(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 550 799**

(51) Int. CI.:

***C12Q 1/68*** (2006.01)
***C12N 5/10*** (2006.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(96) Fecha de presentación y número de la solicitud europea: **27.07.2012 E 12753557 (3)**

(97) Fecha y número de publicación de la concesión europea: **09.09.2015 EP 2737086**

(54) Título: **Análisis de tendencia en secuenciación**

(30) Prioridad:

**29.07.2011 GB 201113214**

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
**12.11.2015**

(73) Titular/es:

**UNIVERSITY OF EAST ANGLIA (100.0%)**
**Norwich**
**Norfolk NR4 7TJ, GB**

(72) Inventor/es:

**SOREFAN, KARIM;**
**DALMAY, TAMAS;**
**MOULTON, VINCENT y**
**PAIS, HELIO ERNESTO CORONEL MACHADO**

(74) Agente/Representante:

**UNGRÍA LÓPEZ, Javier**

## DESCRIPCIÓN

Análisis de tendencia en secuenciación

**Campo de la invención**

La presente invención se refiere a métodos para determinar la tendencia de secuencia de una técnica de secuenciación. Además, la invención se refiere a métodos para reducir o aumentar la tendencia de secuencia durante la secuenciación de ácidos nucleicos por medio de técnicas que implican uniones a un adaptador. Específicamente el método se refiere al uso de una secuencia de ARN degenerado para analizar tendencias de secuencia cuando se generan bibliotecas de pequeños ARN, y al uso de adaptadores modificados para la clonación de pequeños ARN con secuencias degeneradas o específicas para reducir o aumentar la tendencia de la secuenciación, así como varias moléculas de ácidos nucleicos que se relacionan con estas o se derivan de las mismas.

**Antecedentes de la invención**

Muchos métodos biológicos moleculares necesitan una etapa de clonación que necesita el uso de una ADN o ARN ligasa para unir oligonucleótidos adaptadores u otras secuencias a una secuencia de nucleótido diana. La eficacia de esta unión depende de la secuencia del adaptador y la diana. Cuando se crea una biblioteca de secuencias a partir de ARN o ADN a menudo es importante unir todas las secuencias posibles y también que la biblioteca sea representativa de la abundancia relativa de la diana. Estas dos propiedades son importantes cuando se generan bibliotecas de alta calidad de pequeños ARN. Estas bibliotecas se pueden secuenciar utilizando la secuenciación tradicional de Sanger pero ahora se secuencian más comúnmente por secuenciación de alto rendimiento o técnicas de secuenciación de próxima generación (NGS).

La expresión genética en eucariotas está regulada en varias capas y uno de los mecanismos descubierto más recientemente implica moléculas de pequeños ARN no codificantes de 20-25 nucleótidos de longitud (sARN) (Fire, Xu et al. 1998; Voinnet 2002). Hay varias clases de sARN con diferentes rutas de biogénesis y modos de acción. La clase mejor caracterizada es la de los microARN (miARN) que se generan a partir de estructuras tallo-lazo y ARNm dianas *in trans*. La mayoría de los miARN se expresan específicamente en ciertos tejidos y en determinados estadios de desarrollo específicos y su acumulación a menudo cambia debido a señales externas y durante la enfermedad (MicroRNAs in Cancer Translational Research, William.C.S.Cho, 2011, Springer). Varias bases de datos han recopilado la asociación de los miARN con la enfermedad que incluyen la base de datos de microARN enfermedad humana y la mir2enfermedad (http://www.mir2disease.org/;http://202.38.126.151/ hmdd/ mirna/md/) y se han lanzado recientemente nuevos productos para clasificar cánceres (o enfermedades) que se basan en sus perfiles de miARN, tales como Mirview (Rosetta Genomics, documento US 2010/0273172).

Por lo tanto, el perfil preciso del nivel de los miARN (y otras clases de sARN) es muy importante en la investigación básica y clínica. La identificación de un miARN interesante para estudios adicionales es un proceso empírico que a menudo se basa en su alta expresión y clara expresión diferencial. Estos criterios son más importantes cuando falta el contexto biológico del miARN tal como en animales en los que la predicción de la diana es pobre (Dalmay 2008). Además, el nivel de expresión se utiliza a menudo para discernir entre el miARN y el miARN estrella. El perfil de miARN preciso se complica por la naturaleza heterogénea de los miARN tal como las isoformas de secuencias y la longitud de los isomiros ya que se cree que estos tienen actividades diferenciales (Fernandez-Valverde, Taft et al. 2010; Guo y Lu 2010; Starega-Roslan, Krol et al. 2011).

La medición de los microARN con tecnologías de secuenciación tales como de alto rendimiento y Sanger o por PCR cualitativa (QPCR) necesita el uso de enzimas de modificación de ácidos nucleicos. Las ligasas, transcriptasas inversas y ADN polimerasas, son las enzimas más importantes que se utilizan en biología molecular. Para mejorar la actividad de estas enzimas se necesita entender completamente su función, lo que necesita un método para medir su actividad e identificar los determinantes que regulan su función.

El documento WO 2009/088933 describe métodos de mutagénesis que se aplican en particular en la generación de bibliotecas o matrices de proteínas mutantes o ácidos nucleicos que codifican tales proteínas. Los métodos de mutagénesis descritos implican el uso de cebadores que comprenden al menos un codón degenerado de 2 a 12 veces.

El informe de Christian ("RNA Captor: A Tool for RNA Characterization", PLOS ONE, vol. 6, nº 4, Abril 2011) describe un método de “marcaje” de ARN basado en el uso de T3 ARN ligasa 2 y adaptadores. Se describe un protocolo de RACE-PCR que se basa en cebadores específicos del adaptador y específicos genéticos.

