuevos virus a menudo se ve impedido por dificultades en su amplificación en cultivo celular y/o la carencia de su reactividad cruzada en ensayos de hibridación serológicos y de ácido nucleico con virus conocidos. Recientemente se han identificado muchos virus nuevos en muestras del entorno y clínicas usando enfoques metagenómicos, en los que en primer lugar se purifican partículas virales y entonces se amplifican al azar secuencias de ácido nucleico viral antes de subclonar y secuenciar (Delwart, 2007).

La familia Dicer de receptores inmunitarios de huésped media en la inmunidad antiviral en hongos, plantas y animales invertebrados mediante interferencia de ARN (iARN) o silenciamiento de ARN (1-3). En esta inmunidad, un ARN bicatenario (ARNbc) viral se reconoce por Dicer y se corta en trozos para dar ARN de interferencia pequeños (ARNip). Estos ARNip derivados de virus se cargan entonces en un complejo de silenciamiento de ARN para que actúen como determinantes de especificidad y para guiar el corte de los ARN virales diana mediante una proteína Argonauta (AGO) presente en el complejo. Las proteínas Dicer contienen normalmente un dominio ARN helicasa, un dominio PAZ compartido con AGO y dos dominios endorribonucleasa tipo III (ARNasa III) en tándem. Dicer escinde ARNbc con preferencia simple hacia un extremo terminal de ARNbc, produciendo fragmentos de ARN pequeño dobles de tamaños diferenciados progresivamente desde el extremo terminal (4).

Además de ARNip, microARN (miARN) y ARN que interaccionan con PIWI (ARNpi) también guían el silenciamiento de ARN en complejos similares pero con AGO distintas (4-6). En *Drosophila melanogaster*, miARN y ARNip son predominantemente de 22 y 21 nucleótidos de longitud, dependen de Dicer-1 (DCR1) y DCR2 para su biogénesis, y actúan en complejos de silenciamiento que contiene AGO1 y AGO2 en la subfamilia AGO, respectivamente (4-6). Por el contrario, ARNpi de ~24-30 nt son independientes de Dicer y requieren AGO3, aubergina (AUB) y PIWI en la subfamilia PIWI para su biogénesis (4-6). Los análisis genéticos (7-10) han demostrado claramente un papel para DCR2 de *D. melanogaster* en la inmunidad y la biogénesis de ARNip virales que seleccionan como diana diversos virus ARN de hebra positiva (+), incluyendo virus de Flock House (VFH), virus de la parálisis del grillo, virus C de *Drosophila* (VCD) y virus Sindbis (VSIN). La clonación y la secuenciación de ARN pequeños a partir de células de *Drosophila* infectadas con VFH indican además que los productos intermedios de replicación de ARNbc viral (vRI-ARNbc) son el sustrato de DCR2 y el precursor de ARNip virales (11-12). La susceptibilidad de *Drosophila* al virus X de *Drosophila* (VXD), que contiene un genoma de ARNbc, está influida por componentes a partir de las rutas tanto de ARNip (por ejemplo, AGO2 y R2D2) como de ARNpi (por ejemplo, AUB y PIWI) (13). Sin embargo, aún no se ha notificado la detección de ARN pequeños derivados a partir de cualquier virus ARNbc (1, 13).

En primer lugar se detectaron ARN pequeños derivados de virus en plantas infectadas con un virus ARN+ (14). Las proteínas Dicer implicadas en la producción de ARNip que seleccionan como diana tanto virus ARN+ como virus ADN se han identificado en *Arabidopsis thaliana* (2-3), que codifica para AGO en la subfamilia AGO pero no en la subfamilia PIWI (15). La clonación y la secuenciación de ARNip virales de plantas sugiere que pueden procesarse o bien a partir de vRI-ARNbc o bien de regiones de horquilla de precursores de ARN monocatenario (16-20). También se ha demostrado la producción de ARNip virales en hongos, gusanos de seda, mosquitos y nematodos en respuesta a infección con virus ARN+ y recientemente se han clonado y secuenciado ARN de silenciamiento pequeños virales producidos en hongos y mosquitos (21-25).

Por tanto, los datos disponibles ilustran que la acumulación de ARN de silenciamiento pequeños derivados de virus es una característica común de una respuesta inmunitaria activa frente a la infección viral en diversas especies huésped eucariotas.

Sumario

La divulgación proporciona un método para el descubrimiento de virus que no depende ni de la amplificación ni de la purificación de partículas virales. Muchas enfermedades humanas tales como aproximadamente la mitad de todos los casos analizados de encefalitis y gastroenteritis humanas, no tienen una etiología identificada. Por tanto, el descubrimiento de nuevos virus debería facilitar la identificación de virus patógenos humanos, mejorar la

comprensión de su transmisión y proporcionar herramientas de diagnóstico y dianas para el desarrollo de antivirales.

Los métodos se basan en el mecanismo de invertebrados, plantas, algas, hongos, etc. de procesamiento de ARN inhibidores pequeños virales, o ARN de “silenciamiento pequeños”, (ARNip), microARN (miARN) y/o ARN que interaccionan con PIWI (ARNpi), incluyendo inmunidad viral mediada por miARN, ARNpi, ARNip y/o iARN en plantas e invertebrados, incluyendo insectos (*Drosophila melanogaster* y mosquitos) y nematodos (*Caenorhabditis elegans*), y algas, hongos, cianobacterias y similares.

La invención proporciona un método para descubrir un genoma microbiano según la reivindicación 1 en el presente documento.

En realizaciones alternativas, los métodos comprenden además determinar la secuencia del cóntigo ensamblado; o comprende además:

(a) buscar en una base de datos de secuencias virales o de microorganismo usando el al menos un cóntigo para identificar una secuencia, o subsecuencia de la misma, que codifica para proteína, ácido nucleico o genoma viral o de microorganismo, que tiene una homología significativa con respecto al cóntigo ensamblado; o

(b) el método de (a), en el que la base de datos comprende secuencias de nucleótidos no redundantes; o

(c) el método de (a), en el que la base de datos comprende secuencias de traducción *in silico*,

en los que opcionalmente el cóntigo ensamblado tiene una homología significativa con respecto a un género o genoma viral conocido.

En realizaciones alternativas, los métodos comprenden además buscar en una base de datos de secuencias virales o de microorganismo usando el al menos un cóntigo para identificar una secuencia, o subsecuencia de la misma, que codifica para proteína, ácido nucleico o genoma viral o de microorganismo, que tiene un porcentaje de homología de al menos el 50% al 100% con respecto a la totalidad o parte del cóntigo ensamblado.

En realizaciones alternativas, los métodos comprenden además realizar un análisis filogenético de la secuencia que codifica para proteína, ácido nucleico o genoma viral o de microorganismo identificada con el cóntigo.

En realizaciones alternativas, los métodos comprenden además identificar y anotar el análisis filogenético de la secuencia viral identificada con el cóntigo.

En realizaciones alternativas, las secuencias de ARN o nucleótidos obtenidas están sustancialmente purificadas o aisladas de un organismo de interés.

En realizaciones alternativas, los métodos comprenden además purificar sustancialmente ARNpi a partir de un organismo de interés y secuenciar los fragmentos de ARN para obtener una biblioteca de ARN. En realizaciones alternativas, los métodos comprenden además eliminar segmentos secuenciados de la biblioteca que se solapan con la secuencia genómica del organismo de interés a partir de la que se derivó el ARN.

En realizaciones alternativas, los métodos comprenden además el rellenado de los huecos entre los cóntigos. En una realización, el rellenado de los huecos entre los cóntigos comprende el uso de RT-PCR y/o secuenciación para rellenar los huecos entre los cóntigos.

En realizaciones alternativas, los métodos comprenden además completar una secuencia genómica de un virus o un microorganismo que comprende el cóntigo usando 5’-RACE y 3’-RACE.

En una realización, el organismo o los organismos es/son un invertebrado, un insecto *(Anthropoda),* un nematodo *(Nemapoda), Mollusca, Porifera,* una planta, hongos, algas, cianobacterias*;* o el organismo o los organismos se identifican o no se identifican y se derivan a partir de una muestra del entorno. En una realización, la muestra del entorno es una muestra de tierra, una muestra de agua o una muestra de aire.

En una realización, la invención proporciona métodos para identificar un virus, que comprenden:

construir una biblioteca de ARN pequeños a partir de un organismo u organismos;

secuenciar de manera masiva la biblioteca de ARN pequeños;

ensamblar los ARN pequeños secuenciados usando (a) todos los ARNpi; o (b) ARNpi de una longitud definida en

una pluralidad de cóntigos;

identificar y eliminar aquellas secuencias ensambladas mapeadas sobre el genoma del organismo para proporcionar un conjunto enriquecido de cóntigos;

realizar una búsqueda de homología de cóntigos frente a virus conocidos tanto a nivel de nucleótido como de proteína;

opcionalmente usar RT-PCR y secuenciar para rellenar los huecos entre los cóntigos que muestran similitudes limitadas con un virus conocido;

completar la secuencia genómica de longitud completa del virus identificado con 5'-RACE y 3'-RACE; y

anotar el virus identificado.

En una realización, el organismo o los organismos es/son un invertebrado, un insecto *(Anthropoda),* un nematodo *(Nemapoda), Mollusca, Porifera,* una planta, hongos, algas, cianobacterias*;* o el organismo o los organismos se identifican o no se identifican y se derivan a partir de una muestra del entorno. En una realización, la muestra del entorno es una muestra de tierra, una muestra de agua o una muestra de aire.

Los detalles de una o más realizaciones de la divulgación se exponen en los dibujos adjuntos y la descripción a continuación. Otras características, objetos y ventajas de la divulgación resultarán evidentes a partir de la descripción y los dibujos, y a partir de las reivindicaciones.

Descripción de los dibujos

Las figuras 1A y B muestran la distribución de cóntigos de ARNip viral ensamblados en el genoma de ARN tripartito y bipartito de (a) VMP y (b) VFH; tal como se trata en detalle en el ejemplo 1, más adelante.

La figura 2 muestra el descubrimiento de tres nuevos virus mediante el ensamblaje de ARNip virales. Se mapearon un total de 54, 34 y 19 cóntigos ensamblados a partir de ARNip secuenciados en TVDm, BVDm y TRVDm, respectivamente. Se mostró la organización del genoma de VEE como referencia para TRVDm. Se mostraron los % de identidades de secuencia de proteína de cóntigos ensamblados (barras rojas) de los tres virus con respecto a virus relacionados en la parte superior o inferior; tal como se trata en detalle en el ejemplo 1, más adelante.

La figura 3A, la figura 3B, la figura 3C ilustran un análisis filogenético de virus recién identificados (indicados por una flecha roja) según las similitudes de RdRP virales con método Clustal W; tal como se trata en detalle en el ejemplo 1, más adelante.

La figura 4A y la figura 4B ilustran la distribución de cóntigos de ARNip viral ensamblados en el genoma monopartito y el genoma de ARN bipartito de VSIN y un nuevo nodavirus respectivamente; tal como se trata en detalle en el ejemplo 1, más adelante.

La figura 5 ilustra la posición y distribución de cóntigos de ARNip de VFH y VSIN ensamblados a partir de ARN pequeños secuenciados a partir de (figura 5A) células S2 *de Drosophila* infectadas con el mutante de deleción B2 de VFH (11), (figura 5B) una cepa de *C. elegans* transgénica en el contexto de mutante defectuoso para iARN 1 (rde-1) que porta un replicón ARN1 de VFH en el que la secuencia codificante de B2 se remplazó por la de GFP (29), y (figura 5C) mosquitos adultos infectados con VSIN (22); tal como se trata en detalle en el ejemplo 2, más adelante.

La figura 6 ilustra el descubrimiento de los virus ARNbc TVD (figura 6A) y BVD (figura 6B), y los virus ARN+ TrVD (figura 6C) y NVM (figura 6D) de células S2-GMR mediante vdSAR; tal como se trata en detalle en el ejemplo 2, más adelante.

La figura 7 ilustra los cuatro virus ARN infecciosos contenidos en células S2-GMR: la figura 7A ilustra que TVD, BVD, VXD y NVA se detectaron todos mediante RT-PCR en células S2 no contaminadas 4 días tras la inoculación con el sobrenadante de las células S2-GMR; la figura 7B ilustra la detección de un ARN-ID derivado a partir de ARN2 de NVA en células S2-GMR (carril derecho) y en S2 tras la inoculación con el sobrenadante de células S2-GMR (carril izquierdo) mediante hibridaciones de transferencia de tipo Northern usando una sonda que reconoce los 120 nt del extremo 3'-terminal de ARN2; la figura 7C ilustra la estructura del ARN-ID clonado de NVA (parte superior) y el mapeo de los ARNip de 21 nt de apareamiento perfecto secuenciados a partir de células S2-GMR en las hebras positiva (azul) y negativa (roja) de ARN2 de NVA (ventanas de 20 nt) (parte inferior); tal como se trata en detalle en el ejemplo 2, más adelante.

La figura 8 ilustra la distribución del tamaño (figura 8A) y la composición de nucleótidos agregada (figura 8B) de ARN pequeños derivados de virus en células OSS de *Drosophila*; tal como se trata en detalle en el ejemplo 2, más adelante.

Símbolos de referencia similares en los diversos dibujos indican elementos similares.

Descripción detallada

Los métodos para el ensamblaje de genoma viral y descubrimiento de virus pueden usar ARN inhibidores pequeños, o ARN de "silenciamiento pequeños" (ARNip), microARN (miARN) y/o ARN que interaccionan con PIWI (ARNpi), incluyendo ARNip, miARN y/o ARNpi aislados o secuenciados a partir de organismos invertebrados tales como insectos (*Anthropoda*), nematodos *(Nemapoda), Mollusca, Porifera* y otros invertebrados, y/o plantas, hongos o algas, cianobacterias y similares.

Tal como se describe en el ejemplo 2, se encontró que ARN de silenciamiento pequeños virales producidos por animales invertebrados son de secuencia solapante y pueden ensamblarse en fragmentos contiguos largos del genoma viral invasor a partir de bibliotecas de ARN pequeños secuenciadas mediante plataformas de próxima generación. En base a este hallazgo, se desarrolló un enfoque de descubrimiento de virus en invertebrados mediante secuenciación masiva y ensamblaje de ARN pequeños totales (vdSAR, *virus discovery in invertebrates by deep sequencing and assembly of total small RNAs*) aislados a partir de un organismo huésped de interés.

Tal como se describe en el ejemplo 2, realizaciones alternativas de la invención revelaron una infección mixta de líneas celulares de *Drosophila* y mosquitos adultos por múltiples virus ARN, cinco de los cuales eran nuevos. El análisis de ARN pequeños a partir de células de *Drosophila* infectadas mixtas mostró que la infección de los tres virus ARNbc distintos desencadenó la producción de ARNip virales con características similares a ARNip derivados a partir de virus ARN+. El estudio también reveló la producción y el ensamblaje de ARNpi derivados de virus en células de *Drosophila*, lo que sugiere una función novedosa de los ARNpi en la inmunidad viral. Por tanto, las características únicas del vdSAR de la invención pueden descubrir nuevos patógenos virales animales y humanos portados por invertebrados y artrópodos.

Tal como se usa en el presente documento y en las reivindicaciones adjuntas, las formas en singular "un", "una" y "el/la" incluyen referencias en plural a menos que el contexto indique claramente lo contrario. Por tanto, por ejemplo, referencia a "un ARNip" incluye una pluralidad de tales ARNip y una referencia a "el virus" incluye referencia a uno o más virus, etc.

A menos que se defina lo contrario, todos los términos técnicos y científicos usados en el presente documento tienen el mismo significado tal como se entienden comúnmente por un experto habitual en la técnica a la que pertenece esta divulgación. Aunque puede usarse cualquier método y reactivo similar o equivalente a los descritos en el presente documento en la práctica de los métodos y composiciones dados a conocer, se describen ahora los métodos y materiales a modo de ejemplo.

Además, el uso de "o" significa "y/o" a menos que se indique lo contrario. De manera similar, "comprenden", "comprende", "que comprende(n)", "incluyen", "incluye" y "que incluye(n)" son intercambiables y no pretenden ser limitativos.

Debe entenderse además que cuando las descripciones de diversas realizaciones usan el término "que comprende(n)", los expertos en la técnica entenderán que en algunos casos específicos, una realización puede describirse alternativamente usando la expresión "que consiste esencialmente en" o "que consiste en".

La divulgación de la patente estadounidense n.º 7.211.390, describe técnicas asociadas con "secuenciación masiva".

En inmunidad, la infección viral induce la producción de ARN de interferencia pequeños (ARNip) derivados de virus, ARNpi y miARN que posteriormente guían el aclaramiento de ARN viral específico mediante el mecanismo de interferencia de ARN (iARN), ARNpi y miARN. En *D. melanogaster,* por ejemplo, se producen ARNip de 21 nucleótidos de longitud que seleccionan como diana varios virus ARN de hebra positiva (+) mediante Dicer-2 a partir del procesamiento de productos intermedios de replicación de ARNbc sintetizados durante la replicación de ARN viral. Asistidos por la proteína de unión a ARNbc R2D2, estos ARNip virales se cargan entonces en Argonauta-2 para dirigir el aclaramiento de ARN viral (Galiana-Arnoux *et al*., 2006; Wang *et al*., 2006; Zambon *et al*., 2006). Como contradefensa, los virus codifican para proteínas de patogénesis esenciales que son supresores virales de iARN (VSR) (Li y Ding, 2006; Mlotshwa *et al*., 2008). Los VSR pueden inhibir o bien la producción o bien la actividad de ARNip virales seleccionando como diana los precursores de ARNbc, ARNip o proteínas Argonauta. Varios virus ADN de replicación nuclear producen microARN derivados de virus tras la infección de sus células huésped de mamífero y muchas proteínas codificadas por virus ARN y ADN de mamífero presentan actividad VSR. Sin embargo, el consenso actual es que en los vertebrados el ARNbc viral desencadena respuestas de PKR e interferón en lugar de

la respuesta iARN.