Perelle et al. ("Comparison of PCR-ELISA and LightCycler real-time PCR assays for detecting Salmonella spp. in milk and meat samples", Molecular and Cellular Probes, vol. 18, nº 6, Diciembre 2004, páginas 409-420) expone los resultados de una comparación entre dos técnicas de PCR específicas para detectar *Salmonella* spp. Los oligonucleótidos cebadores y las sondas que se utilizan en este estudio se muestran en la Tabla 1.

Tian Geng et al. ("Sequencing bias: comparison of different protocols of microRNA library construction" BMC Biotechnology, vol. 10, nº 1, Septiembre 2010, página 64) describe una comparación de tres protocolos diferentes de preparación de una biblioteca de ARN. El protocolo SOLID utiliza adaptadores de doble cadena con un segmento protuberante N6 aleatorio degenerado.

**Sumario de la invención**

La presente invención proporciona un método para reducir la tendencia de secuencia de una técnica de secuenciación que implica la unión de un adaptador, comprendiendo el método:

(a) proporcionar un grupo de oligonucleótidos de cadena sencilla de secuencia conocida con extremos 3' bloqueados (moléculas adaptadoras 3') y un grupo de oligonucleótidos de cadena sencilla de secuencia conocida con extremos 5' bloqueados (moléculas adaptadoras 5') donde las moléculas adaptadoras 3' y 5' comprenden uno o más nucleótidos degenerados;
(b) unir las moléculas adaptadoras 3' a los extremos 3' de las moléculas de ácido nucleico diana utilizando una ligasa;
(c) unir las moléculas adaptadoras 5' a los extremos 5' de las moléculas de ácido nucleico diana de la etapa (b) utilizando una ligasa; y
(d) determinar la secuencia de las moléculas de ácido nucleico diana que se obtienen en la etapa (c) utilizando un cebador capaz de hibridarse con el complemento de la molécula adaptadora 5' y un cebador capaz de hibridarse con la molécula adaptadora 3'.

En otro aspecto más, la presente invención proporciona un método para detectar preferentemente una molécula de ácido nucleico en una biblioteca de moléculas de ácido nucleico, comprendiendo el método:

(a) proporcionar un grupo de oligonucleótidos de cadena sencilla de secuencia conocida con extremos 3' bloqueados (moléculas adaptadoras 3') y un grupo de oligonucleótidos de cadena sencilla de secuencia conocida con extremos 5' bloqueados (moléculas adaptadoras 5') donde las moléculas adaptadoras 3' y 5' comprenden uno o más nucleótidos degenerados y que se pueden unir preferentemente a la molécula de ácido nucleico diana;
(b) unir las moléculas adaptadoras 3' a los extremos 3' de las moléculas de ácido nucleico diana utilizando una ligasa;
(c) unir las moléculas adaptadoras 5' a los extremos 5' de las moléculas de ácido nucleico diana de la etapa (b) utilizando una ligasa;
(d) crear una biblioteca de moléculas de ácido nucleico amplificada por PCR de las moléculas de ácido nucleico que se obtienen en la etapa (c) utilizando un cebador capaz de hibridarse con el complemento de la molécula adaptadora 5' y un cebador capaz de hibridarse con la molécula adaptadora 3'.
(e) secuenciar la biblioteca amplificada resultante de moléculas de ácido nucleico; y
(f) analizar los resultados de la secuenciación para determinar si la molécula de ácido nucleico está presente en la biblioteca de moléculas de ácido nucleico.

Preferentemente, el ácido nucleico diana se asocia con una enfermedad o estado de pre-enfermedad. Preferentemente, el ácido nucleico diana se asocia con un organismo en particular. Preferentemente, el ácido nucleico diana se asocia con un tipo de tejido particular. Preferentemente, el ácido nucleico diana se asocia con un estado de desarrollo particular.

En realizaciones preferidas de los aspectos descritos anteriormente, las moléculas de ácido nucleico son moléculas de ARN. En otras realizaciones preferidas de los aspectos descritos anteriormente, las moléculas de ácido nucleico son moléculas de ADN.

En otro aspecto, la presente invención proporciona un método para generar una biblioteca de ADNc a partir de una biblioteca de moléculas de ARN, comprendiendo el método:

(a) proporcionar un grupo de oligonucleótidos de cadena sencilla de secuencia conocida con extremos 3' bloqueados (moléculas adaptadoras 3') y un grupo de oligonucleótidos de cadena sencilla de secuencia conocida con los extremos 5' bloqueados (moléculas adaptadoras 5') donde las moléculas adaptadoras 3' y 5' comprenden uno o más nucleótidos degenerados;

(b) unir las moléculas adaptadoras 3' a los extremos 3' de las moléculas de ARN utilizando una ligasa;

(c) unir las moléculas adaptadoras 5' a los extremos 5' de las moléculas de ARN de la etapa (b) utilizando una ligasa;

(d) crear moléculas de estructura híbrida de ARN/ADN a partir de moléculas de ARN obtenidas en la etapa (c) utilizando una enzima transcriptasa inversa y un cebador capaz de hibridarse con la molécula adaptadora 3'; y

(e) crear una biblioteca de ADNc por PCR de las moléculas de estructura híbrida de ARN/ADN obtenidas en (d) utilizando un cebador capaz de hibridarse con el complemento de la molécula adaptadora 5' y un cebador capaz de hibridarse con la molécula adaptadora 3'.

Los oligonucleótidos de secuencia conocida con extremos 3' bloqueados (moléculas adaptadoras 3') y los oligonucleótidos de secuencia conocida con extremos 5' bloqueados (moléculas adaptadoras 5'), que contienen uno o más nucleótidos degenerados también se denominan en el presente documento adaptadores de Alta Definición (HD).

En otro aspecto, la presente invención proporciona un grupo de oligonucleótidos de cadena sencilla de secuencia conocida con extremos 3' bloqueados (moléculas adaptadoras 3') y un grupo de oligonucleótidos de cadena sencilla de secuencia conocida con extremos 5' bloqueados (moléculas adaptadoras 5') donde las moléculas adaptadoras 3' y 5' comprenden uno o más nucleótidos degenerados, para su uso en los métodos descritos en el presente documento.

En un aspecto más, la presente invención proporciona un grupo de oligonucleótidos de cadena sencilla de secuencia conocida con extremos 3' bloqueados (moléculas adaptadoras 3') y un grupo de oligonucleótidos de cadena sencilla de secuencia conocida con extremos 5' bloqueados (moléculas adaptadoras 5'), donde las moléculas adaptadoras 3' y 5' comprenden uno o más nucleótidos degenerados, donde dichos oligonucleótidos se pueden unir preferentemente a una secuencia diana, para su uso en los métodos descritos en el presente documento.

En realizaciones preferidas, los oligonucleótidos descritos en los aspectos anteriores pueden tener 1, 2, 3, 4, 5, 6 o más nucleótidos degenerados. Los nucleótidos degenerados se pueden agrupar en las regiones 3', 5' o centrales del oligonucleótido. De manera alternativa, se pueden distribuir a lo largo de la longitud del oligonucleótido en cualquier configuración.

**Descripción de los dibujos**

La presente invención se entenderá mejor por referencia a los dibujos adjuntos, en los que:

**Figura 1:** Análisis de los datos de la secuencia de muestras sintéticas N9 (un oligonucleótido ARN completamente degenerado que consiste en nueve N nucleótidos donde N es o A, C, G o U con un 25% de probabilidad para cada nucleótido). **A-C.** Gráficos de frecuencia de nucleótidos que muestran la proporción de cada nucleótido en cada posición de la secuencia. **A.** Los adaptadores Illumina muestran grandes tendencias por secuencias con el nucleótido Guanina en la posición 8, lo que ilustra el problema de tendencia. **B.** Los adaptadores HD tienen tendencias significativamente reducidas y las líneas están cercanas al óptimo teórico de frecuencia de 0,25. **C.** La muestra de 'no unión' muestra tendencias causadas por la PCR de los nucleótidos. Las tendencias de la PCR eran casi cero con líneas en el óptimo teórico de frecuencia de 0,25.

**Figura 2:** Análisis de los datos de la secuencia de muestras sintéticas N21 (un oligonucleótido ARN completamente degenerado que consiste en nueve nucleótidos N donde N es o A, C, G o U con un 25% de probabilidad para cada nucleótido). **A.** La proporción de lecturas con un número de recuentos particular. Se llevó a cabo HD e Illumina con dos ARN sintéticos independientes y se secuenciaron por duplicado. La muestra de 'no unión' se secuenció por duplicado. El óptimo teórico es un recuento por secuencia. Más del 95% de las secuencias se encontraron solo una vez en la muestra de adaptador HD. Los adaptadores Illumina favorecían algunas secuencias, dando como resultado múltiples lecturas de la misma secuencia. **B-D** Gráficos de frecuencias de nucleótidos que muestran la proporción de cada nucleótido en cada posición de la secuencia. **B.** Como ilustración de la tendencia de secuenciación, los adaptadores de Illumina muestran grandes tendencias para las secuencias con el nucleótido Guanina en la posición 20. **C**. Los adaptadores HD tienen tendencias significativamente reducidas y las líneas están más cercanas al óptimo teórico de frecuencia de 0,25. **D.** La muestra de 'no unión' muestra tendencias producidas por la PCR de los oligonucleótidos. Las secuencias con una composición alta en Guanina y Adenosina se amplifican y secuencian preferentemente.

**Figura 3:** La estructura secundaria de cada secuencia sintética N9 individual con las correspondientes secuencias adaptadoras Illumina se plegaron con un software RNAfold. Se representó la Energía Mínima Libre calculada por el RNAfold contra los recuentos de cada secuencia. Cuanto menor energía mínima libre de la estructura secundaria con los adaptadores más probable es que la secuencia se secuencie múltiples veces (coeficiente de correlación de Pearson igual a 0,67, valor p < 10e-15).

**Figura 4:** Gráficos de frecuencia de nucleótidos para las secuencias clonadas por el adaptador 3' HD **GAGA**TCGTATGCCGTCTTCTGCTTG (SEC ID Nº 1) (54.023 recuentos) y **ATTG**TCGTATGCCGTCTTCTGCTTG (SEC ID Nº 2) (51.928 recuentos).

**Figura 5:** Análisis de los datos de secuencia de muestras biológicas que mapean el genoma. Los gráficos de frecuencia de nucleótidos muestran la proporción de cada nucleótido en cada posición de las secuencias 22nt. **A** y **B** son muestras MCF7 y **C** y **D** son muestras MCF10a. **E** es una muestra completa de ratón. Los adaptadores Illumina muestran tendencias para secuencias con el nucleótido Guanina en los dos penúltimos nucleótidos 3' (posiciones 20 y 21) (**A** y **C**), que es similar a la tendencia observada en las muestras sintéticas. Los adaptadores HD tienen tendencias significativamente reducidas y las líneas están cercanas al óptimo teórico de frecuencia de 0,25 (**B** y **D**). La muestra clonada de ratón con los adaptadores HD tenían tendencias bajas similares a las

muestras con el adaptador HD MCF7 y MCF10a.