La divulgación proporciona un método para el descubrimiento de virus que es independiente tanto de la amplificación como de la purificación de partículas virales. Muchas de las enfermedades humanas tales como aproximadamente la mitad de todos los casos analizados de encefalitis y gastroenteritis humanas, no tienen una etiología identificada. Por tanto, el descubrimiento de nuevos virus o la identificación de la presencia de infección viral pueden facilitar la identificación de virus patógenos humanos, mejorar la comprensión de su transmisión y proporcionar herramientas de diagnóstico y dianas para el desarrollo de antivirales.

La divulgación se basa, en parte, en la comprensión del mecanismo de la inmunidad viral mediada por iARN, incluyendo basada en ARNpi, miARN y ARNip. En esta inmunidad, la infección viral induce la producción de ARN de interferencia pequeños (ARNip) derivados de virus, ARNpi y miARN, que posteriormente guían el aclaramiento de ARN viral específico mediante el mecanismo de interferencia de ARN (iARN) (basado en ARNpi, miARN y ARNip). En *D. melanogaster,* por ejemplo, se producen ARNip de 21 nucleótidos de longitud que seleccionan como diana varios virus ARN de hebra positiva (+) mediante Dicer-2 a partir del procesamiento de productos intermedios de replicación de ARNbc sintetizados durante la replicación de ARN viral. Asistidos por la proteína de unión a ARNbc R2D2, estos ARNip virales se cargan entonces en Argonauta-2 para dirigir el aclaramiento de ARN viral (Galiana-Arnoux *et al*., 2006; Wang *et al*., 2006; Zambon *et al*., 2006). Como contradefensa, los virus codifican para proteínas de patogénesis esenciales que son supresores virales de iARN (VSR) (Li y Ding, 2006; Mlotshwa *et al*., 2008). Los VSR pueden inhibir o bien la producción o bien la actividad de ARNip, ARNpi y miARN virales seleccionando como diana los precursores de ARNbc, ARNip, ARNpi y miARN o proteínas Argonauta. Varios virus ADN de replicación nuclear producen microARN derivados de virus tras la infección de sus células huésped de mamífero y muchas proteínas codificadas por virus ARN y ADN de mamífero presentan actividad VSR (Ding y Voinnet, 2007). Sin embargo, el consenso actual es que en los vertebrados el ARNbc viral desencadena respuestas de PKR e interferón en lugar de la respuesta de iARN (ARNip, ARNpi y miARN).

La divulgación demuestra mediante las tecnologías de secuenciación de próxima generación ARNip, ARNpi y miARN virales producidos en plantas y moscas de la fruta infectadas con virus ARN de hebra positiva, que están relacionados estrechamente con virus patógenos humanos tales como poliovirus y virus del Nilo occidental. Los resultados muestran que ARNip virales producidos por el sistema inmunitario del huésped en respuesta a infección viral son de secuencia solapante y pueden volver a ensamblarse en fragmentos contiguos largos (cóntigos) del genoma de ARN viral infectivo usando programas de ensamblaje desarrollados para secuenciación de genoma de lectura corto. A diferencia de los ARNip, ARNpi y miARN individuales, los cóntigos ensamblados a partir de ARNip, ARNpi y miARN virales pueden traducirse para dar secuencias de proteína *in silico* para búsquedas de homología para identificar nuevos virus que pueden estar relacionados sólo de manera lejana con virus conocidos.

La divulgación demuestra que la secuenciación masiva mediante las tecnologías de próxima generación y el ensamblaje de ARNip, ARNpi y miARN derivados de virus pueden emplearse como un nuevo enfoque para el descubrimiento y la identificación de virus. De hecho, el examen de una biblioteca de ARN pequeños secuenciada recientemente (Flynt *et al*., 2009) preparada a partir de una línea celular de *Drosophila* encontró que la línea celular está infectada con al menos cinco virus ARN distintos. Éstos incluyen dos virus conocidos y tres nuevos virus que pertenecen a diferentes géneros no descritos anteriormente. Puesto que la infección por virus de plantas e invertebrados inevitablemente da como resultado la producción de ARNip, ARNpi y miARN derivados de virus, esta invención no depende de la capacidad ni de amplificar el virus ni de purificar la partícula viral para enriquecer los ácidos nucleicos virales, que es esencial para las tecnologías actuales. De manera importante, cualquier virus detectado por el método de la divulgación está vivo y puede replicarse debido a que los ARNip, ARNpi y miARN virales son productos de una respuesta inmunitaria del huésped activa frente a la infección viral.

La observación de que ARNip, ARNpi y miARN virales individuales pueden volver a ensamblarse en fragmentos de genoma más largos del virus invasor proporciona un excitante nuevo método para el descubrimiento de virus mediante secuenciación masiva y ensamblaje de ARNip, ARNpi y miARN virales. A diferencia de ARNip, ARNpi y miARN individuales, los cóntigos ensamblados a partir de ARNip, ARNpi y miARN virales pueden traducirse para dar secuencias de proteína *in silico* para búsquedas de homología para identificar nuevos virus que pueden estar relacionados de manera lejana con virus conocidos.

La divulgación proporciona un marco del VDsiR compuesto por análisis bioinformático y verificación experimental. El ensamblaje de ARN pequeño es un componente útil del sistema, el número de secuencias de entrada y programas distintos tienen impacto sobre el resultado. En un estudio piloto (descrito en el presente documento), se encontró que Velvet era un programa útil para el proyecto, que emplea el principio de gráficos de Bruijn para desarrollar secuencias continuas a partir de lecturas cortas en un tiempo de ejecución corto (Zerbino *et al*, 2008).

Por tanto, la divulgación proporciona en una realización, un método que comprende (i) obtener secuencias de nucleótidos a partir de bibliotecas de ARN pequeños que comprenden una pluralidad de fragmentos de ARN de 18-28 nucleótidos que se producen de manera natural para obtener una biblioteca de ARN pequeños secuenciada; (ii) ensamblar las secuencias en la biblioteca de ARN pequeños secuenciada para dar al menos una secuencia contigua

que comprende una pluralidad de secuencias de fragmentos de ARN de nucleótidos; opcionalmente rellenar los huecos en una secuencia mediante técnicas de RT-PCR; (iii) buscar en una base de datos de secuencias virales usando la al menos una secuencia contigua para identificar una secuencia viral que tiene un porcentaje de homología de al menos el 50%-100% con respecto a la secuencia contigua; (iv) identificar y anotar el análisis filogenético de la secuencia viral identificada con la secuencia contigua.

Se entenderá que puede proporcionarse una biblioteca de secuencias por un tercero o puede hacerse disponible para un usuario de varias maneras (es decir, Internet, medio legible por ordenador y similares) y por tanto el procedimiento descrito anteriormente puede adaptarse para llevar a cabo el procedimiento e identificar o anotar un virus, por consiguiente. En alguna realización, sin embargo, la biblioteca puede ser una biblioteca de muestra que comprende sustancialmente ARN purificado de un organismo de interés. En tales casos, se llevan a cabo técnicas de secuenciación masiva y se crea una biblioteca de secuencias. Aún en otra realización, se proporciona una muestra que comprende sustancialmente purificar fragmentos de ARN pequeño a partir de un organismo de interés en cuyo caso se realiza la secuenciación de los fragmentos de ARN para obtener la biblioteca de ARN pequeños.

Aún en otra realización, si se proporciona una muestra de ARN en bruto de un organismo o si se desea un aumento de la búsqueda de homología, el método puede incluir opcionalmente eliminar segmentos secuenciados de la biblioteca que se solapan con la secuencia genómica del organismo de interés a partir de la que se derivó el ARN.

Aún en otra realización, el método comprende además completar una secuencia genómica de un virus que comprende la secuencia contigua usando 5'-RACE y 3'-RACE.

Por ejemplo, la divulgación demuestra en las realizaciones específicas y prueba del principio que un método que incluye las etapas de construcción de una biblioteca de ARN pequeños a partir de cultivo celular o insectos adultos tales como mosquitos o moscas de la fruta; secuenciación masiva de las bibliotecas de ARN pequeños con un 2G Analyse de Illumina; ensamblaje de los ARN pequeños secuenciados mediante Velvet usando o bien la totalidad de los ARN pequeños secuenciados de 18-28 nucleótidos de longitud o bien ARN pequeños de longitudes específicas tales como 21 nt y 22 nt, que lo más probablemente representan los productos de *Drosophila* Dicer-2 y Dicer-1, respectivamente, para generar un cóntigo/cóntigos, los cóntigos de ARNip derivados de virus pueden incluir características tales como enriquecimiento específico de ARN pequeños de 21 a 22 nt, la presencia de ARN pequeños de ambas polaridades y la alta densidad de ARNip (número de ARNip/longitud de cóntigos); identificación y eliminación de aquellas secuencias ensambladas mapeadas sobre el genoma del huésped cuando se conoce la secuencia del genoma, lo que reduce el número de los candidatos para las siguientes etapas; búsqueda de homología de cóntigos con virus conocido tanto a nivel de nucleótido como de proteína; en una realización opcional, pueden usarse RT-PCR y secuenciación para rellenar los huecos entre los cóntigos que muestran similitudes limitadas con un virus conocido; opcionalmente completar la secuencia genómica de longitud completa del virus identificado con 5'-RACE y 3'-RACE; y anotación y análisis filogenético del virus identificado, dio como resultado la identificación de 2 virus conocidos y 3 virus novedosos a partir de una muestra de *D. melanogaster*.

Tal como se usa en el presente documento, una muestra es cualquier muestra que puede contener un virus. Por tanto, la muestra puede obtenerse del entorno, de un organismo específico (incluyendo plantas, insectos y mamíferos). Una muestra del entorno puede obtenerse de cualquier número de fuentes (tal como se describió anteriormente), incluyendo, por ejemplo, heces de insecto, fuentes termales, tierra y similares. Puede utilizarse cualquier fuente de ácidos nucleicos en forma purificada o no purificada como material de partida. Por tanto, los ácidos nucleicos pueden obtenerse de cualquier fuente que esté contaminada por un organismo infeccioso (por ejemplo un virus). La muestra puede ser un extracto de cualquier muestra corporal tal como sangre, orina, líquido cefalorraquídeo, tejido, frotis vaginal, deposiciones, líquido amniótico o enjuague bucal de cualquier organismo mamífero. Para organismos no mamíferos (por ejemplo, invertebrados), la muestra puede ser una muestra de tejido, muestra de saliva, material fecal o material en el tracto digestivo del organismo. Por ejemplo, en pruebas de horticultura y agricultura la muestra puede ser una planta, tierra, líquido u otro producto de horticultura o agricultura; en pruebas de alimentos la muestra puede ser alimento fresco o alimento procesado (por ejemplo leche artificial para lactantes, marisco, productos frescos y alimentos envasados); y en pruebas del entorno la muestra puede ser líquido, tierra, tratamiento de aguas residuales, lodo y cualquier otra muestra en el entorno.

La muestra puede procesarse usando técnicas conocidas en la técnica para secuenciación masiva. En algunas realizaciones, la muestra se trata con un inhibidor de RNasa para impedir la degradación de oligonucleótidos de ARN en la muestra. En la técnica se conocen cócteles e inhibidor de RNasa y están disponibles comercialmente.

En la técnica se conocen técnicas de secuenciación masiva tal como se describió anteriormente y en otra parte en el presente documento. Una muestra que comprende fragmentos de ARN pequeño ya sea purificado, sustancialmente purificado o no purificado, puede someterse a secuenciación masiva para secuenciar un gran número de fragmentos de ARN en la muestra.

Tal como se describe en el presente documento, las secuencias pueden alinearse entonces y aparearse para generar al menos uno, normalmente una pluralidad, de cóntigos que comprenden una pluralidad de fragmentos de

ARN solapantes y adyacentes. En la técnica se conocen algoritmos y programas informáticos para realizar tal apareamiento y generación de cóntigos (véase, por ejemplo, Velvet: algorithms for de novo short read assembly using de Bruijn graphs. D.R. Zerbino y E. Birney. Genome Research 18:821-829; e información en la World Wide Web en (www-ebi.ac.uk/-zerbino/velvet/velvet_poster.pdf y http:~~www-ebi.ac.uk/-zerbino/velvet/) (obsérvese que la url se ha modificado con "-" para evitar hipervínculos). Los resultados de, por ejemplo, Velvet que comprenden cóntigos más largos de entre 30-50, 50-100 y varios cientos de bases pueden usarse para examinar bases de datos de secuencias de ácido nucleico directamente o puede traducirse para examinar bases de datos de aminoácidos.

Están disponibles varias bases de datos fuente que contienen o bien una secuencia de ácido nucleico y/o una secuencia de aminoácidos deducida para identificar o determinar secuencias u homólogos relacionados. Puede usarse la totalidad o una parte representativa de las secuencias para buscar en una base de datos de secuencias (por ejemplo, GenBank, PFAM o ProDom), o bien simultáneamente o bien individualmente. En la técnica se conocen varios métodos diferentes de realizar tales búsquedas de secuencia. Las bases de datos pueden ser específicas para un organismo particular o una colección de organismos. Los datos de secuencia se alinean con las secuencias en la base de datos o las bases de datos usando algoritmos diseñados para medir la homología entre dos o más secuencias.

Tales métodos de alineación de secuencias incluyen, por ejemplo, BLAST (Altschul *et al*., 1990), BLITZ (MPsrch) (Sturrock y Collins, 1993) y FASTA (Person y Lipman, 1988). La secuencia sonda (por ejemplo, los datos de secuencia del clon) puede ser de cualquier longitud y se reconocerá como homóloga basándose en un valor de homología umbral. El valor umbral puede predeterminarse, aunque esto no se requiere. El valor umbral puede basarse en la longitud de polinucleótido particular. Para alinear las secuencias pueden usarse varios procedimientos diferentes. Normalmente, se usan los algoritmos de Smith-Waterman o Needleman-Wunsch. Sin embargo, tal como se mencionó pueden usarse procedimientos más rápidos tales como BLAST, FASTA, PSI-BLAST.

Por ejemplo, la alineación óptima de secuencias para alinear una ventana de comparación puede realizarse mediante el algoritmo de homología local de Smith (Smith y Waterman, Adv Appl Math, 1981; Smith y Waterman, J Teor Biol, 1981; Smith y Waterman, J Mol Biol, 1981; Smith *et al*, J Mol Evol, 1981), mediante el algoritmo de alineación de homología de Needleman (Needleman y Wunsch, 1970), mediante el método de búsqueda de similitud de Pearson (Pearson y Lipman, 1988), mediante implementaciones computarizadas de estos algoritmos (GAP, BESTFIT, FASTA y TFASTA en el paquete de software de Wisconsin Genetics, versión 7.0, Genetics Computer Group, 575 Science Dr., Madison, Wis., o el paquete de software Sequence Analysis del Genetics Computer Group, Universidad de Wisconsin, Madison, Wis.) o mediante inspección, y se selecciona la mejor alineación (es decir, la que da como resultado el mayor porcentaje de homología con respecto a la ventana de comparación) generada por los diversos métodos. Entonces puede predecirse la similitud de las dos secuencias (es decir, la secuencia sonda y la secuencia de base de datos).

Tal software hace coincidir secuencias similares asignando grados de homología con respecto a diversas deleciones, sustituciones y otras modificaciones. Los términos "homología" e "identidad" en el contexto de dos o más ácidos nucleicos o secuencias de polipéptido, se refieren a dos o más secuencias o subsecuencias que son iguales o que tienen un porcentaje especificado de residuos de aminoácido o nucleótidos que es igual cuando se comparan y se alinean para una correspondencia máxima con respecto a una ventana de comparación o región designada tal como se mide usando cualquier número de algoritmos de comparación de secuencias o mediante alineación manual e inspección visual.

Para la comparación de secuencias, normalmente una secuencia actúa como secuencia de referencia, con la que se comparan las secuencias de prueba. Cuando se usa un algoritmo de comparación de secuencias, se introducen las secuencias de prueba y de referencia en un ordenador, se designan coordinados de subsecuencia, si es necesario, y se designan parámetros de programa de algoritmo de secuencia. Pueden usarse los parámetros del programa por defecto, o pueden designarse parámetros alternativos. El algoritmo de comparación de secuencias calcula entonces el porcentaje de identidades de secuencia para las secuencias de prueba con respecto a la secuencia de referencia, basándose en los parámetros del programa.

Una "ventana de comparación", tal como se usa en el presente documento, incluye la referencia a un segmento de una cualquiera del número de posiciones contiguas seleccionadas del grupo que consiste en desde 20 hasta 600, habitualmente de aproximadamente 50 a aproximadamente 200, más habitualmente de aproximadamente 100 a aproximadamente 150 en las que puede compararse una secuencia con una secuencia de referencia del mismo número de posiciones contiguas tras alinearse de manera óptima las dos secuencias.

Un ejemplo de un algoritmo útil son los algoritmos BLAST y BLAST 2.0, que se describen en Altschul *et al*., Nuc. Acids Res. 25:3389-3402 (1977) y Altschul *et al*., J. Mol. Biol. 215:403-410 (1990), respectivamente. El software para realizar los análisis de BLAST está disponible al público a través del National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/). Este algoritmo implica identificar en primer lugar pares de secuencias de alta puntuación (HSP) identificando palabras cortas de longitud W en la secuencia de consulta, que o bien coinciden o bien satisfacen alguna puntuación de umbral de valor positivo cuando se alinea con una palabra de la misma

longitud en una secuencia de la base de datos. T se denomina el umbral de puntuación de palabra de vecindad (Altschul *et al*., citado anteriormente). Estos resultados positivos de palabra de vecindad iniciales actúan como simientes para iniciar búsquedas para encontrar HSP más largos que las contienen. Los resultados positivos de palabra se extienden en ambos sentidos a lo largo de cada secuencia siempre que pueda aumentarse la puntuación de alineación acumulativa. Las puntuaciones acumulativas se calculan usando, para secuencias de nucleótidos, los parámetros M (puntuación de recompensa para un par de residuos coincidentes; siempre >0). Los parámetros del algoritmo BLAST W, T y X determinan la sensibilidad y la velocidad de la alineación. El programa BLASTN (para secuencias de nucleótidos) usa como parámetros por defecto una longitud de palabra (W) de 11, una expectativa (E) de 10, M=5, N=-4 y una comparación de ambas hebras.

El algoritmo BLAST también realiza un análisis estadístico de la similitud entre dos secuencias (véase, por ejemplo, Karlin y Altschul, Proc. Natl. Acad. Sci. USA 90:5873 (1993)). Una medida de la similitud proporcionada por el algoritmo BLAST es la mínima probabilidad de suma (P(N)), que proporciona una indicación de la probabilidad a la que se producirá una coincidencia entre dos secuencias de nucleótidos al azar. Por ejemplo, un ácido nucleico se considera similar a una secuencia de referencia si la mínima probabilidad de suma en una comparación del ácido nucleico de prueba con el ácido nucleico de referencia es menor de aproximadamente 0,2, más preferiblemente menor de aproximadamente 0,01 y lo más preferiblemente menor de aproximadamente 0,001.

La homología de secuencia significa que dos secuencias de polinucleótido son homólogas (es decir, en una base nucleótido a nucleótido) con respecto a la ventana de comparación. Un porcentaje de identidad u homología de secuencia se calcula comparando dos secuencias alineadas de manera óptima con respecto a la ventana de comparación, determinando el número de posiciones en las que aparece la base de ácido nucleico idéntica (por ejemplo, A, T, C, G, U o I) en ambas secuencias para proporcionar el número de posiciones coincidentes, dividiendo el número de posiciones coincidentes entre el número total de posiciones en la ventana de comparación (es decir, el tamaño de ventana) y multiplicando el resultado por 100 para proporcionar el porcentaje de homología de secuencia. Esta homología sustancial indica una característica de una secuencia de polinucleótido, en la que el polinucleótido comprende una secuencia que tiene una homología de secuencia de al menos el 60 por ciento, normalmente una homología de al menos el 70 por ciento, a menudo una homología de secuencia del 80 al 90 por ciento y lo más comúnmente una homología de secuencia de al menos el 99 por ciento en comparación con una secuencia de referencia de una ventana de comparación de al menos 25-50 nucleótidos, en la que el porcentaje de homología de secuencia se calcula comparando la secuencia de referencia con la secuencia de polinucleótido que puede incluir deleciones o adiciones que en total son el 20 por ciento o menos de la secuencia de referencia con respecto a la ventana de comparación.

Secuencias que tienen una homología suficiente pueden identificarse adicionalmente mediante cualquier anotación contenida en la base de datos, incluyendo, por ejemplo, información de la especie y la fuente. Por consiguiente, en una muestra típica, se obtendrán una pluralidad de secuencias de ácido nucleico, se secuenciarán, se generarán cóntigos y se identificarán secuencias homólogas correspondientes a partir de una base de datos. Esta información proporciona un perfil de los polinucleótidos presentes en la muestra, incluyendo una o más características asociadas con el polinucleótido incluyendo el organismo y otra información de la fuente asociada con esa secuencia o cualquier polipéptido codificado por esa secuencia basándose en la información de la base de datos.

En algunos casos puede ser deseable someter a RT-PCR partes de regiones en blanco entre cóntigos o fragmentos. Tales métodos pueden usarse para rellenar los espacios en blanco y mejorar la fidelidad de cualquier búsqueda de homología. En la técnica se conocen métodos de RT-PCR.

En algunos casos puede ser deseable realizar una amplificación de la secuencia de ácido nucleico presente en una muestra o un clon particular que se ha aislado. En esta realización, la secuencia de ácido nucleico se amplifica mediante reacción de PCR o reacción similar conocida por los expertos en la técnica. Están disponibles kits de amplificación disponibles comercialmente para llevar a cabo tales reacciones de amplificación.

Además, puede usarse RACE, o amplificación rápida de extremos de ADNc, para obtener la secuencia de longitud completa de un transcrito de ARN encontrado en una célula. RACE da como resultado la producción de una copia de ADNc de la secuencia de ARN de interés, producida mediante transcripción inversa, seguido por amplificación mediante PCR de las copias de ADNc. Las copias de ADNc amplificadas se secuencian entonces y, si son lo suficientemente largas, deberían mapearse en un único ARNm ya descrito, cuya secuencia completa se conoce. RACE puede proporcionar la secuencia de un transcrito de ARN a partir de una secuencia conocida pequeña dentro del transcrito hasta el extremo 5' (5' RACE-PCR) o el extremo 3' (3' RACE-PCR) del ARN. Los protocolos para 5' o 3' RACE difieren ligeramente. 5' RACE-PCR comienza usando ARNm como molde para una primera ronda de reacción de síntesis de ADNc (o transcripción inversa) usando un cebador de oligonucleótido antisentido (inverso) que reconoce una secuencia conocida en el gen de interés; el cebador se denomina cebador específico para gen (GSP) y copia el molde de ARNm en el sentido de 3' a 5' para generar un producto de ADNc monocatenario específico. Tras la síntesis de ADNc, se usa la enzima desoxinucleotidil transferasa terminal (TdT) para añadir una sucesión de nucleótidos idénticos, conocida como cola homopolimérica, en el extremo 3' del ADNc. Entonces se lleva a cabo una reacción de PCR, que usa un segundo cebador específico para gen (GSP2) antisentido que se une

a la secuencia conocida, y un cebador universal (UP) sentido (directo) que se une a la cola homopolimérica añadida a los extremos 3' de los ADNc para amplificar un producto de ADNc desde el extremo 5'. 3' RACE-PCR usa la cola poliA natural que existe en el extremo 3' de todos los ARNm de eucariotas para cebar durante la transcripción inversa, de manera que este método no requiere la adición de nucleótidos mediante TdT. Los ADNc se generan usando un cebador Oligo-dT de transferencia que complementa el tramo poliA y añade una secuencia de transferencia especial al extremo 3' de cada ADNc. Entonces se usa PCR para amplificar ADNc 3' desde una región conocida usando un GSP sentido y un cebador antisentido complementario a la secuencia de transferencia.

Los siguientes ejemplos pretenden ilustrar pero no limitar la divulgación. Aunque son típicos de aquéllos que pueden usarse, alternativamente pueden usarse otros procedimientos conocidos por los expertos en la técnica.

Ejemplos

EJEMPLO 1: Descubrimiento de nuevos virus

Descubrimiento de nuevos virus a partir de línea celular de *Drosophila* mediante VDsiR. Se ejecutaron los métodos de la divulgación (VDsiR) en la biblioteca de ARN pequeños (número de registro de GEO: GSM361908 y GSM272652) construida a partir de la línea celular Schneider 2 (S2) notificada por Flynt *et al* (2009). La secuenciación mediante la plataforma de Illumina de esta línea celular S2, que procedía del laboratorio de Gerald Rubin (S2-GMR) y se mostró que estaba infectada de manera latente por VFH, proporcionó 6.922.433 de secuencias pequeñas en total, de las cuales 1.254.333 eran únicas. Se ensamblaron un total de 1639 cóntigos mediante el programa Velvet. La longitud de los cóntigos oscila entre 33 y 813 nt, y el tamaño medio y la mediana del tamaño de los cóntigos son de 92 y 65 respectivamente. Pueden mapearse 1032 cóntigos en el genoma de *D. melanogaster* (cobertura >=90%, similitud >= 90%) y 635 de estos cóntigos se solapan con loci de transposón. 51 cóntigos mostraron similitud con respecto al genoma de ARN (+) bipartito de VFH lo que apoya la conclusión de Flynt *et al* (2009). Sin embargo, el análisis mostró además que el aislado persistente de VFH en S2-GMR es idéntico al aislado de TNCL notificado anteriormente en una línea celular S2 (Li *et al*, 2007). La longitud de los cóntigos de virus VFH varía desde 33 pb hasta 339 pb con un tamaño medio de 70 pb. Se mapearon 28 y 23 en el ARN1 y ARN2 genómico de TNCL, respectivamente, cubriendo el 73% y el 91% de los ARN de longitud completa.

En primer lugar se compararon todos los cóntigos restantes con las secuencias de nucleótidos de virus conocidos en las bases de datos de NCBI. 46 cóntigos (de 33 a 401 nt de longitud) mostraron similitud con el genoma de ARNbc bipartito de virus X de *Drosophila* (VXD). El 75% del segmento A de VXD (VXD-A) estaba cubierto por 21 cóntigos, mientras que el 99% de VXD-B estaba cubierto por 26 cóntigos.

Además del mapeo de cóntigos en el genoma de *Drosophila* y de virus conocidos (VFH y VXD), todavía hay 510 cóntigos no asignados. La traducción *in silico* de estos cóntigos se comparó adicionalmente con proteínas codificadas por virus conocidos. 19 cóntigos (indicados como barras rojas en la figura 2) muestran escasa similitud con respecto a y pueden agruparse dentro de proteínas virales codificadas por miembros de familias de virus *Totiviridae*, *Birnaviridae* y *Tetraviridae*, siendo el virus conocido más cercano el virus de la mionecrosis infecciosa del camarón peneido (VMICP), virus de la enfermedad de bursitis infecciosa (VEBI), y virus de *Euprosterna elaeasa* (VEE), respectivamente (figura 2). VMICP y VEBI son ambos virus ARNbc mientras que VEE es un virus ARN de hebra positiva. Se denominó provisionalmente a los virus ARNbc como totivirus de *Drosophila melanogaster* (TVDm) y birnavirus de *Drosophila melanogaster* (BVDm), y al virus ARN de hebra (+) como tetravirus de *Drosophila melanogaster* (TRVDm).

Basándose en la orientación y las posiciones relativas de estos cóntigos con respecto a los virus relacionados, se diseñaron cebadores y se amplificó el genoma completo para los dos virus ARNbc mediante RT-PCR y 5'/3' RACE y se secuenció completamente. Se mapearon 31 y 20 cóntigos no asignados adicionales en los ARN genómicos secuenciados de TVDm y BVDm, respectivamente (indicados como barras grises en la figura 2). Se ensamblaron los ARNip secuenciados de TRVDm en un cóntigo de 3 kb, que codifica para la mayoría de la RdRP viral. Hasta ahora había 403 cóntigos no asignados. Estos cóntigos no asignados pueden corresponder a virus novedosos que no muestran similitud detectable con ninguno de los virus conocidos aunque varios de ellos pueden mapear hasta la mitad 3' del genoma de TRVDm que aún debe obtenerse mediante RT-PCR.

Análisis filogenéticos de las secuencias de RdRP viral mostraron que los tres virus identificados son nuevos. Estos análisis, tal como se ilustra en la figura 3A, figura 3B y figura 3C, también sugirieron que TVDm y TRVDm pueden definir dos nuevos géneros de virus y que TVDm forma un nuevo género con VMICP, que no se había asignado anteriormente. La inoculación de células S2 no contaminadas con el sobrenadante de GMR-S2 y análisis de transferencia de tipo Northern posteriores mostraron que los tres virus recién identificados son infecciosos. VMICP, VEBI y el virus de la necrosis pancreática infecciosa (VNPI) son todos patógenos importantes en agricultura y pesca (Muller *et al*. 2003, Poulos *et al*. 2006). Por tanto, establecer un modelo de *Drosophila* para virus relacionados facilitaría la comprensión de la patogénesis.

Para la figura 3, se calcularon los árboles filogenéticos usando el método de unión a vecino y se evaluó la fiabilidad de cada rama con análisis de remuestreo (repetición de 1000 veces). Los géneros de cada virus se enumeraron a la derecha y "?" indica un virus con un estado taxonómico no asignado. Los virus usados para el análisis filogenético se enumeran a continuación. Virus ARN de *Leishmania* 2 – 1 (VAL-2-1), virus ARN de *Leishmania* 1 - 1 (VAL-1-1), virus ARN de *Leishmania* 1 - 4 (VAL-1-4), virus 190S de *Helminthosporium victoriae* (V-190SHv), totivirus de *Botryotinia fuckeliana* 1 (TFVBf-1), virus ARN de *Sphaeropsis sapinea* 1 (VASs-1), virus de *Magnaporthe oryzae* 1 (VMo-1), virus de *Helicobasidium mompa* n.º 17 (VHm-1-17), micovirus de *Coniothyrium minitans* (MVCm), virus ARN de *Sphaeropsis sapinea* 2 (VASs-2), virus de *Magnaporthe oryzae* 2 (VMo-2), virus ARN de *Gremmeniella abietina* L1 (VAGa-1), virus ARN de *Gremmeniella abietina* L2 (VAGa-2), virus de *Ustilago maydis* H1 (VUm-H1), virus de la mionecrosis infecciosa del camarón peneido (VMICP), virus de *Giardia lamblia* (VGL), virus Z de *Zygosaccharomyces bailii* (VZZb), micovirus asociado con la enfermedad de la cereza de Amasya (MVAEcA), virus de la ascitis de la seriola (VAS), birnavirus marino (BVM), birnavirus de *Paralichthys olivaceus* (BVPo), virus de la necrosis pancreática infecciosa (VNPI), virus de la enfermedad de la bursitis infecciosa (VEBI), virus x de *Drosophila* (VXD), virus del pez cabeza de serpiente moteado (VPm), virus de *Euprosterna elaeasa* (VEE), virus de *Thosea asigna* (VTA), betavirus de *Nudaurelia capensis* (BVNC), tetravirus de *Dendrolimus punctatus* (TVDP), virus del raquitismo de *Helicoverpa armigera* (VRHA).

Descubrimiento de un nuevo virus a partir de mosquitos adultos mediante VDsiR. También se usó el sistema VDsiR para examinar la biblioteca de ARN pequeños notificada preparada a partir de mosquitos adultos infectados con VSIN (Miles *et al*. 2008). Tal como se esperaba, los cóntigos que correspondían a VSIN se ensamblaron fácilmente a partir de la biblioteca de ARN pequeños (figura 4A). Además, tres cóntigos (que cubren el 80% de la proteína) muestran similitudes con la región codificante de la proteína precursora de cápside de nodavirus de Wuhan (figura 4B), que es un miembro no asignado de *Nodoviridae* (Liu *et al*., 2006) (denominado en el presente documento nodavirus A de mosquito). Estos hallazgos sugieren que es factible usar el sistema VDsiR para el descubrimiento de virus en mosquitos. En la figura 4, se mostró el genoma de nodavirus de Wuhan (NVW) como referencia para el nuevo virus; y se muestra el porcentaje (%) de identidades de secuencia de proteína de cóntigos ensamblados (barras rojas) del nuevo virus con respecto a NVW.

Usando los métodos descritos en el presente documento, los ARNip virales producidos por el sistema inmunitario del huésped son de secuencia solapante y pueden volver a ensamblarse en fragmentos largos del genoma de ARN viral infectivo. Los ARN pequeños individuales de 21 a 24 nt de longitud derivados de virus conocidos pueden identificarse fácilmente mediante programas convencionales (Aravin *et al*., 2003). Sin embargo, estos programas se vuelven menos informativos o no funcionan cuando los ARNip virales se derivan de un virus infectivo que difiere significativamente de cualquiera de los virus conocidos. Recientemente se han secuenciado los ARNip virales producidos por el sistema inmunitario iARN de *Drosophila* en respuesta a infección mediante un virus de Flock House (VFH) mutante deficiente en VSR (Aliyari *et al*., 2008). VFH contiene un genoma de ARN de hebra positiva y codifica para una proteína VSR B2 que inhibe tanto la producción como la actividad de ARNip virales uniéndose a ARNbc y ARNip. El uso de la plataforma 454 proporcionó 4371 secuencias de ARN pequeño en total (18-28 nt), de entre las que 1177 (27%) son específicas para VFH. El precursor principal de ARNip de VFH es el ARNbc de -400 pb que corresponde a la región 5'-terminal del genoma de ARN de VFH de manera que la mayoría de los ARNip virales secuenciados son idénticos o complementarios en secuencia a esta región del genoma. A diferencia de la producción "en fases" de ARNip desde un extremo definido de sustratos de ARNbc sintéticos y precursores de ARNip trans-actuadores, sin embargo, el procesamiento de Dicer-2 del ARNbc viral en células de *Drosophila* infectadas se inicia desde múltiples posiciones. Como resultado, los ARNip virales producidos son de secuencia solapante y pueden volver a ensamblarse en fragmentos más largos del genoma de ARN de VFH (figura 1B) usando programas informáticos para el ensamblaje de genoma completo a partir de lecturas cortas (Warren *et al*. 2007, Jeck *et al*. 2007, Zerbino *et al*, 2008).