**Figura 6:** Número de miARN conocidos que se identificaron con un aumento de umbral en las muestras MCF7 clonadas con adaptadores Illumina y adaptadores HD. Las secuencias con números de lectura por debajo del umbral se excluyeron del recuento. Los adaptadores HD identificaron más miARN en todos los umbrales.

**Figura 7:** Cuantificación absoluta por análisis de transferencia de Northern. **A** y **B**. Se utilizó el gráfico de densitometría para comparar la intensidad de bandas de la señal de las muestras biológicas MCF7 y MCF10a con una curva de concentración de referencia que utilizaba un oligonucleótido con la secuencia idéntica al miARN. **C**. Número de lecturas para hsa-mir-103 y hsa-mir-25 utilizando adaptadores Illumina y HD. **D**. Cuantificación absoluta que muestra que hsa-mir-25 es ~ 10 veces más abundante que hsa-mir-103. Esto es similar a la cuantificación por adaptadores HD.

**Figura 8:** Frecuencia del emparejamiento de bases de nucleótidos prevista por posición utilizando los parámetros de plegamiento de ADN, temperatura igual a 50 ºC. **A**. el producto de N9 de unión artificial tiene un emparejamiento de bases de nucleótido aumentado en el extremo 3' del cebador inverso (flecha). La línea sólida y los círculos representan un grupo de secuencias con un número superior de 5.000 lecturas. En comparación las líneas discontinuas y los triángulos representan grupos de secuencias con un número inferior de 5000 lecturas. **B**. El producto N21 de unión artificial también tiene un emparejamiento de bases de nucleótidos aumentado en el extremo 3' del cebador inverso. La línea sólida y los círculos representan un grupo de secuencias con un número superior de 10.000 lecturas. En comparación las líneas discontinuas y los triángulos representan un grupo de secuencias generadas aleatoriamente.

**Figura 9:** Estructuras en horquilla en el extremo 3' del cebador inverso preferido en los experimentos qPCR que utilizan una Taq polimerasa. Se sintetizaron varias estructuras de ADN que o tenían una estructura secundaria prevista no fuerte o tenían estructuras secundarias previstas en posiciones o en los sitios de imprimación en los extremos de los sitios del cebador o entre los sitios de imprimación. Como las estructuras se añadieron en cantidades equimolares los resultados de la QPCR muestran la eficacia de la PCR. La QPCR era más eficaz para la estructura con una horquilla en el extremo 3' del cebador inverso. Las estructuras se cuantificaron espectrofotométricamente por triplicados y las diluciones se hicieron equimolares. Las diluciones de estructura se hicieron por triplicado (n = 3) con hasta cuatro replicados técnicos.

**Figura 10:** Frecuencia del emparejamiento de bases de nucleótidos prevista por posición para la muestra N21. La línea sólida y los círculos representan grupos de secuencias que se encuentran dos o más veces en el grupo de datos HD; las líneas discontinuas y los triángulos representan grupos de 5.000 secuencias generadas aleatoriamente. **A.** Secuencia que contienen la inserción y el adaptador 3'. **B.** Secuencia que contienen la inserción del adaptador 5' y el adaptador 3'. Las flechas indican el punto de unión.

**Figura 11:** Los protocolos de preparación de la biblioteca de ADNc distorsiona la investigación de miARN. **(a)** Comparación del cambio en el nivel de miARN entre las células tipo silvestre y Dicer KO DLD obtenidos en muestras Illumina (eje x) y HD (eje y). R2 = 0,62 **(b)** Número de miARN conocidos que se encuentran en las células DLD en diferentes umbrales utilizando adaptadores Illumina o HD. Independientemente del umbral escogido, los adaptadores HD identifican más miARN. (c) Cuantificación absoluta de ocho miARN conocidos (let-7i, miR-10a, miR-19b, miR-21, miR-25, miR-29b, miR-93, miR-375) que se obtiene por transferencia de Northern en comparación con el número de veces que estos miARN se secuenciaron utilizando adaptadores Illumina o HD en la línea celular DLD. Los datos obtenidos con adaptadores HD se correlacionan mejor con las cuantificaciones absolutas (R2 = 0,70) que los datos con Illumina (R2 = 0,12). (d) Número de citas PubMed y números de lecturas por experimentos (datos obtenidos de la miRbase v17) de miARN humanos. Los miARN con mayor número de lecturas tienden a estudiarse más extensamente.

**Figura 12:** Cuantificación de microARN conocidos en células DLD-1 (células de adenocarcinoma epitelial), tanto de tipo silvestre como Dicer -/-. Se analizó el ARN total (10 μg) por medio de transferencia de Northern y se cuantificaron las muestras experimentales de miARN utilizando una dilución en serie de secuencias de oligonucleótidos sintéticos que corresponden con los miR de interés. Se utilizó el pequeño ARN U6 nuclear como control de carga, y esta imagen de U6 se duplicó donde se analizaron los microARN en la misma membrana (miR 10a y 29b y miR 25 y Let 7i).