Los métodos se demostraron con ARNip virales usando una biblioteca de ARN pequeños secuenciada recientemente en el laboratorio mediante la plataforma Illumina, que proporciona muchas más lecturas en una sola ejecución que la plataforma 454. Se construyó la biblioteca a partir de ARN pequeños de plantas *Arabidopsis thaliana* infectadas con virus del mosaico del pepino (VMP), que contiene una genoma de ARN de hebra (+) tripartito. Se recogieron 2.036.929 secuencias de ARN pequeños que oscilaban entre 19-25 nt de longitud. Se ejecutaron estas secuencias mediante el programa de ensamblaje y se obtuvieron 114 cóntigos. 46 cóntigos tienen identidad con el genoma de VMP, del 96% a aproximadamente el 99% de las regiones de ARN1, ARN2 y ARN3 se han cubierto mediante estos cóntigos. El cóntigo más grande procede de ARN1 y es de 2741 pb de longitud (figura 2A).

Aliyari, R., y Ding, S. W. (2009). RNA-based viral immunity initiated by the Dicer family of host immune receptors. Immunol Rev 227, en formato impreso.

Aliyari R, Wu Q, Li HW, Wang XH, Li F, Green LD, Han CS, Li WX, Ding SW. Mechanism of induction and suppression of antiviral immunity directed by virus-derived small RNAs in *Drosophila*. Cell Host Microbe. 16 de octubre de 2008; 4(4):387-97.

Delwart EL. Viral metagenomics. Rev Med Virol. Marzo-abril de 2007; 17(2):115-31.

Ding, S. W., y Voinnet, O. (2007). Antiviral immunity directed by small RNAs. Cell 130, 413-426.

Flynt A, Liu N, Martin R, Lai EC. Dicing of viral replication intermediates during silencing of latent *Drosophila* viruses. Proc. Natl. Acad. Sci. USA. 27 de febrero de 2009. [publicación electrónica antes de impresión]

Jeck, W.R., Reinhardt, J.A., Baltrus, D.A., Hickenbotham, M.T., Magrini, V., Mardis, E.R., Dangl, J.L., Jones, C.D. (2007) Extending assembly of short DNA sequences to handle error. Bioinformatics 23:2942-2944

Lang AS, Rise ML, Culley AI, Steward GF. RNA viruses in the sea. FEMS Microbiol Rev. Marzo de 2009; 33(2):295-323.

Li H, Li WX, Ding SW. Induction and suppression of RNA silencing by an animal virus. Science. 17 de mayo de 2002; 296(5571):1319-21.

Li, W. X., Li, H., Lu, R., Li, F., Dus, M., Atkinson, P., Brydon, E. W., Johnson, K. L., Garcia-Sastre, A., Ball, L. A., *et al*. (2004). Interferon antagonist proteins of influenza and vaccinia viruses are suppressors of RNA silencing. Proc Natl Acad Sci U S A 101, 1350-1355.

Li, F., y Ding, S. W. (2006). Virus counterdefense: diverse strategies for evading the RNA-silencing immunity. Annu. Rev. Microbiol 60, 503-531.

Li TC, Scotti PD, Miyamura T, Takeda N. Latent infection of a new alphanodavirus in an insect cell line. J Virol. Octubre de 2007; 81(20):10890-6.

Myles KM, Wiley MR, Morazzani EM, Adelman ZN. Alphavirus-derived small RNAs modulate pathogenesis in disease vector mosquitoes. Proc Natl Acad Sci USA. 16 de diciembre de 2008; 105(50):19938-43. Publicación electrónica del 1 de diciembre de 2008.

Mlotshwa, S., Pruss, G. J., y Vance, V. (2008). Small RNAs in viral infection and host defense. Trends Plant Sci 13, 375-382.

Muller H, Islam MR, Raue R. Research on infectious bursal disease--the past, the present and the future. Vet Microbiol. 2 de diciembre de 2003; 97(1-2):153-65.

Poulos BT, Tang KF, Pantoja CR, Bonami JR, Lightner DV. Purification and characterization of infectious myonecrosis virus of penaeid shrimp. J Gen Virol. Abril de 2006; 87(Pt 4):987-96.

Sanchez-Vargas, I., Scott, J. C., Poole-Smith, B. K., Franz, A. W., Barbosa-Solomieu, V., Warren, R.L., Sutton, G.G., Jones, S.J.M., Holt, R.A. (2007) Assembling millions of short DNA sequences using SSAKE. Bioinformatics 4:500-501.

Wang XH, Aliyari R, Li WX, Li HW, Kim K, Carthew R, Atkinson P, Ding SW. RNA interference directs innate immunity against viruses in adult *Drosophila*. Science. 21 de abril de 2006; 312(5772):452-4.

Wilusz, J., Olson, K. E., y Blair, C. D. (2009). Dengue virus type 2 infections of *Aedes acgypti* are modulated by the mosquito's RNA interference pathway. PLoS Pathog 5, e1000299.

Zambon, R. A., Vakharia, V. N., y Wu, L. P. (2006). RNAi is an antiviral immune response against a dsRNA virus in *Drosophila melanogaster*. Cell Microbial 8, 880-889.

Zerbino, D.R., Birney, E. (2008) Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 18:821-829.

EJEMPLO 2: Descubrimiento de nuevos virus y genomas virales

Los métodos para el ensamblaje de genoma viral y el descubrimiento viral pueden usar ARN inhibidores pequeños, o ARN "de silenciamiento pequeños", (ARNip), microARN (miARN) y/o ARN que interaccionan con PIWI (ARNpi), incluyendo ARNip, miARN y/o ARNpi aislados o secuenciados a partir de organismos invertebrados tales como insectos (*Anthropoda*), nematodos (*Nemapoda*), *Mollusca*, *Porifera* y otros invertebrados, y/o plantas, hongos o algas, cianobacterias y similares.

Aunque la invención no está limitada por ningún mecanismo de acción o proceso natural particular, la invención se basa en la respuesta de plantas e invertebrados a una infección. Los invertebrados procesan genomas de ARN viral en replicación para dar ARN de interferencia pequeños (ARNip) de tamaños diferenciados para guiar el aclaramiento de virus mediante interferencia de ARN. En el presente documento se muestra que ARNip virales secuenciados a partir de células de mosca de la fruta, mosquito y nematodo eran todos de secuencia solapante, y los métodos de la invención usan estos ARNip para el ensamblaje de genoma viral y el descubrimiento de virus.

Para demostrar cómo funcionan las realizaciones de la invención, se examinaron cóntigos ensamblados de bibliotecas de ARN pequeños publicadas y se descubrieron cinco nuevos virus de células de *Drosophila* cultivadas y mosquitos adultos, incluyendo tres con un genoma de ARN de hebra positiva y dos con un genoma de ARNbc. Particularmente, cuatro de los virus identificados presentaron sólo escasas similitudes de secuencia con respecto a virus conocidos de manera que no pudo asignarse ninguno a un género de virus existente. También se describe la detección de los primeros ARN que interaccionan con PIWI (ARNpi) derivados de virus en *Drosophila* y se demuestra el ensamblaje de genoma viral a partir de ARNpi virales en ausencia de ARNip virales.

Por tanto, esta invención proporciona un enfoque independiente de cultivo potente para el descubrimiento de virus en invertebrados mediante el ensamblaje de genomas virales directamente a partir de productos de la respuesta inmunitaria del huésped sin enriquecimiento o amplificación de virus anterior. Los virus, incluyendo virus de plantas, hongos, algas y/o cualquier invertebrado descubiertos mediante un método de esta invención pueden incluir nuevos patógenos virales de seres humanos, plantas y vertebrados que se transmiten por vectores de artrópodos y plantas.

Resultados:

Secuenciación del genoma de virus mediante el ensamblaje de ARNip virales producidos en huéspedes invertebrados. La ribonucleasa específica para ARNbc de tipo III Dicer escinde preferiblemente sustratos de ARNbc largos desde un extremo terminal de manera que los precursores de ARNbc con un extremo terminal definido se procesan para dar ARNip en fases de 21 nucleótidos (nt) (26-27). Sin embargo, la secuenciación de ARN pequeños mediante la plataforma 454 a partir de células de *Drosophila* infectadas de manera aguda con VFH mostró recientemente que los ARNip virales de 21 nt, dependientes de DCR2 no se producen en fase (11). Por tanto, se sometió a prueba la idea de que los ARNip virales producidos mediante el sistema inmunitario del huésped podían ser de secuencia solapante determinando si los fragmentos de ARNip de VFH secuenciados podían volver a ensamblarse en el genoma de ARN de VFH.

Se eligió el programa VELVET™ (28) desarrollado para el ensamblaje de genoma a partir de lecturas cortas y se fijaron 17 nucleótidos como la longitud de solapamiento mínima (*k-mero: 17*) requerida para unir dos ARN pequeños en un cóntigo. El ensamblaje de los 1177 ARNip de VFH secuenciados (11) mediante VELVET™ proporcionó tres cóntigos de 54, 73 y 52 nucleótidos de longitud, que contenían 27, 47, 35 ARNip, respectivamente (figura 5A). Esto indicaba que ARNip virales producidos en células de mosca de la fruta infectadas eran de hecho de secuencia solapante. Se agruparon los tres cóntigos ensamblados en la región 5'-terminal del ARN1 genómico de VFH (figura 5A), en el que se mapeó anteriormente más del 60% de los ARNip específicos para ARN1 (11).

El nematodo *Caenorhabditis elegans* que porta un ARN1 genómico autorreplicante de VFH produce ARNip virales detectables mediante hibridación de transferencia de tipo Northern (29). Se secuenciaron un total de 1.236.800 de ARN pequeños de 19-25 nucleótidos a partir de los nematodos mediante la plataforma Illumina. Además de 321.568 (26%) miARN de C. *elegans* conocidos, la biblioteca contenía 5.957 (0,48%) y 1455 (0,12%) de lecturas que eran idénticas al 100%/complementarias a y diferían en un nucleótido del genoma de VFH replicante, respectivamente. Se dividieron los ARNip virales igualmente en polaridades (+) y (-) y los ARNip virales más abundantes de ambas polaridades eran de 23 nucleótidos de longitud (figura 5), lo que concuerda con el tamaño determinado mediante la hibridación de transferencia de tipo Northern (29). Particularmente, el ensamblaje de los ARNip virales clonados a partir de animales C. *elegans* proporcionó 29 cóntigos que cubrían el 93% del ARN1 de VFH 36 veces (figura 5B). También había solapamiento con el cóntigo vecino para 21 de los 29 cóntigos de VFH y la carencia de ensamblaje adicional mediante Velvet se debió a que la longitud de solapamiento era más corta que el valor *k-mero* definido (17 nt).

La clonación y secuenciación de ARNip virales producidos en mosquitos (*Aedes aegypti*) infectados con el virus Sindbis (VSIN) portado por artrópodos se notificaron recientemente (22). La biblioteca contenía 525.457 ARN pequeños de VSIN de apareamiento perfecto y 68.669 ARN pequeños de VSIN con un apareamiento erróneo. Se encontró que el ensamblaje de los ARN pequeños de VSIN contenidos en la biblioteca generó 19 cóntigos de ARNip con sólo 5 huecos verdaderos y que el 99% del genoma de 10 kb de VSIN se cubrió 1029 veces (véase la figura 5C).

Estos hallazgos ilustran que la secuenciación de gran volumen de ARN pequeños totales de huéspedes infectados proporcionó ensamblajes de ARNip viral para cubrir casi los genomas virales completos múltiples veces. Por tanto, se concluyó que ARNip virales producidos por los tres huéspedes invertebrados son de secuencia solapante y podrían usarse para el ensamblaje de genoma viral. Puesto que la producción de ARNip virales es una respuesta

inmunitaria inevitable de muchos huéspedes eucariotas frente a la infección por virus (1-3), los métodos de la invención pueden usarse para descubrir nuevos virus mediante secuenciación masiva y ensamblaje de ARN pequeños totales acumulados en un organismo de interés.

Descubrimiento de cuatro nuevos virus ARN a partir de una línea celular S2 de *Drosophila*. Para demostrar que el método vdSAR de esta invención funcionaba, se analizaron dos bibliotecas de ARN pequeños por duplicado construidas a partir de una línea celular Schneider 2 (S2) de *Drosophila* mantenida anteriormente en el laboratorio de Gerald Rubin, denominada S2-GMR (12). Estas bibliotecas contienen 6.454.759 ARN pequeños de 18 a 28 nucleótidos de longitud en total, de los que 1.092.833 moléculas son únicas. En primer lugar se ejecutó el programa Velvet de ensamblaje de ARN pequeño usando k-mero de 17 y se obtuvieron 1639 cóntigos en total. BLASTN identificó tres grupos de cóntigos que eran idénticos o altamente homólogos con respecto a las entradas de secuencia de nucleótidos de las bases de datos no redundantes de NCBI. Se mapeó el primer grupo de 1032 cóntigos en el genoma de *D. melanogaster,* y el 62% de esos cóntigos solaparon loci de transposón, lo que sugiere que ARNip endógenos de la mosca de la fruta también son de secuencia solapante.

Se encontró que el segundo grupo de 49 cóntigos (de 33 a 401 nt de longitud) correspondía al genoma de ARNbc bipartito de virus X de *Drosophila* (VXD). Se obtuvieron el 78% del segmento A del genoma de VXD (23 cóntigos) y el 91% del segmento B (26 cóntigos) a partir de los ensamblajes de ARN pequeño, proporcionando una secuencia de ARN consenso de 5,55 kb en total con 73 veces de cobertura. Puesto que el genoma viral ensamblado a partir de los ARN pequeños secuenciados mostró una identidad del 98% con respecto a VXD (NC_004177, NC_004169) a nivel de la secuencia tanto del nucleótido como de la proteína, se concluyó que la línea celular S2-GMR estaba infectada de manera persistente con VXD.

El tercer grupo incluyó 57 cóntigos que presentaban una fuerte homología con respecto al genoma ARN+ bipartito del virus de la línea celular Tn5 (VLCT). VLCT es un miembro descrito recientemente del género *Alphanodavirus* y comparte un identidad de secuencia de nucleótidos del 89,3% y el 84% con ARN1 y ARN2 de VFH en el mismo género (30). El ensamblaje de ARNip proporcionó 2.196 y 1.048 nucleótidos de longitud para ARN1 y ARN2, respectivamente. Las partes restantes del genoma bipartito se obtuvieron mediante RT-PCR y RACE-PCR a partir de células S2-GMR proporcionadas por el laboratorio de Lai, produciendo las moléculas de ARN1 y ARN2 completas de 3107 y 1416 nucleótidos de longitud. El virus identificado, designado nodavirus americano (NVA), era el que se relaciona de la manera más estrecha con VLCT, con identidades del 94% y del 92% con respecto a ARN1 y ARN2 de VLCT. NVA también compartía el 89% y el 82% con respecto a ARN1 y ARN2 de VFH, lo que explicaba por qué se pensaba anteriormente que las células estaban infectadas de manera persistente por VFH (12). Además de la ARN polimerasa dependiente de ARN (RdRP) y la proteína de cobertura (CP), tanto VLCT como NVA codifican para el supresor de iARN (proteína B2) de 106 residuos de aminoácido (aa). Sin embargo, las tres proteínas virales presentaban niveles similares de variaciones de secuencia entre NVA, VLCT y VFH, lo que sugiere que NVA representa una nueva especie de *Alphanodavirus.*

Se identificaron tres agrupaciones adicionales de cóntigos específicos para virus entre los 501 cóntigos ensamblados restantes mediante comparación por BLASTX con las proteínas virales conocidas en NCBI (punto de corte: 1e-3). Ocho cóntigos en la primera agrupación (figura 6A) codificaron para proteínas con identidades del 34-62% frente a o bien RdRP (5 cóntigos de 1410 nucleótidos de longitud) o bien la proteína estructural (3 cóntigos de 899 nucleótidos de longitud) de virus de la mionecrosis infecciosa del camarón peneido (VMNICP) (31). VMNICP es un miembro no asignado en los *Totiviridae,* que incluye tres géneros establecidos de virus con un genoma de ARNbc lineal (32). Se obtuvo el genoma completo del virus identificado, designado totivirus de *Drosophila melanogaster* (TVD), a partir de células S2-GMR mediante RT-PCR usando cebadores designados según las secuencias de los ocho cóntigos y sus posiciones relativas mapeadas en el genoma de VMNICP (figura 6A), y mediante RACE-PCR. El genoma de TVD era de 6.780 nucleótidos de longitud y codificaba para ORF de CP y RdRP que se solapaban en 205 nucleótidos. Aunque las RdRP de TVD y VMNICP compartían identidades de sólo el 37,6%, el análisis filogenético de las RdRP virales en los *Totiviridae* mostró que TVD y VMNICP formaban una agrupación distinta fuera de los tres géneros conocidos (figura 6A). Por tanto se sugiere un nuevo género en los *Totiviridae* para que incluya TVD y VMNICP.

La segunda agrupación de ARNip también contenía 8 cóntigos (figura 6B) que codifican para proteínas con homología con diversos miembros de *Birnaviridae,* que contiene un genoma de ARNbc bipartito (33). Se mapearon cuatro de esos cóntigos con una longitud combinada de 1.224 nucleótidos en total en la región codificante de RdRP (VP1) mientras que se mapearon los cóntigos restantes de 888 nucleótidos de longitud con el segundo segmento del genoma birnaviral que codifica para las proteínas estructurales (figura 6B). Se recuperó el genoma bipartito completo del birnavirus identificado, designado birnavirus de *Drosophila melanogaster* (BVD), de células S2-GMR mediante PCR tal como se describió anteriormente y se clonó. El segmento A de BVD era de 3.258 nucleótidos de longitud y codificaba para una poliproteína homóloga con respecto a las proteínas estructurales birnavirales conocidas y una proteína solapante N-terminal, que sin embargo no presentaba similitudes con las proteínas solapantes N-terminales codificadas por diversos birnavirus (33). El segmento B era de 3.014 nucleótidos de longitud y codificaba para la RdRP viral (figura 6B). Análisis de secuencia y filogenéticos indican que BVD es claramente distinto de todos los birnavirus conocidos incluyendo VXD, el único birnavirus notificado aislado de un insecto huésped (33). Por ejemplo,

ni la RdRP predicha ni las proteínas estructurales de BVD comparten identidades de más del 31% con ningún miembro de los tres géneros birnavirales conocidos (figura 6B). Por tanto, se sugiere que BVD representa una especie y un género nuevos en *Birnaviridae.*

La última agrupación de ARNip contenía dos cóntigos (figura 6C) que codifican para proteínas homólogas con respecto a RdRP de virus de *Euprosterna elaeasa* (VEE) de *Tetraviridae,* cuyos miembros contienen un genoma ARN+. La longitud combinada de los dos cóntigos era de 892 nucleótidos. Los intentos repetidos de recuperar el genoma viral de la línea celular S2-GMR establecida en UC-Riverside mediante RT-PCR no fueron satisfactorios. Sin embargo, la inclusión de una biblioteca de ARN pequeños adicional (NCBI-GEO: GSM 272653) construida a partir de la línea celular Kc de *Drosophila* por el laboratorio de Lai (12) en el ensamblaje proporcionó un cóntigo contiguo largo de 3.005 nucleótidos de longitud. Este cóntigo largo contenía los dos cóntigos identificados inicialmente y 17 cóntigos adicionales en las bibliotecas de S2-GMR que no presentaban una homología detectable con respecto a proteínas virales conocidas (barras rojas y grises en la figura 6C). La secuencia consenso de ARNip ensamblado codificaba para una proteína de 984 residuos, que compartía identidades de aproximadamente el 29% con la RdRP tanto de VEE como de virus de *Thosea asigna* (VTA) en *Tetraviridae.* El análisis filogenético adicional de las RdRP en *Tetraviridae* (figura 6C) sugiere que el virus identificado, designado tetravirus de *Drosophila melanogaster* (TrVD), representa una nueva especie en *Tetraviridae.*

Estos resultados indican que las células S2-GMR usadas para la construcción de la biblioteca estaban infectadas de manera persistente con cinco virus ARN, que pertenecen a cuatro familias de virus diferentes. Tal como se esperaba, se mostró que la línea celular S2-GMR establecida posteriormente en UC-Riverside contenía de hecho VXD, NVA, TVD y BVD infecciosos, pero no TrVD, mediante inoculación de células S2 sanas seguido por análisis mediante RT-PCR (véase la figura 7A). Estos resultados explicaron que aunque se tuvo éxito en la obtención de las secuencias genómicas de longitud completa de NVA, TVD y BVD, se fracasó en los intentos repetidos de recuperar TrVD. 388.289 (6%) de las 6.454.759 lecturas totales de las células S2-GMR y 220 de los 1639 cóntigos ensamblados se mapearon en los cinco virus. La especie más predominante de ARN pequeños derivados de cualquiera de los tres virus ARNbc (TVD, BVD y VXD) o los dos virus ARN+ (NVA y TrVD) era de 21 nucleótidos (figura 7B), y las razones de ARNip virales (+) y (-) de 21 nt eran aproximadamente iguales (figura 7B). Estas características de ARN pequeños derivados de virus eran similares a las de ARNip derivados de VFH producidos mediante DCR-2 (11), lo que sugiere una ruta de biogénesis compartida para ARNip virales que seleccionan como diana virus ARN+ y ARNbc en *D. melanogaster.*

Había diferencias importantes en la abundancia relativa de ARNip derivados de cada uno de los cinco virus, con el 56%, el 18,1%, el 17,1%, el 5,7% y el 3,4% del ARNip viral total asignado a NVA, TVD, BVD, VXD y TrVD, respectivamente. El análisis adicional indicó que la mayor densidad de ARNip que selecciona como diana NVA se debía lo más probablemente a la presencia de un ARN de interferencia defectuoso de 591 nt (ARN-ID) derivado de NVA. Se clonó el ARN-ID y se encontró que el 51% de los ARNip virales totales de las células S2-GMR se mapearon en los tres picos de ARNip de ARN2 de NVA, que correspondían precisamente a las regiones de ARN2 (nucleótidos 1-245, 515-712, 1250-1277 y 1297-1416) presentes en el ARN-ID (figura 7B y figura 7C). La hibridación de transferencia de tipo Northern (figura 7B) reveló que el ARN-ID se replicó a altos niveles en células S2-GMR (carril derecho), pero a un nivel mucho más bajo en células S2 sanas nuevas inoculadas con el sobrenadante de las células S2-GMR (carril izquierdo), lo que indica que los altos niveles de replicación de ARN-ID pueden ser una característica clave de la infección viral mixta en las células S2-GMR. Además de los picos de ARNip derivados de ARN-ID en ARN2 de NVA, la distribución de ARNip virales tampoco era uniforme a lo largo de los ARN genómicos virales restantes (figura 4), tal como se indicó anteriormente (16-20). Sin embargo, los análisis indicaron que las regiones de densidad de ARNip alta de los genomas de ARN viral no estaban asociadas ni con contenido en AU no habitual (53%) ni con estructuras secundarias fuertes.

Ensamblaje de ARNip y descubrimiento de virus en mosquitos y *C. elegans.* A continuación se demostró que el método vdSAR de esta invención también funcionaba en otros invertebrados. La biblioteca de ARN pequeños de mosquito notificada por Miles y colegas contenía 3.771.297 lecturas de 18-26 nucleótidos de longitud, lo que representa 756.219 secuencias únicas (22). Excepto para los 19 cóntigos mapeados en el genoma viral Sindbis, no se identificaron cóntigos específicos para virus adicionales mediante BLASTN. Sin embargo, las búsquedas de BLASTX de los 435 cóntigos restantes de ARN pequeños identificaron dos cóntigos (figura 6D), que codificaban para proteínas que presentan similitudes del 54% y del 72% con respecto al precursor CP de nodavirus de Wuhan (NVW). El NVW, un virus de insecto identificado recientemente, es un miembro no asignado de *Nodaviridae* (34-35). La longitud combinada de los dos cóntigos era de 1103 nucleótidos y la proteína codificada cubría el 83% de, y compartía una identidad del 41,6% con, el precursor CP de NVW. Por tanto, el virus identificado puede representar un nuevo virus, designado como nodavirus de mosquito (NVM). El análisis filogenético de CP nodavirales indica que NVM no pertenece a ninguno de los géneros establecidos de *Nodaviridae* (figura 6D).

Los ARN pequeños totales se secuenciaron a partir de la cepa de *C. elegans* N2, se ensamblaron en 117 cóntigos en total. Sin embargo, excepto para los 29 cóntigos específicos para VFH, no se identificaron cóntigos específicos para virus adicionales mediante o bien BLASTN o bien BLASTX. De manera similar, ninguno de los 172 cóntigos ensamblados de una biblioteca grande de 10.964.021 ARN pequeños construidos a partir de estadios mixtos de *C.*

elegans (36) presentó similitudes detectables frente a virus conocidos. Esto sugiere que la cepa de laboratorio común de *C. elegans* puede no estar infectada de manera persistente con un virus ARN de homología suficiente detectable mediante vdSAR.

Detección y ensamblaje de ARNpi derivados de virus en células de la hoja somática de ovario de *Drosophila*. Se llevó a cabo adicionalmente el ensamblaje de los ARN pequeños secuenciados recientemente a partir de una línea celular de hoja somática de ovario (OSS) de *Drosophila* (37). A diferencia de las células S2 aisladas originalmente a partir de estadios embrionarios tardíos que no expresan ninguno de los tres miembros de la subfamilia PIWI, las células OSS producen abundantes ARNpi primarios de 24-30 nucleótidos además de ARNip y miARN debido a la expresión de la proteína PIWI (37-39). Las búsquedas de BLASTN de los cóntigos ensamblados identificaron fácilmente seis virus ARN en las células OSS. Éstos incluyen VXD, NVA, BVD y TrVD, detectándose todos ellos también en células S2-GMR, así como virus C de *Drosophila* (VCD) y noravirus. VCD y noravirus pertenecen a diferentes familias de virus ARN+ y ambos comparten similitudes con picornavirus. Una fuente común de contaminación por virus para las dos líneas celulares pueden ser extractos de moscas infectadas usadas en cultivo celular (37). Las búsquedas de BLASTX de los cóntigos ensamblados restantes no identificaron virus adicionales. El 3,3% de las 36.389.371 lecturas totales de las células OSS se mapearon en los seis virus. Entre los 1.184.811 ARNip virales en total, el 31,4%, el 26,9%, el 17%, el 13,5%, el 7,1% y el 4% procedían de VCD, Nora, VXD, BVD, VFH y TrVD respectivamente. Por tanto, NVA no era la diana predominante para el corte en trozos en las células OSS y por consiguiente, el mapeo de los ARNip en ARN genómicos individuales no identificó los tres picos de ARNip que corresponden a las regiones específicas para el ARN-ID de ARN2 detectado en las células S2-GMR.

Particularmente, se encontró una nueva población de ARN pequeños derivados de virus en las células OSS (figura 8A) que no se detectó en las células S2-GMR. Se sugiere que representan ARNpi derivados de virus debido a las siguientes tres características compartidas con los ARNpi primarios endógenos detectados en células OSS (37-39). En primer lugar, estos ARNpi virales eran de 24 a 30 nucleótidos de longitud con dos picos a 27 y 28 nucleótidos (figura 8A). En segundo lugar, los ARNpi virales presentaban un fuerte sesgo de uridina 5' (aproximadamente el 63%) pero no preferencia por adenina en la décima posición (figura 8B) y por tanto eran distintos de ARNpi secundarios de ovario de *Drosophila* cargados en AGO3, el 73% de los cuales tienen adenina en la 10ª posición (38). En tercer lugar, los ARNpi virales eran casi exclusivamente (95%) de una polaridad (figura 8A). En comparación, los ARNip virales eran más cortos que los ARNpi virales y no presentaban sesgo de hebra o preferencia por un nucleótido particular en ninguna posición. Además, la abundancia relativa de ARNpi virales era altamente variable entre los seis virus ARN que infectan de manera persistente las células OSS. Los ARNpi virales que seleccionan como diana NVA y VCD eran mucho más abundantes que los que seleccionan como diana los cuatro virus restantes. Sorprendentemente, ARNpi específicos para NVA eran más de dos veces más abundantes que ARNip virales en las células OSS. Sin embargo, ARNpi de los seis virus estaban claramente sesgados para lecturas de sentido, correspondiendo a o bien el ARN genómico de virus ARN+ (NVA, VCD, TrVD y noravirus) o bien a la hebra sentido de ARNm de los virus ARNbc (VXD y BVD), y estos ARNp virales presentaban sesgos de U 5' en lecturas sólo sentido pero no antisentido (figura 8B).

A continuación se determinó si estos seis virus podían identificarse mediante el ensamblaje de ARNpi virales en ausencia de ARNip virales. Para este fin, se clasificaron 19.334.507 lecturas (3.298.838 secuencias únicas) de desde 25 hasta 30 nucleótidos de las bibliotecas de células OSS. Se encontró que los seis virus se identificaron mediante el ensamblaje de estas lecturas de ARNpi libres de ARNip seguidas por BLASTN, independientemente de su abundancia relativa de ARNpi. Se mapearon 28 cóntigos en NVA, que cubrían el 94% y el 99% de ARN1 y ARN2, respectivamente. El 92% del genoma de VCD estaba representado por 68 cóntigos ensamblados. Para los cuatro virus restantes que produjeron menos ARNpi, se identificaron un total de 205 cóntigos de ARNpi específico de virus, que cubrían del 83% al 95% de o bien los genomas completos de BVD, VXD y noravirus o bien el genoma parcial de TrVD (figura 8). Sin embargo, las búsquedas de BLASTX de los cóntigos ensamblados restantes no identificaron virus adicionales incluyendo TVD, lo que concuerda con los resultados del ensamblaje de ARNip virales.

Discusión

En este estudio, se describe una realización de esta invención, un enfoque de vdSAR para el descubrimiento de virus en invertebrados (y plantas, algas y similares) mediante secuenciación masiva y ensamblaje de ARN de silenciamiento pequeños virales producidos por la maquinaria inmunitaria del huésped en respuesta a infección. El vdSAR se basó en la observación de que ARN de silenciamiento pequeños virales producidos por células de mosca de la fruta, mosquito y nematodo eran todos de secuencia solapante. En esta realización, se aislaron ARN pequeños totales a partir de un huésped, se secuenciaron en un carril de Illumina individual y se ensamblaron en cóntigos mediante Velvet. Se identificaron cóntigos específicos para virus mediante la búsqueda de las entradas de secuencia de nucleótidos no redundantes de NCBI tanto antes (BLASTN) como después de la traducción *in silico* (BLASTX), y los genomas completos de los virus identificados pudieron recuperarse posteriormente mediante PCR y se clonaron. El uso de vdSAR reveló infección mixta persistente de líneas celulares S2-GMR y OSS de *Drosophila* por cinco y seis virus ARN, respectivamente. También se ensamblaron los cóntigos de ARNip viral y se identificaron a partir de células S2 de *Drosophila* infectadas de manera aguda y mosquitos adultos (figura 5A y figura 5C). Sin embargo, no se identificó ningún virus mediante vdSAR de la cepa de laboratorio N2 de *C. elegans.* Por tanto, puede

ser necesario examinar aislados de campo (40) para el descubrimiento de virus en *C. elegans* puesto que el mantenimiento en laboratorio de cepas de gusano a menudo implica múltiples rondas de tratamiento con lejía para comenzar los cultivos a partir de huevos eliminando las larvas y animales adultos y la contaminación microbiana asociada.

Cinco de los virus ensamblados a partir de ARN pequeños de mosca y mosquito eran nuevos e incluían tres virus ARN+ y dos virus ARNbc. Excepto para NVA, BVD, TVD, TrVD y NVM todos presentaban bajas identidades de secuencia (25-42%) con respecto a virus conocidos que eran detectables sólo en regiones cortas de las proteínas virales codificadas. Como resultado, ninguno de los cuatro virus pudo asignarse a un género de virus existente. Esto sugiere que vdSAR puede descubrir nuevos virus que están relacionados sólo de manera distante con virus conocidos. Debe señalarse que los virus descubiertos mediante vdSAR a partir de invertebrados pueden incluir aquellos patógenos virales de seres humanos y vertebrados que se transmiten por vectores de artrópodos. La identificación de dos nuevos virus ADN se ha notificado en plantas de boniato mediante un enfoque similar a vdSAR (véase la referencia 41), lo que indica que vdSAR funciona tanto en plantas como en invertebrados.

El análisis de las bibliotecas de ARN pequeños recientemente notificadas preparadas en células OSS de *Drosophila* identificó ARNpi derivados de virus. Este hallazgo sugiere por primera vez que los ARNpi pueden desempeñar un papel antiviral, además de su papel en la defensa del genoma frente a transposones (4-6). Sin embargo, es interesante observar que estos ARNpi virales también son de secuencia solapante y pueden usarse para el ensamblaje de genoma viral en ausencia de ARNip virales. Esto sugiere que es probable que vdSAR sea eficaz para huéspedes o tejidos de huéspedes que sólo pueden producir ARNpi virales.

El descubrimiento de nuevos virus de animales a menudo se ve impedido por dificultades en su amplificación en cultivo celular y/o carencia de su reactividad cruzada en ensayos de hibridación serológicos y de ácidos nucleicos frente a virus conocidos. Recientemente se han identificado muchos nuevos virus en muestras del entorno y clínicas usando enfoques metagenómicos, en los que las partículas virales en primer lugar se purifican parcialmente y las secuencias de ácido nucleico viral se amplifican al azar antes de subclonar y secuenciar (42-44). Tanto los enfoques metagenómicos como vdSAR son independientes de cultivo y pueden identificar virus que comparten sólo escasas similitudes de secuencia con los virus conocidos. En comparación, los métodos de esta invención no requieren ni la purificación de la partícula viral ni la amplificación de la secuencia de ácido nucleico viral. Además, en algunas realizaciones, los métodos de esta invención implican la secuenciación de la fracción de ARN pequeños de huésped y la extracción de datos de sólo aquellos ARN pequeños que pueden ensamblarse en cóntigos de manera que tanto la cantidad de secuenciación como la complejidad de datos se reducen en gran medida. Las realizaciones de los métodos de esta invención pueden ensamblar genomas virales a partir de los productos de una respuesta inmunitaria del huésped activa frente a infección. En algunas realizaciones, sólo los virus en replicación e infecciosos que inducen la respuesta inmunitaria pueden identificarse mediante vdSAR.