**Descripción detallada**

Las tecnologías de secuenciación de alto rendimiento son candidatas ideales para perfilar sARN debido a que tienen la capacidad de identificar sARN no anotados anteriormente y cuantificar su nivel de acumulación. El supuesto es que el número de veces que se encuentra una cierta lectura corta se correlaciona con el nivel de acumulación del sARN en las células. Sin embargo, recientemente los presentes inventores y otros (Tian, Yin et al. 2010; Linsen, de Wit et al. 2009; Willenbrock, Salomon et al. 2009, McCormick, Willmann et al. 2011, Hafner et al 2011) descubrieron que los diferentes protocolos de preparación de bibliotecas tienen preferencias por ciertos tipos de secuencias cortas, que da lugar a perfiles de sARN imprecisos. Algunas secuencias se encuentran más a menudo de lo que se esperaría, algunas secuencias se encuentran menos frecuentemente de lo esperado y puede haber algunas secuencias que no se encuentran a pesar del hecho de estar presentes en las células. Si se observara una sobre-presentación de algunas secuencias en los datos biológicos, se reduciría la representación media de otras secuencias. Por lo tanto, el potencial de secuenciación de pequeños ARN con baja abundancia se reduciría y se presenta la posibilidad de que algunos pequeños ARN sean 'inclonables' utilizando un protocolo de referencia y por tanto no se identifican. Por lo tanto, existe la necesidad de métodos más eficaces y eficientes para generar bibliotecas de clonación sin tendencias para los métodos de secuenciación de alto rendimiento y otros.

Se describe en el presente documento un método para determinar la tendencia de secuencia de una técnica de secuenciación, comprendiendo el método:

(a) proporcionar una biblioteca degenerada de moléculas de ácido nucleico;
(b) determinar la secuencia de las moléculas de ácido nucleico en la biblioteca degenerada de moléculas de ácido nucleico proporcionada en (a) que utiliza la técnica de secuenciación; y
(c) analizar los resultados de la secuenciación para determinar la sobre- o bajo-representación de moléculas de ácido nucleico particulares.

El significado de la expresión "biblioteca degenerada" estará claro para los expertos en la técnica. En general, se debería tomar como que significa un grupo de moléculas de ácido nucleico en el que se representa cualquier combinación posible de nucleótidos. Típicamente, aunque no siempre, la biblioteca será de nucleótidos de una determinada longitud. Por ejemplo, una biblioteca degenerada de moléculas de dinucleótidos tiene 16 miembros: AA; AG; AC; AT; GA; GG; GC; GT; TA; TG; TC; TT; CA; CG; CC; y CT. El número de miembros en una biblioteca degenerada de una longitud fija se determina por la fórmula $4^L$, donde L es la longitud del oligonucleótido. De manera similar, la expresión "nucleótido degenerado" estará clara para los expertos en la técnica. Cuando un oligonucleótido se describe como que tiene un nucleótido degenerado en una posición particular, el experto apreciará que se refiere a un grupo de oligonucleótidos que tiene o A, G, C o T en la posición degenerada, cada uno presente en una concentración aproximadamente igual. El número de oligonucleótidos distintos se determina por la fórmula $4^n$, donde *n* es el número de nucleótidos degenerados de la secuencia. Por lo tanto, un oligonucleótido que tiene cuatro posiciones degeneradas es de hecho un grupo de 256 oligonucleótidos únicos. Habitualmente, cada oligonucleótido estará presente en una concentración aproximadamente igual. Además, el experto en la técnica apreciará que en el caso del ARN, 'T' se puede remplazar por 'U'.

El significado de la expresión "se une" estará claro para el experto en la técnica y tiene un significado constante a lo largo de la presente solicitud. En general, se pretende que englobe la unión covalente de dos moléculas de ácido nucleico por medio de enlaces fosfodiéster. Habitualmente, esto se consigue con el uso de una enzima ligasa, incluyendo pero sin limitarse a las ADN ligasas I a IV o T4 ARN ligasa 1 o 2. Las condiciones de unión óptimas de ácidos nucleicos serán conocidas por los expertos en la técnica.

Se pretende que las expresiones "extremo 5' bloqueado" y "extremo 3' bloqueado" indiquen que la molécula no es capaz de tener otra molécula nucleica unida o ligada a su extremo 5' o 3', respectivamente. Los métodos de conseguir esto son bien conocidos por los expertos en la técnica e incluyen, pero no se limitan al uso de un nucleótido difosfato en la posición 5' o 3'. El significado de estas expresiones son constantes a lo largo de la presente solicitud.

El significado de la frase "capaz de hibridarse con" estará claro para los expertos en la técnica y tiene un significado constante a lo largo de la presente solicitud. En general, se pretende que englobe las condiciones que se encuentran durante la etapa de hibridación de una PCR típica o una PCR de transcriptasa inversa (rtPCR). El grado de hibridación solo necesita ser suficiente para asegurar que la reacción en cadena de la polimerasa tenga lugar y no necesita que sea sobre la longitud completa de la región diana. Como apreciará un experto, un oligonucleótido puede contener un número de no coincidencias y aún se considera capaz de hibridarse con una diana particular. Las condiciones que se encuentran durante las etapas de hibridación de una PCR serán conocidas en general por un experto en la técnica aunque las condiciones de hibridación precisas variarán entre una reacción y otra (véase Sambrook et al., 2001, Molecular Cloning, A Laboratory Manual, 3ª Ed, Cold Spring Harbor Laboratory Press, NY; Current Protocols, eds Ausubel et al.). Típicamente, tales condiciones pueden comprender, pero no se limitan a, (después de una etapa de desnaturalización a una temperatura de aproximadamente 94 ºC durante aproximadamente 1 minuto) la exposición a una temperatura en el intervalo de desde 40 ºC a 72 ºC (preferentemente 50-68 ºC) durante un periodo de aproximadamente 1 minuto en tampón de reacción de PCR de referencia.