Dada la diversidad genética y estructural de los virus caracterizados, es posible que haya tipos novedosos de patógenos virales y subvirales que no presentan similitud con respecto a ninguno de los virus conocidos detectables mediante las herramientas bioinformáticas disponibles. Estos virus novedosos escaparían fácilmente a la detección mediante los enfoques metagenómicos dependientes de homología y vdSAR actuales. De hecho, varios cóntigos ensamblados de las células de *Drosophila* no presentan similitud detectable con respecto a entradas en las bases de datos de NCBI. A este respecto, las características únicas de vdSAR pueden facilitar el desarrollo de nuevas herramientas bioinformáticas para seleccionar cóntigos particulares para el descubrimiento de virus. Por ejemplo, densidades de ARN pequeños, patrones de distribución de ARN pequeños, y razones de hebra positiva/negativa de ARN pequeños en los cóntigos ensamblados que concuerdan con ARN de silenciamiento pequeños virales, pueden considerarse indicadores de cóntigos con un origen viral.

Materiales y métodos

Cultivo celular. El cultivo, la infección por virus de células S2 y la hibridación de transferencia de tipo Northern fueron tal como se describió (11). La línea celular S2-GMR se facilitó amablemente por Eric Lai. Se usó el sobrenadante de células S2-GMR establecido en UC-Riverside para la infección de células S2 sanas nuevas.

Secuenciación, ensamblaje y análisis de bibliotecas de ARN pequeños. Se construyó la biblioteca de ARN pequeños de *C. elegans* tal como se describió (45) y se secuenció mediante 2G ANALYZER™ de Illumina en las instalaciones centrales en el campus Genomics Institute para genómica. Se recuperaron otras bibliotecas de ARN pequeños a partir de la base de datos de GEO. Se descargaron la secuencia de genoma de *D. melanogaster* y el archivo de anotación de repetición de UCSC (http://genome.ucsc.edu/). Se descargaron las bases de datos de nr y nt de NCBI (actualizadas en enero de 2009). Se descargó Velvet de EBI (http:// www.ebi.ac.uk/-zerbino/velvet). Se realizó el mapeo de ARN pequeños y cóntigos ensamblados en genomas de mosca y virales mediante el programa BLASTN usando los parámetros convencionales en el ensamblaje de genoma (cóntigos o cóntigo viral es similitud >=90% y cobertura >= 90% de cóntigos). También se examinaron los cóntigos ensamblados para determinar la similitud de sus proteínas codificadas con respecto a las bases de datos usando el programa BLASTX. Se llevaron a cabo análisis de datos adicionales con las secuencias de comandos de Perl domésticas. Se llevaron a cabo los análisis de

computación usando las instalaciones centrales en el campus Genomics Institute para bioinformática.

RT-PCR, RACE-PCR y secuenciación. Se usaron transcripción inversa (RT) y PCR para rellenar los huecos entre cóntigos de ARNip usando cebadores diseñados según las secuencias consenso de los cóntigos específicos implicados y se mapearon sus posiciones relativas en el genoma viral estrechamente relacionado. Se secuenciaron los productos de RT-PCR directamente mediante secuenciación con ABI didesoxilo convencional. Se llevó a cabo 5' RACE siguiendo las instrucciones del fabricante (Invitrogen). Para 3'-RACE, se aisló el ARN total a partir de células de mosca mediante el protocolo Trizol, se desnaturalizó a 65ºC durante 5 min y se ligó a una secuencia de transferencia 3' preadenilada ppACACTCGGGCACCAAGGA (SEQ ID NO:1) (ligador 2, IDT Company, EE.UU.) con fragmento truncado de T4 ARN ligasa (New England Biolabs, EE.UU.) (46). Tras la precipitación con etanol, se sometieron a transcripción inversa los productos de ligación mediante SUPERSCRIPT III™ (Invitrogen, Carlsbad, CA), se amplificaron mediante PCR. Se clonaron los productos de 5'-RACE y 3' RACE en el vector pGEM-T easy (Promega, EE.UU.) antes de la secuenciación con ABI didesoxilo. Se usó el paquete Phred-Phrap-Consed para el ensamblaje del genoma del virus.

Análisis filogenético. Se usó el paquete Mega 4 para construir los árboles filogenéticos. Se realizó la alineación de proteínas con el método Clustal W, y se calculó el árbol filogenético usando el método de unión a vecino. Se evaluó la fiabilidad de cada rama con análisis de remuestreo (repetición de 1000 veces).

Leyendas de las figuras para el ejemplo 2:

La figura 5 ilustra la posición y distribución de cóntigos de ARNip de VFH y VSIN ensamblados a partir de ARN pequeños secuenciados a partir de (figura 5A) células S2 de *Drosophila* infectadas con el mutante de deleción B2 de VFH (11), (figura 5B) una cepa de C. *elegans* transgénica en el contexto de mutante defectuoso para iARN 1 (*rde-1*) que porta un replicón ARN1 de VFH en el que se reemplazó la secuencia codificante de B2 por la de GFP (29), y (figura 5C) mosquitos adultos infectados con VSIN (22). Obsérvese que la longitud de genomas de ARN no se dibujó a escala.

La figura 6 ilustra el descubrimiento de los virus ARNbc TVD (figura 6A) y BVD (figura 6B) y los virus ARN+ TrVD (figura 6C) y NVM (figura 6D) a partir de células S2-GMR mediante vdSAR. Las barras rojas se refieren a los cóntigos específicos para virus identificados inicialmente mediante el % de similitudes de secuencia de sus proteínas codificadas con respecto a una proteína viral en las bases de datos. Los cóntigos de TVD, TrVD y NVM mostraron las mayores similitudes con respecto a VMICP, VEE y NVW, respectivamente. Sin embargo, se identificaron cuatro miembros diferentes de *Birnaviridae* como los más cercanos a cóntigos de BVD: a - virus de la enfermedad de la bursitis infecciosa (VEBI); b - VXD; c - birnavirus marino (VMA); d - virus del pez cabeza de serpiente moteado (VPM). Las barras grises se refieren a los cóntigos que se ensamblaron a partir de ARN pequeños de células S2-GMR y se mapearon posteriormente en virus específicos tras obtenerse los genomas completos. Obsérvese que la longitud de genomas de ARN se dibujó a escala y los marcos de lectura abiertos codificados por el genoma parcial de TrVD (3005 nt) y NVM (1130 nt) estaban incompletos.

La figura 7 ilustra los cuatro virus ARN infecciosos contenidos en células S2-GMR. La figura 7A ilustra que TVD, BVD, VXD y NVA se detectaron todos mediante RT-PCR en células S2 no contaminadas 4 días tras la inoculación con el sobrenadante de las células S2-GMR. Se usaron células S2 sanas y las células S2-GMR como controles. Se esperaba que los cebadores usados para RT-PCR proporcionaran productos específicos de 1.087, 1.030, 865 y 1.212 pb de longitud a partir de TVD, BVD, VXD y NVA, respectivamente. La figura 7B ilustra la detección de un ARN-ID derivado de ARN2 de NVA en células S2-GMR (carril derecho) y en S2 tras la inoculación con el sobrenadante de células S2-GMR (carril izquierdo) mediante hibridaciones de transferencia de tipo Northern usando una sonda que reconoce los 120 nt del extremo 3'-terminal de ARN2. La figura 7C ilustra la estructura del ARN-ID clonado de NVA (parte superior) y el mapeo de los de ARNip de 21 nt de apareamiento perfecto secuenciados a partir de células S2-GMR en sus hebras positiva (azul) y negativa (roja) de ARN2 de NVA (ventanas de 20 nt) (parte inferior).

La figura 8 ilustra la distribución del tamaño (figura 8A) y la composición de nucleótidos agregada (figura 8B) de ARN pequeños derivados de virus en células OSS de *Drosophila*. Para cada genoma o segmento de genoma viral, en la figura 8A se mostró el % de ARN pequeños virales o bien sentido (barra roja) o bien antisentido (barra azul) de tamaños distintos con respecto a las lecturas totales de longitud de 18-31 nt con apareamiento perfecto. Se calculó el % de composiciones de nucleótidos agregadas para todas las lecturas virales de ARNip de 21 nt o ARNpi de 27 nt + 28 nt con los números totales de lecturas en cada tamaño mostrado entre paréntesis.

Bibliografía (ejemplo 2):

1. Aliyari R y Ding SW (2009) RNA-based viral immunity initiated by the Dicer family of host immune receptors. Immunol Rev 227:176-188.

2. Mlotshwa S, Pruss GJ, y Vance V (2008) Small RNAs in viral infection and host defense. Trends Plant Sci 13(7):375-382.

3. Ding SW y Voinnet O (2007) Antiviral immunity directed by small RNAs. Cell 130(3):413-426.

4. Ghildiyal M y Zamore PD (2009) Small silencing RNAs: an expanding universe. Nat Rev Genet 10(2):94-108.

5. Siomi MC, Saito K, y Siomi H (2008) How selfish retrotransposons are silenced in *Drosophila* germline and somatic cells. FEBS Lett 582(17):2473-2478.

6. Malone CD y Hannon GJ (2009) Small RNAs as guardians of the genome. Cell 136(4):656-668.

7. Galiana-Arnoux D, Dostert C, Schneemann A, Hoffmann JA, e Imler JL (2006) Essential function in vivo for Dicer-2 in host defense against RNA viruses in *drosophila*. Nat Immunol 7(6):590-597.

8. Wang XH, *et al*. (2006) RNA interference directs innate immunity against viruses in adult Drosophila. Science 312(5772):452-454.

9. van Rij RP, *et al*. (2006) The RNA silencing endonuclease Argonaute 2 mediates specific antiviral immunity in *Drosophila melanogaster*. Genes Dev 20(21):2985-2995.

10. Li HW, Li WX, y Ding SW (2002) Induction and suppression of RNA silencing by an animal virus. Science 296(5571):1319-1321.

11. Aliyari R, *et al*. (2008) Mechanism of induction and suppression of antiviral immunity directed by virus-derived small RNAs in *Drosophila*. Cell Host Microbe 4(4):387-397.

12. Flynt A, Liu N, Martin R, y Lai EC (2009) Dicing of viral replication intermediates during silencing of latent *Drosophila* viruses. Proc Natl Acad Sci USA. 106:5270-5

13. Zambon RA, Vakharia VN, y Wu LP (2006) RNAi is an antiviral immune response against a dsRNA virus in *Drosophila melanogaster*. Cell Microbiol 8(5):880-889.

14. Hamilton AJ y Baulcombe DC (1999) A species of small antisense RNA in posttranscriptional gene silencing in plants. Science 286(5441):950-952.

15. Vaucheret H (2008) Plant ARGONAUTES. Trends Plant Sci 13(7):350-358.

16. Molnar A, *et al*. (2005) Plant virus-derived small interfering RNAs originate predominantly from highly structured single-stranded viral RNAs. J Virol 79(12):7812-7818.

17. Ho T, Pallett D, Rusholme R, Dalmay T, y Wang H (2006) A simplified method for cloning of short interfering RNAs from *Brassica juncea* infected with Turnip mosaic potyvirus and Turnip crinkle carmovirus. J Virol Methods 136(1-2):217-223.

18. Qi X, Bao FS, y Xie Z (2009) Small RNA deep sequencing reveals role for *Arabidopsis thaliana* RNA-dependent RNA polymerases in viral siRNA biogenesis. PLoS ONE 4(3):e4971.

19. Donaire L, *et al*. (2009) Deep-sequencing of plant viral small RNAs reveals effective and widespread targeting of viral genomes. Virology 392(2):203-214.

20. Wang XB, *et al*. (2009) RNAi-mediated viral immunity requires amplification of virus-derived siRNAs in *Arabidopsis thaliana*. Proc Natl Acad Sci USA. En formato impreso.

21. Brackney DE, Beane JE, y Ebel GD (2009) RNAi targeting of West Nile virus in mosquito midguts promotes virus diversification. PLoS Pathog 5(7):e1000502.

22. Myles KM, Wiley MR, Morazzani EM, y Adelman ZN (2008) Alphavirus-derived small RNAs modulate pathogenesis in disease vector mosquitoes. Proc Natl Acad Sci USA 105(50):19938-19943.

23. Sanchez-Vargas I, *et al*. (2009) Dengue virus type 2 infections of *Aedes aegypti* are modulated by the mosquito's RNA interference pathway. PLoS Pathog 5(2):e1000299.

24. Segers GC, Zhang X, Deng F, Sun Q, y Nuss DL (2007) Evidence that RNA silencing functions as an antiviral defense mechanism in fungi. Proc Natl Acad Sci USA 104(31):12902-12906.

25. Zhang X, Segers GC, Sun Q, Deng F, y Nuss DL (2008) Characterization of hypovirus-derived small RNAs generated in the chestnut blight fungus by an inducible DCL-2-dependent pathway. J Virol 82(6):2613-2619.

26. Zhang H, Kolb FA, Brondani V, Billy E, y Filipowicz W (2002) Human Dicer preferentially cleaves dsRNAs at their termini without a requirement for ATP. EMBO J. 21(21):5875-5885.

27. Vagin VV, *et al*. (2006) A distinct small RNA pathway silences selfish genetic elements in the germline. Science 313(5785):320-324.

28. Zerbino DR y Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18(5):821-829.

29. Lu R, Yigit E, Li WX, y Ding SW (2009) An RIGI-Like RNA helicase mediates antiviral RNAi downstream of viral siRNA biogenesis in *Caenorhabditis elegans*. PLoS Pathog 5(2):e1000286.

30. Li TC, Scotti PD, Miyamura T, y Takeda N (2007) Latent infection of a new alphanodavirus in an insect cell line. J Virol 81(20):10890-10896.

31. Poulos BT, Tang KF, Pantoja CR, Bonami JR, y Lightner DV (2006) Purification and characterization of infectious myonecrosis virus of penaeid shrimp. J Gen Virol 87(Pt 4):987-996.

32. Hanizlik TN, *et al*. (2005) Totiviridae. Virus taxonomy - Eighth report of the international committee on taxonomy of viruses, editores Fauquet CM, Mayo MA, Maniloff J, Desselberger U, y Ball LA (Academic Press, San Diego), págs 873-883.

33. Delmas B, *et al*. (2005) Birnaviridae. Virus taxonomy - Eighth report of the international committee on taxonomy of viruses, editores Fauquet CM, Mayo MA, Maniloff J, Desselberger U, y Ball LA (Academic Press, San Diego), págs 561-569.

34. Liu C, *et al*. (2006) Isolation and RNA1 nucleotide sequence determination of a new insect nodavirus from *Pieris rapae* larvae in Wuhan city, China. Virus Res 120(1-2):28-35.

35. Liu C, *et al*. (2006) Sequence analysis of coat protein gene of Wuhan nodavirus isolated from insect. Virus Res 121(1):17-22.

36. Batista PJ, *et al*. (2008) PRG-1 and 21U-RNAs interact to form the piRNA complex required for fertility in *C. elegans*. Mol Cell 31(1):67-78.

37. Lau NC, *et al*. (2009) Abundant primary piRNAs, endo-siRNAs, and microRNAs in a *Drosophila* ovary cell line. Genome Res.

38. Brennecke J, *et al*. (2007) Discrete small RNA-generating loci as master regulators of transposon activity in *Drosophila*. Cell 128(6):1089-1103.

39. Saito K, *et al*. (2009) A regulatory circuit for piwi by the large Maf gene traffic jam in *Drosophila*. Nature 461(7268):1296-1299.

40. Troemel ER, Felix MA, Whiteman NK, Barriere A, y Ausubel FM (2008) Microsporidia are natural intracellular parasites of the nematode *Caenorhabditis elegans*. PLoS Biol 6(12):2736-2752.

41. Kreuze JF, *et al*. (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):1-7.

42. Culley AI, Lang AS, y Suttle CA (2006) Metagenomic analysis of coastal RNA virus communities. Science 312(5781): 1795-1798.

43. Victoria JG, Kapoor A, Dupuis K, Schnurr DP, y Delwart EL (2008) Rapid identification of known and new RNA viruses from animal tissues. PLoS Pathog 4(9):c1000163.

44. Cox-Foster DL, *et al*. (2007) A metagenomic survey of microbes in honey bee colony collapse disorder. Science

318(5848):283-287.

45. Mi S, *et al*. (2008) Sorting of small RNAs into *Arabidopsis* argonaute complexes is directed by the 5' terminal nucleotide. Cell 133(1):116-127.

46. Wu Q, *et al*. (2008) Poly A- transcripts expressed in HeLa cells. PLoS One 3(7):e2803.

Se han descrito varias realizaciones de la divulgación. Sin embargo, se entenderá que pueden realizarse diversas modificaciones sin apartarse del alcance de las reivindicaciones. Por consiguiente, otras realizaciones están dentro del alcance de las siguientes reivindicaciones.

Lista de secuencias

<110> Rectorado de la Universidad de California

<120> DESCUBRIMIENTO DE VIRUS MEDIANTE SECUENCIACIÓN Y ENSAMBLAJE DE ARNip, miARN, ARNpi DERIVADOS DE VIRUS

<130> 00013-030WO1/UCR-2009-660-PCT

<140> aún no asignado

<141> 01-06-2010

<150> documento 61/183.377

<151> 02-06-2009

<150> documento 61/286.742

<151> 15-12-2009

<160> 1

<170> PatentIn versión 3.5

<210> 1

<211> 18

<212> ADN

<213> artificial

<220>

<223> secuencia de transferencia 3'

<400> 1

acactcgggc accaagga 18

REIVINDICACIONES

1. Método para descubrir un genoma microbiano que comprende:

(a)

(i) obtener una pluralidad de ARN que interaccionan con PIWI que se producen de manera natural, para generar una biblioteca de ARN, u obtener una pluralidad de ARN que interaccionan con PIWI a partir de un organismo u organismos, o una planta o plantas; y

(ii) determinar la secuencia de los ARN que interaccionan con PIWI, y usar esas secuencias para ensamblar los ARN que interaccionan con PIWI en al menos un cóntigo que comprende una pluralidad de los ARN que interaccionan con el nucleótido PIWI; o

(b) el método de (a), en el que los cóntigos se ensamblan usando la ayuda de un programa informático.

2. Método según la reivindicación 1, que comprende además determinar la secuencia del cóntigo ensamblado.

3. Método según la reivindicación 1, que comprende además:

(a) buscar en una base de datos de secuencias virales o de microorganismo usando el al menos un cóntigo para identificar una secuencia, o subsecuencia de la misma, que codifica para proteína, ácido nucleico o genoma viral o de microorganismo, que tiene una homología significativa con respecto al cóntigo ensamblado; o

(b) el método de (a), en el que la base de datos comprende secuencias de nucleótidos no redundantes; o

(c) el método de (a), en el que la base de datos comprende secuencias de traducción *in silico*;

en el que opcionalmente el cóntigo ensamblado tiene una homología significativa con respecto a un género o genoma viral conocido.

4. Método según la reivindicación 3, en el que la secuencia, o subsecuencia de la misma, que codifica para proteína, ácido nucleico o genoma viral o de microorganismo, tiene un porcentaje de homología de al menos el 50% al 100% con respecto a la totalidad o parte del cóntigo ensamblado.

5. Método según la reivindicación 3 o la reivindicación 4, que comprende además:

(iv) realizar un análisis filogenético de la secuencia que codifica para proteína, ácido nucleico o genoma viral o de microorganismo identificada con el cóntigo.

6. Método según la reivindicación 5, que comprende además:

(v) identificar y anotar el análisis filogenético de la secuencia viral identificada con el cóntigo.

7. Método según la reivindicación 1, en el que las secuencias de ARN que interaccionan con PIWI obtenidas están sustancialmente purificadas o aisladas de un organismo de interés.

8. Método según la reivindicación 1, que comprende además purificar sustancialmente ARN que interaccionan con PIWI, a partir de un organismo de interés y secuenciar los fragmentos de ARN para obtener una biblioteca de ARN.

9. Método según cualquiera de las reivindicaciones 1 a 4, que comprende además eliminar segmentos secuenciados de la biblioteca que se solapan con la secuencia genómica del organismo de interés a partir de la que se derivó el ARN.

10. Método según cualquiera de las reivindicaciones 1 a 4, que comprende además rellenar los huecos entre los cóntigos.

11. Método según la reivindicación 10, en el que el rellenado de los huecos entre los cóntigos comprende el uso de RT-PCR y/o secuenciación para rellenar los huecos entre los cóntigos.

12. Método según cualquiera de las reivindicaciones 1 a 4, que comprende además completar una secuencia genómica de un virus o un microorganismo que comprende el cóntigo usando 5'-RACE y 3'-RACE.

13. Método según la reivindicación 1, en el que el organismo o los organismos es/son un invertebrado, un insecto (*Anthropoda*), un nematodo (*Nemapoda*), *Mollusca*, *Porifera*, una planta, hongos, algas, cianobacterias; o el organismo o los organismos se identifican o no se identifican y se derivan a partir de una muestra del entorno, en el que opcionalmente la muestra del entorno es una muestra de tierra, una muestra de agua o una muestra de aire.

14. Método para identificar un virus, que comprende:

construir una biblioteca de ARN pequeños a partir de un organismo u organismos;

secuenciar de manera masiva la biblioteca de ARN pequeños;

ensamblar los ARN pequeños secuenciados usando

(a) todos los ARN que interaccionan con PIWI; o

(b) ARN que interaccionan con PIWI, de una longitud definida en una pluralidad de cóntigos;

identificar y eliminar aquellas secuencias ensambladas mapeadas sobre el genoma del organismo para proporcionar un conjunto enriquecido de cóntigos;

realizar una búsqueda de homología de cóntigos frente a virus conocidos tanto a nivel de nucleótido como de proteína;

opcionalmente usar RT-PCR y secuenciar para rellenar los huecos entre los cóntigos que muestran similitudes limitadas con un virus conocido;

completar la secuencia genómica de longitud completa del virus identificado con 5'-RACE y 3'-RACE; y

anotar el virus identificado.

15. Método según la reivindicación 14, en el que el organismo o los organismos es/son un invertebrado, un insecto (*Anthropoda*), un nematodo (*Nemapoda*), *Mollusca*, *Porifera*, una planta, hongos, algas, cianobacterias; o el organismo o los organismos se identifican o no se identifican y se derivan a partir de una muestra del entorno, en el que opcionalmente la muestra del entorno es una muestra de tierra, una muestra de agua o una muestra de aire.

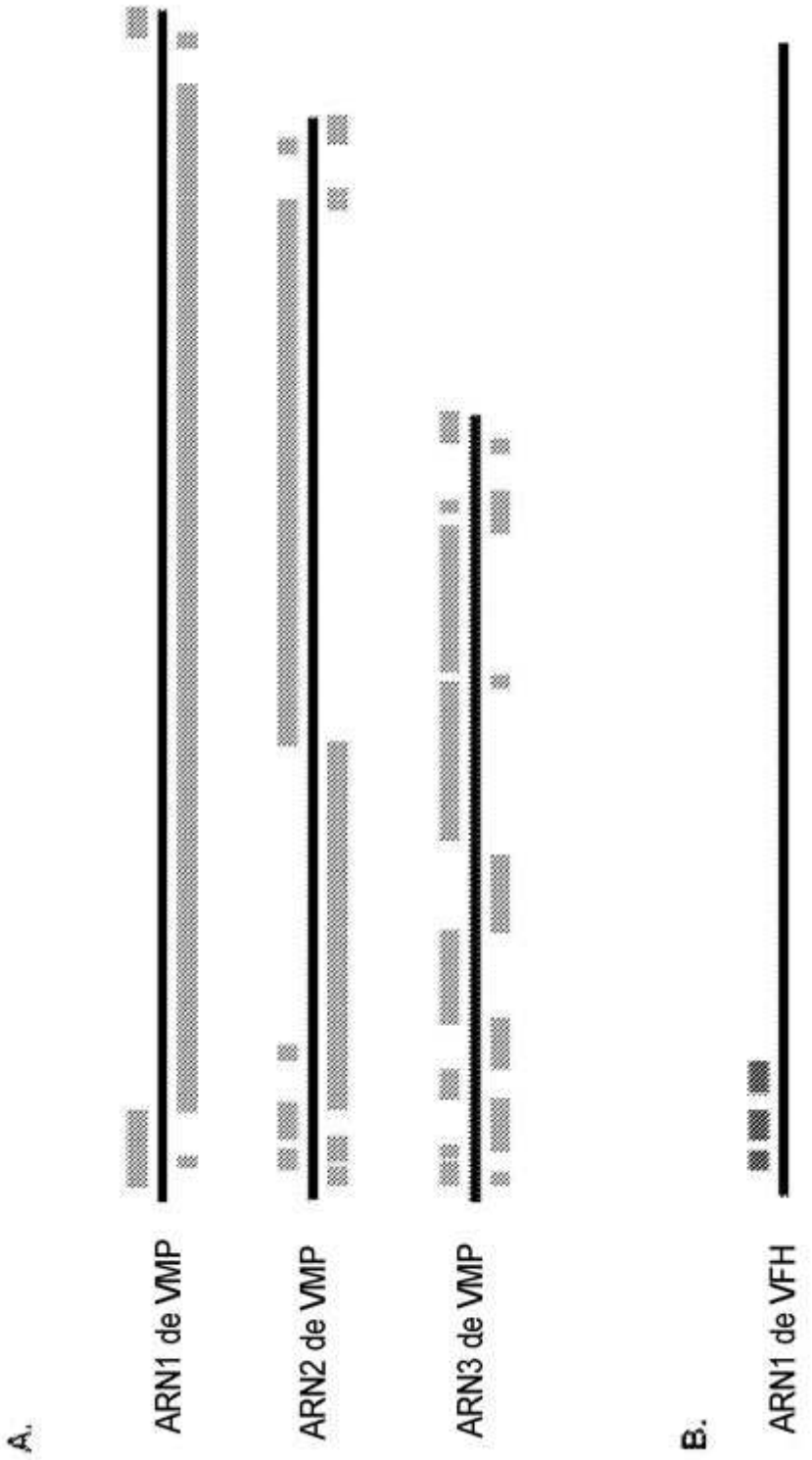
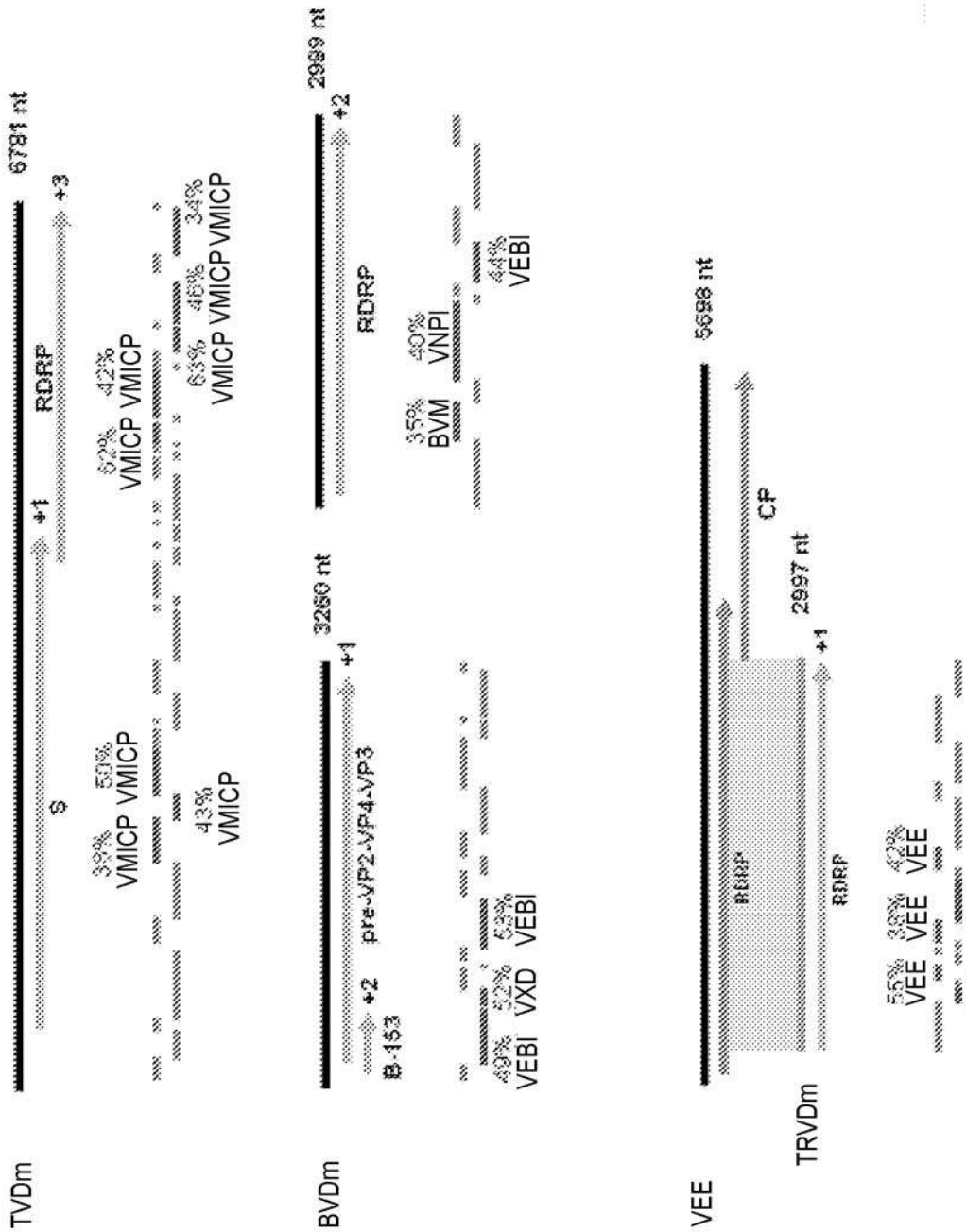


FIGURA 1A-B

FIGURA 2


TVDm
6781 nt
+1
+3
S
RDRP
39%
VMICP
50%
VMICP
62%
VMICP
42%
VMICP
43%
VMICP
65%
VMICP
48%
VMICP
34%
VMICP
BVDm
3260 nt
+1
+2
pre-vP2-vP4-vP3
B-153
RDRP
2999 nt
+2
35%
BVM
40%
VNPI
49%
VEBI
52%
VXD
53%
VEBI
44%
VEBI
VEE
5698 nt
RDRP
CP
TRVDm
2997 nt
+1
RDRP
55%
VEE
38%
VEE
42%
VEE
45%
VEE
51%
VEE