Preferentemente, la técnica de secuenciación implica la unión de moléculas adaptadoras a los extremos 5', 3' o 5' y 3' de moléculas de ácido nucleico de la biblioteca degenerada de moléculas de ácido nucleico.

El método descrito en el presente documento puede comprender:

(a) proporcionar una biblioteca degenerada de moléculas de ácido nucleico;
(b) unir un oligonucleótido de secuencia conocida con un extremo 3' bloqueado (molécula adaptadora 3') con los extremos 3' de las moléculas de ácido nucleico utilizando una ligasa;
(c) unir un oligonucleótido de secuencia conocida con un extremo 5' bloqueado (molécula adaptadora 5') con los extremos 5' de las moléculas de ácido nucleico de la etapa (b) utilizando una ligasa;
(d) secuenciar la biblioteca resultante de moléculas de ácido nucleico unidas; y
(e) analizar los resultados de la secuenciación para determinar la sobre- o bajo-representación de moléculas de ácido nucleico particulares;

donde las etapas (b) y (c) se pueden llevar a cabo en cualquier orden.

El método puede comprender:

(a) proporcionar una biblioteca degenerada de moléculas de ácido nucleico, donde las moléculas de ácido nucleico tienen una región 3' constante y un extremo 3' bloqueado;
(b) unir un oligonucleótido de secuencia conocida con un extremo 5' bloqueado (molécula adaptadora 5') con los extremos 5' de las moléculas de ácido nucleico de la etapa (a) utilizando una ligasa;
(c) secuenciar la biblioteca amplificada resultante de moléculas de ácido nucleico; y
(d) analizar los resultados de secuenciación para determinar la sobre- o bajo-representación de moléculas de ácido nucleico particulares.

El método puede comprender:

(a) proporcionar una biblioteca degenerada de moléculas de ácido nucleico, donde las moléculas de ácido nucleico tienen una región 5' constante y un extremo 5' bloqueado;
(b) unir un oligonucleótido de secuencia conocida con un extremo 3' bloqueado (molécula adaptadora 3') con los extremos 5' de las moléculas de ácido nucleico de la etapa (a) utilizando una ligasa;
(c) secuenciar la biblioteca amplificada resultante de moléculas de ácido nucleico; y
(d) analizar los resultados de secuenciación para determinar la sobre- o bajo-representación de moléculas de ácido nucleico particulares.

El método puede comprender:

(a) proporcionar una biblioteca degenerada de moléculas de ácido nucleico, donde las moléculas de ácido nucleico tiene una región 5' constante y una región 3' constante;
(b) secuenciar la biblioteca amplificada resultante de moléculas de ácido nucleico; y
(c) analizar los resultados de la secuenciación para determinar la sobre- o bajo-representación de moléculas de ácido nucleico particulares.

El método puede comprender:

(a) proporcionar una biblioteca degenerada de moléculas de ARN;
(b) unir un oligonucleótido de secuencia conocida con un extremo 3' bloqueado (molécula adaptadora 3') con los extremos 3' de las moléculas de ARN utilizando una ligasa;
(c) unir un oligonucleótido de secuencia conocida con un extremo 5' bloqueado (molécula adaptadora 5') con los extremos 5' de la molécula de ARN de la etapa (b) utilizando una ligasa;
(d) transcribir de manera inversa las moléculas de ARN en ADNc utilizando una transcriptasa inversa y un cebador capaz de hibridarse con la molécula adaptadora 3';
(e) si es necesario, crear una biblioteca de moléculas de ácido nucleico amplificada por PCR de moléculas de ADNc que se obtiene en la etapa (d) utilizando un cebador capaz de hibridarse con el complemento de la molécula adaptadora 5' y un cebador capaz de hibridarse con la molécula adaptadora 3';
(d) secuenciar la biblioteca resultante de moléculas de ácido nucleico; y
(e) analizar los resultados de la secuenciación para determinar la sobre- o bajo-representación de moléculas de ácido nucleico particulares;

Donde las etapas (b) y (c) se pueden llevar a cabo en cualquier orden.

El método puede comprender:

(a) proporcionar una biblioteca degenerada de moléculas de ADN;
(b) unir un oligonucleótido de secuencia conocida con un extremo 3' bloqueado (molécula adaptadora 3') con los extremos 3' de las moléculas de ADN utilizando una ligasa;
(c) unir un oligonucleótido de secuencia conocida con un extremo 5' bloqueado (molécula adaptadora 5') con los extremos 5' de las moléculas de ADN de la etapa (b) utilizando una ligasa;
(d) amplificar las moléculas de ADN unidas por PCR asimétrica;
(e) si es necesario, crear una biblioteca amplificada de moléculas de ácido nucleico por PCR de las moléculas de ADN que se obtienen en la etapa (d) utilizando un cebador capaz de hibridarse con el complemento de la molécula adaptadora 5' y un cebador capaz de hibridarse con la molécula adaptadora 3';
(d) secuenciar la biblioteca resultante de las moléculas de ácido nucleico; y
(e) analizar los resultados de la secuenciación para determinar la sobre- o bajo-representación de moléculas de ácido nucleico particulares;

donde las etapas (b) y (c) se pueden llevar a cabo en cualquier orden.

La presente invención proporciona un método para reducir la tendencia de secuencia de una técnica de secuenciación que implica la unión a un adaptador, comprendiendo el método:

(a) proporcionar un grupo de oligonucleótidos de secuencia conocida con extremos 3' bloqueados (moléculas adaptadoras 3') y un grupo de oligonucleótidos de secuencia conocida con extremos 5' bloqueados (moléculas adaptadoras 5'), donde las moléculas adaptadoras 3' y 5' comprenden uno o más nucleótidos degenerados;
(b) unir las moléculas adaptadoras 3' con los extremos 3' de las moléculas de ácido nucleico diana utilizando una ligasa;
(c) unir las moléculas adaptadoras 5' con los extremos 5' de las moléculas de ácido nucleico diana de la etapa (b) utilizando una ligasa; y
(d) determinar la secuencia de las moléculas de ácido nucleico que se obtienen en la etapa (c) utilizando un cebador capaz de hibridarse con el complemento de la molécula adaptadora 5' y un cebador capaz de hibridarse con la molécula adaptadora 3'.

En otro aspecto más, la presente invención proporciona un método para detectar preferentemente una molécula de ácido nucleico diana en una biblioteca de moléculas de ácido nucleico, comprendiendo el método:

(a) proporcionar un grupo de oligonucleótidos de secuencia conocida con extremos 3' bloqueados (moléculas adaptadoras 3') y un grupo de oligonucleótidos de secuencia conocida con extremos 5' bloqueados (moléculas adaptadoras 5'), donde las moléculas adaptadoras 3' y 5' comprenden uno o más nucleótidos degenerados y se pueden unir preferentemente a la molécula de ácido nucleico diana;
(b) unir la molécula adaptadora 3' con los extremos 3' de las moléculas de ácido nucleico de la biblioteca de moléculas de ácido nucleico utilizando una ligasa;
(c) unir la molécula adaptadora 5' con los extremos 5' de las moléculas de ácido nucleico de la etapa (b) utilizando una ligasa;
(d) crear una biblioteca amplificada de moléculas de ácido nucleico por PCR de las moléculas de ácido nucleico que se obtienen en la etapa (c) utilizando un cebador capaz de hibridarse con el complemento de la molécula adaptadora 5' y un cebador capaz de hibridarse con la molécula adaptadora 3';
(e) secuenciar la biblioteca amplificada resultante de moléculas de ácido nucleico; y
(f) analizar los resultados de la secuenciación para determinar si la molécula de ácido nucleico diana está presente en la biblioteca de moléculas de ácido nucleico.

También se describe en el presente documento un método de detectar preferentemente una molécula de ácido nucleico diana en una biblioteca de moléculas de ácido nucleico, comprendiendo el método:

(a) proporcionar un oligonucleótido de secuencia conocida (molécula adaptadora), que comprende uno o más nucleótidos degenerados, donde dicho oligonucleótido se puede unir preferentemente a la molécula de ácido nucleico diana;
(b) unir la molécula adaptadora a las moléculas de ácido nucleico de la biblioteca utilizando una ligasa;
(c) llevar a cabo una PCR cuantitativa de las moléculas que se obtienen en la etapa (b) utilizando un cebador específico para la molécula adaptadora y un cebador específico para el ácido nucleico diana.

Preferentemente, el ácido nucleico diana se asocia con una enfermedad o un estado de pre-enfermedad. Preferentemente, el ácido nucleico diana se asocia con un organismo particular. Preferentemente, el ácido nucleico diana se asocia con un tipo de tejido particular, Preferentemente, el ácido nucleico diana se asocia con un estado de desarrollo particular.

En realizaciones preferidas de los aspectos descritos anteriormente, las moléculas de ácido nucleico son moléculas de ARN. En otras realizaciones preferidas de los aspectos descritos anteriormente, las moléculas de ácido nucleico son moléculas de ADN.

En otro aspecto, la presente invención proporciona un método para generar una biblioteca de ADNc a partir de una biblioteca de moléculas de ARN, comprendiendo el método:

(a) proporcionar un grupo de oligonucleótidos de secuencia conocida con extremos 3' bloqueados (moléculas adaptadoras 3') y un grupo de oligonucleótidos de secuencia conocida con extremos 5' bloqueados (moléculas adaptadoras 5'), donde las moléculas adaptadoras 3' y 5' comprenden uno o más nucleótidos degenerados:
(b) unir las moléculas adaptadoras 3' con los extremos 3' de las moléculas de ARN utilizando una ligasa;
(c) unir las moléculas adaptadoras 5' con los extremos 5' de las moléculas de ARN de la etapa (b) utilizando una ligasa;
(d) crear moléculas de estructura híbrida ARN/ADN a partir de moléculas de ARN que se obtienen en la etapa (d) utilizando un cebador capaz de hibridarse con el complemento de la molécula adaptadora 5' y un cebador capaz de hibridarse con la molécula adaptadora 3'.

En un aspecto, la presente invención proporciona un grupo de oligonucleótidos de secuencia conocida con extremos 3' (moléculas adaptadoras 3') y/o un grupo de oligonucleótidos de secuencia conocida con extremos 5' bloqueados

(moléculas adaptadoras 5'), donde las moléculas adaptadoras 3' y 5' comprenden uno o más nucleótidos degenerados, para su uso en los métodos descritos en el presente documento.

En un aspecto más, la presente invención proporciona un grupo de oligonucleótidos de secuencia conocida con extremos 3' bloqueados (moléculas adaptadoras 3') y/o un grupo de oligonucleótidos de secuencia conocida con extremos 5' bloqueados (moléculas adaptadoras 5'), donde las moléculas adaptadoras 3' y 5' comprenden uno o más nucleótidos degenerados, donde dichos oligonucleótidos se pueden unir preferentemente a una secuencia diana para su uso en los métodos descritos en el presente documento.