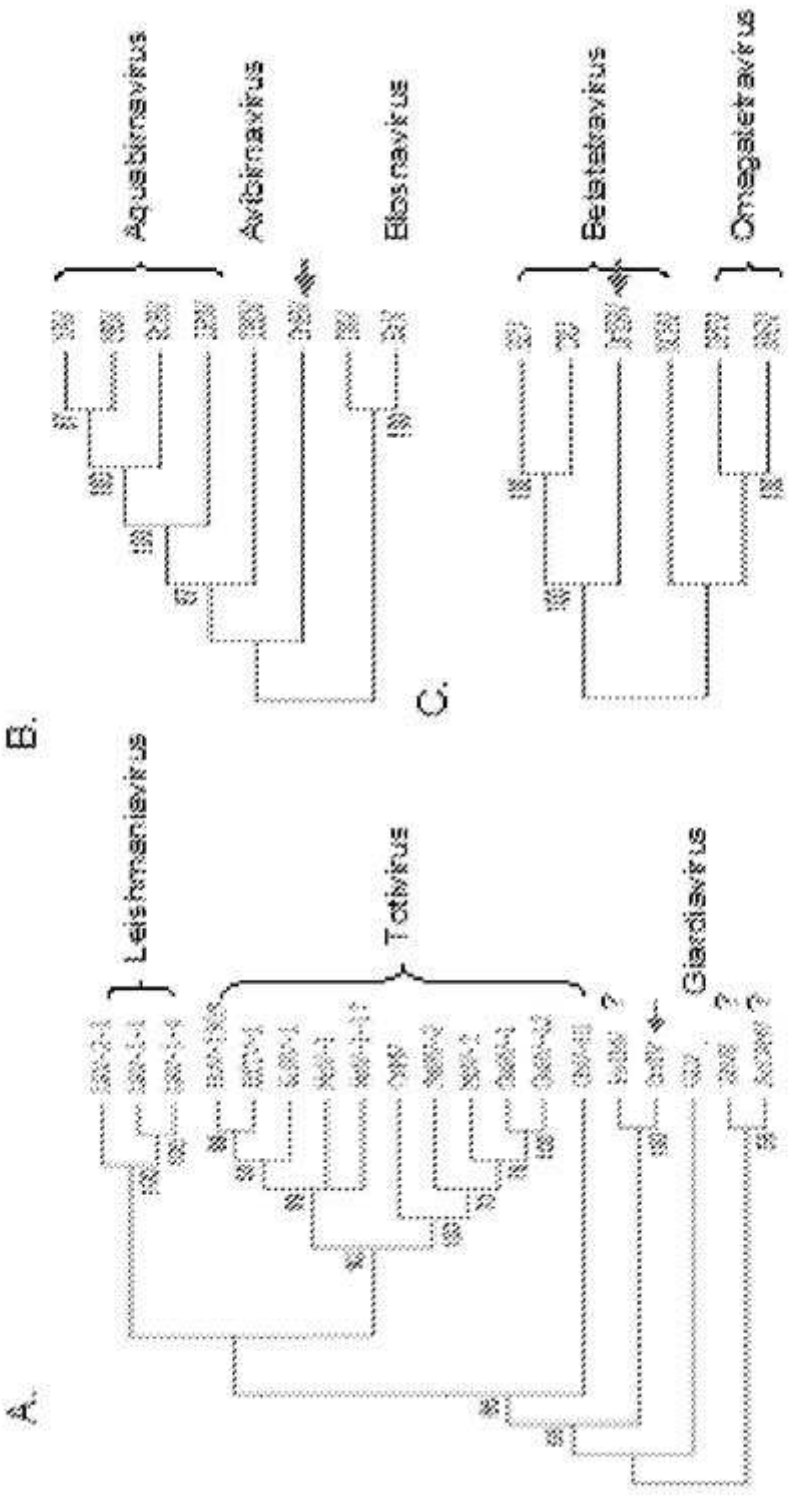

A.
B.
C.
Leishmaniavirus
Totivirus
Giardiavirus
Aquabirnavirus
Avibirnavirus
Ebirnavirus
Betatetravirus
Omegatetravirus

FIGURA 3A-C

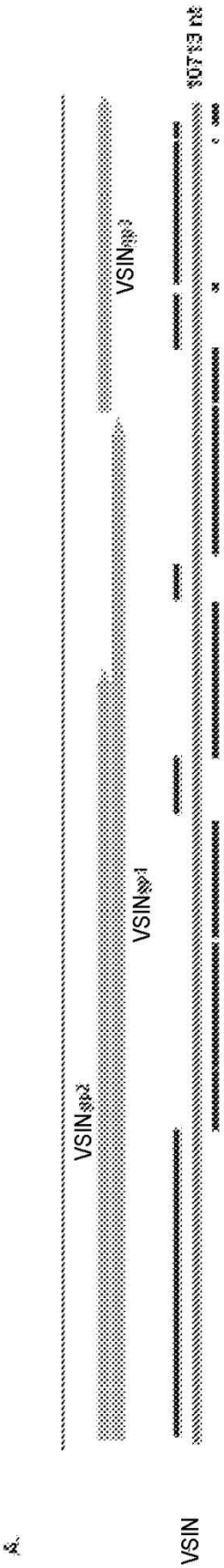


FIGURA 4A-B

Figura 5

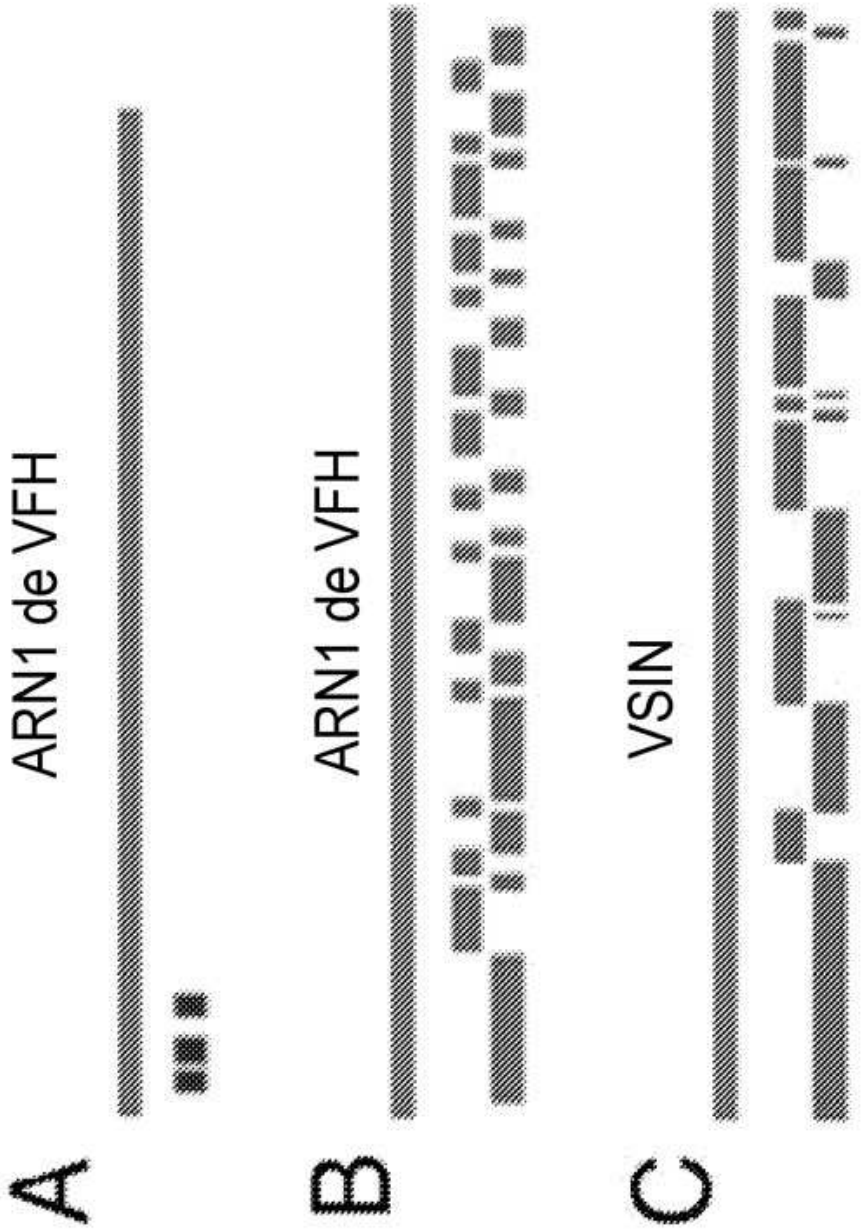


Figura 6

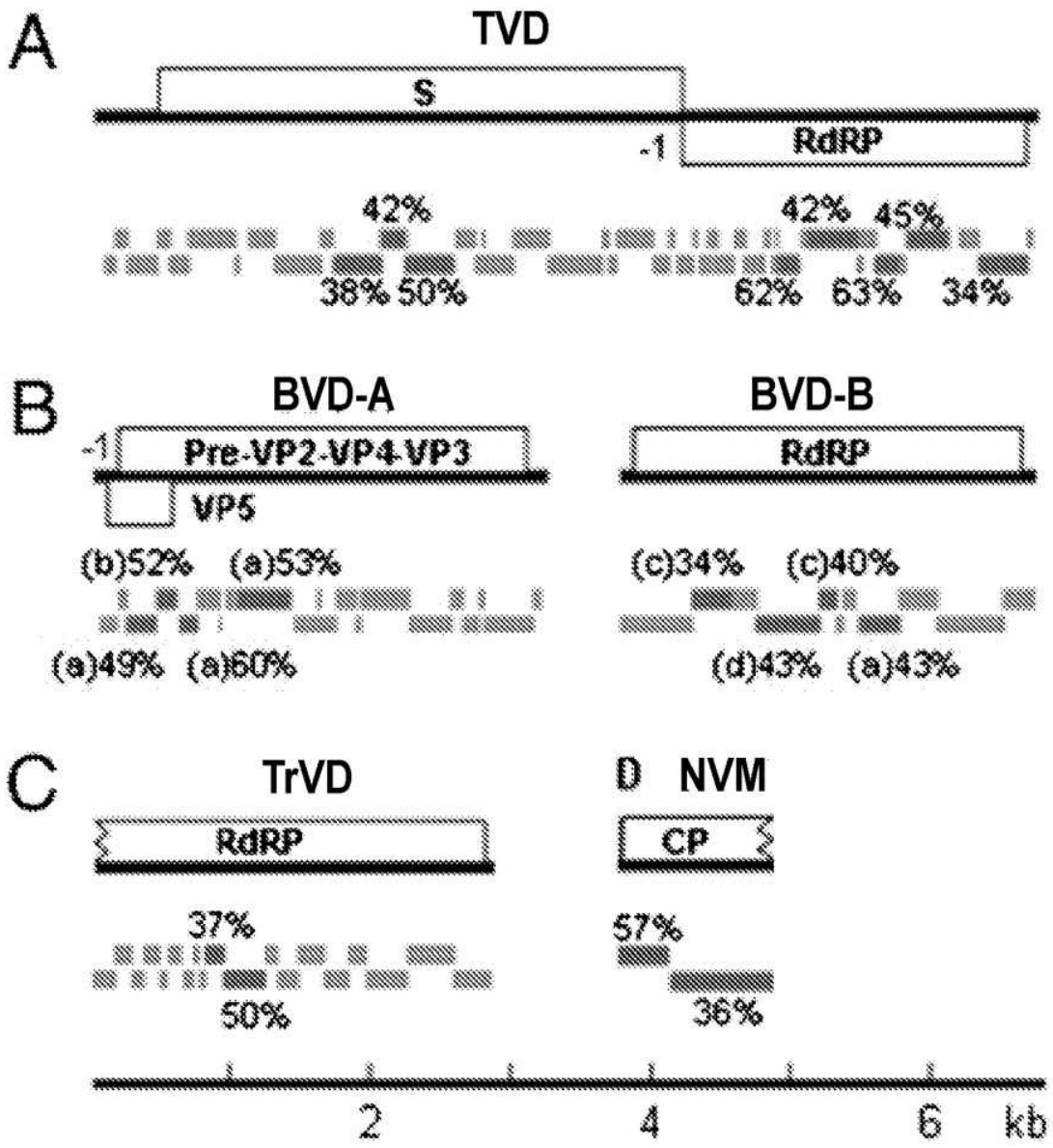

A
TVD
S
-1
RdRP
42%
42%
45%
38%
50%
62%
63%
34%
B
BVD-A
BVD-B
-1
Pre-VP2-VP4-VP3
RdRP
VP5
(b)52%
(a)53%
(c)34%
(c)40%
(a)49%
(a)60%
(d)43%
(a)43%
C
TrVD
D
NVM
RdRP
CP
37%
57%
50%
36%
2
4
6
kb

Figura 7

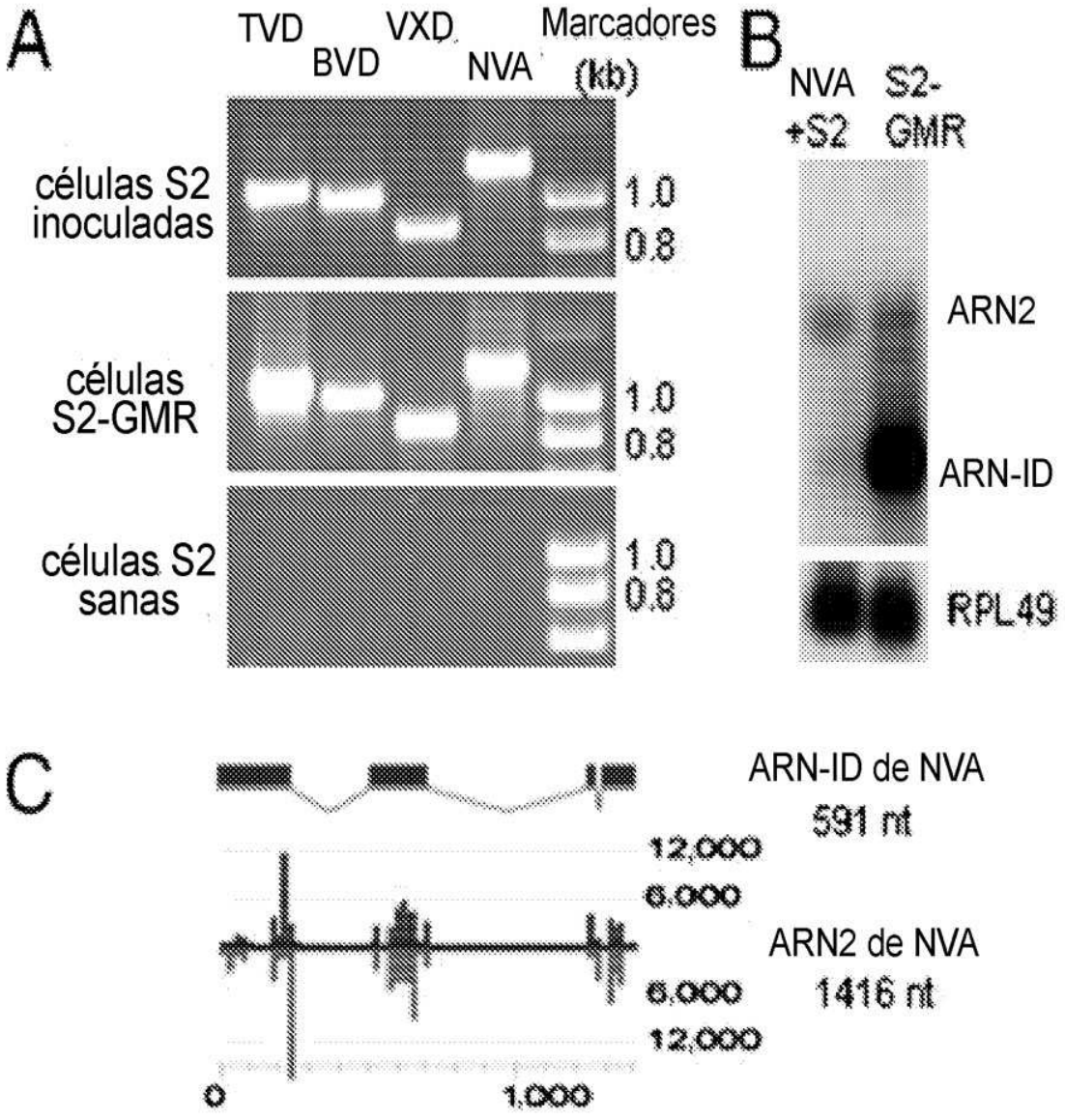
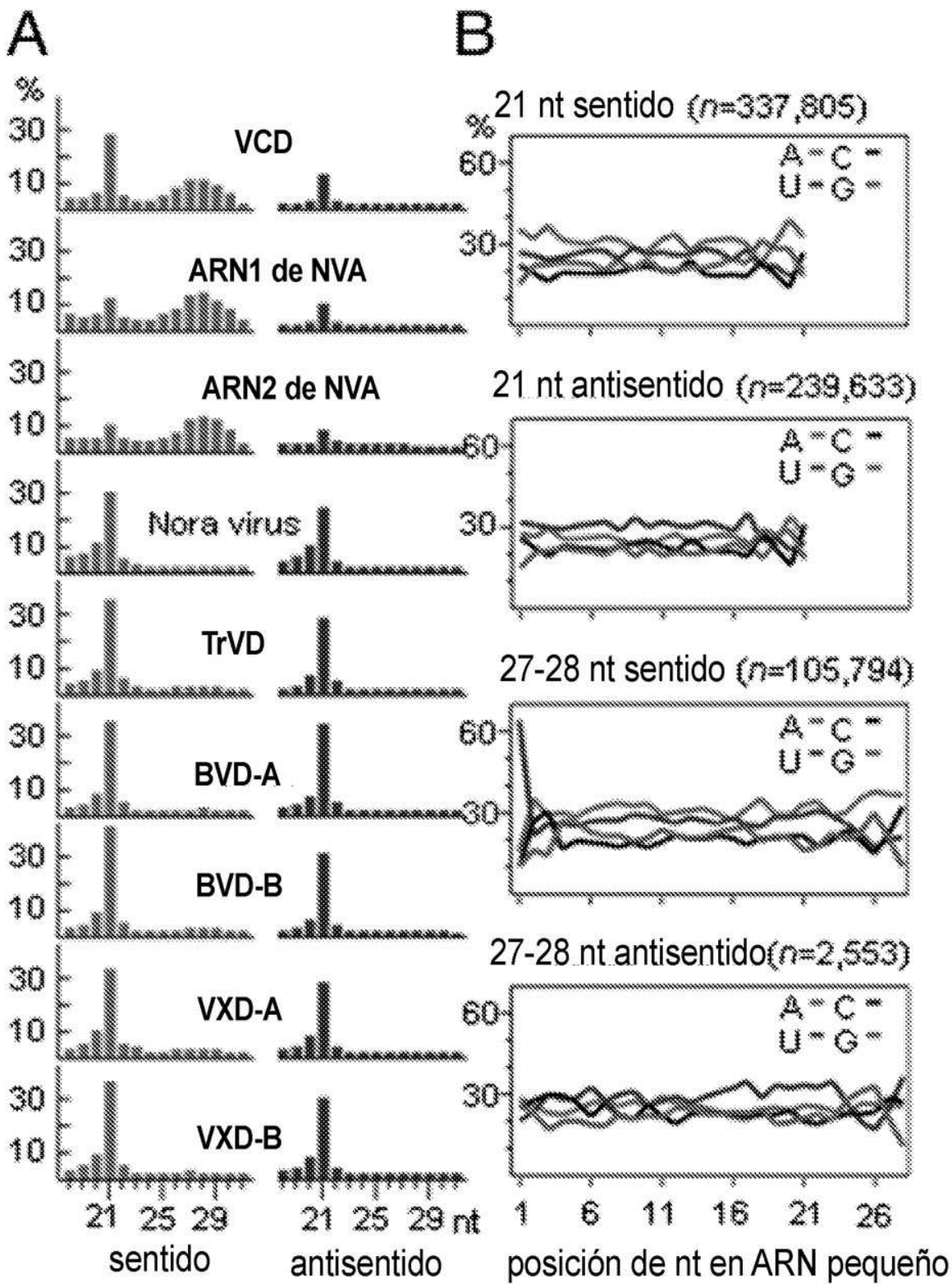

A
TVD
BVD
VXD
NVA
Marcadores (kb)
células S2 inoculadas
células S2-GMR
células S2 sanas
1.0
0.8
B
NVA +S2
S2-GMR
ARN2
ARN-ID
RPL49
C
ARN-ID de NVA
591 nt
12,000
6,000
ARN2 de NVA
1416 nt
6,000
12,000
0
1,000

Figura 8


A
B
%
30
10
VCD
ARN1 de NVA
ARN2 de NVA
Nora virus
TrVD
BVD-A
BVD-B
VXD-A
VXD-B
21 25 29
sentido
21 25 29 nt
antisentido
21 nt sentido (n=337,805)
21 nt antisentido (n=239,633)
27-28 nt sentido (n=105,794)
27-28 nt antisentido(n=2,553)
60
30
A
C
U
G
1 6 11 16 21 26
posición de nt en ARN pequeño