En realizaciones preferidas, los oligonucleótidos descritos en los aspectos anteriores pueden tener 1, 2, 3, 4, 5, 6 o más nucleótidos degenerados. Los nucleótidos degenerados se pueden agrupas en las regiones 3', 5' o centrales del oligonucleótido. De manera alternativa, se pueden distribuir a lo largo de la longitud del oligonucleótido en cualquier configuración.

Los presentes inventores pretenden evaluar la preferencia de secuencia para la preparación de biblioteca por secuenciación de alto rendimiento, que se enfoca en el protocolo Illumina de clonación de pequeños ARN ya que es la plataforma que se utiliza más a menudo para los sARN. En vez de ensayar la preferencia de miARN conocidos (Linsen, de Wit et al. 2009; Willenbrock, Salomon et al. 2009) los presenten inventores desarrollan un nuevo ensayo para ensayar todas las secuencias posibles con el fin de entender la razón que hay detrás de la preferencia. Por lo tanto, los presentes inventores generaron bibliotecas de ADNc utilizando dos oligonucleótidos de ARN 21mero (N21) completamente degenerados, que contenían o Adenosina (A), Guanina (G), Citosina (C) o Uracilo (U) en cada posición con una probabilidad del 25% para dada tipo de nucleótidos ya que se presume que cada secuencia se sintetiza con la misma concentración (se utilizaron dos lotes independientes de N21 para minimizar el riesgo de tendencia durante la síntesis). Los oligonucleótidos N21 se unen a un adaptador con un extremo 5' pre-adenilado y un extremo 3' bloqueado de forma que este adaptador (adaptador 3') solo se puede unir al extremo 3' del oligonucleótido N21. Luego los productos de la unión se unen a un adaptador diferente (adaptador 5') que tiene un extremo 5' bloqueado y un extremo 3' hidroxilo. Estos productos de unión se utilizan entonces como matrices en una reacción de transcripción inversa que se inicia con un cebador complementario al adaptador 3' seguido por una reacción PCR que utiliza cebadores que se pueden hibridar con los adaptadores 5' y 3', respectivamente. Los productos PCR se secuencian entonces en la plataforma Illumina GAII. Si no había tendencia de secuencia, tras la etapa de PCR todas las secuencias deberían estar presentes en la biblioteca un número de veces similar. Sin embargo, como el número de posibles secuencias en la biblioteca N21 es de 4.398 trillones y solamente aproximadamente 20-25 millones lecturas se pueden completar, muchas secuencias no se pueden secuenciar del todo y solo habrá 1-2 lecturas para las que se secuencian. Para superar este problema, los presentes inventores generaron también una biblioteca para un oligonucleótido ARN 9mero (N9) degenerando, que contienen 262.144 secuencias diferentes. Por lo tanto, se puede esperar un número de lecturas significativo para cada secuencia si no hay tendencia de secuencia. Los tres oligonucleótidos degenerados (dos N21 y un N9) se utilizaron para generar dos bibliotecas independientes.

En paralelo a estos 'experimentos de unión', los presentes inventores investigaron la línea base de la tendencia de nucleótido producida por la PCR y la máquina de secuenciación. Se sintetizó un grupo de oligonucleótidos ADN que simulaba un miARN clonado y transcrito inversamente. Esta secuencia incluía una secuencia adaptadora 5', una secuencia degenerada 21 N central y una secuencia adaptadora 3'. Tras la amplificación PCR los productos se secuenciaron.

**Resultados**

Los adaptadores Illumina no eran uniformes cuando clonaban ARN pequeños. Las lecturas obtenidas en las bibliotecas de adaptadores Illumina N9 no mostraban una distribución uniforme a lo largo de todas las secuencias posibles. De hecho, el 56% de las secuencias posibles no se encontraron en absoluto y las lecturas mostraban una tendencia de nucleótidos específica de la posición muy fuerte (Figura 1 A). Las bibliotecas N21 también mostraban tendencias muy fuertes: se encontraron 35 secuencias más de 100 veces y 548 se encontraron más de 50 veces en vez de las 1-2 veces que se esperaban (Figura 2A). Los gráficos de frecuencia de nucleótidos también mostraban fuertes tendencias en la mayoría de las posiciones de nucleótidos (Figura 2B).

La forma truncada de T4 ARN ligasa 2, que se utiliza para unir los sARN (en nuestro caso los oligonucleótidos N9 y N21) con el adaptador 3', puede reparar muescas en el ARN de cadena doble (dsARN) *in vitro* (Nandakumar and Shuman, 2004). Sin el deseo de quedar ligados por una teoría en particular, los presentes inventores hacen la hipótesis de que secuencias que pueden formar una estructura tipo dsARN con el adaptador 3' debería estar sobre-representada en las lecturas de secuencias. Todas las lecturas de secuencias se unieron a la secuencia adaptadora 3' y se calculó la energía libre de las secuencias de ARN resultantes para cada molécula. La abundancia de una secuencia en la biblioteca mostraba una fuerte correlación con el valor de la energía mínima libre; cuanto menor era la energía libre para una secuencia determinante, más abundante era la secuencia en la biblioteca (Figura 3). En base a esta observación los presentes inventores hicieron la hipótesis de que añadiendo nucleótidos degenerados a los adaptadores, sería menos probable que las moléculas adaptadoras ligeramente diferentes formaran estructuras secundarias estables con diferentes tipos de secuencias de sARN. Esto podía permitir; 1, la secuenciación de sARN

que normalmente no están presentes en las bibliotecas generadas por adaptadores tradicionales y 2, la abundancia de secuencias reflejaría mejor la concentración de sARN en la muestra. Para ensayar este hipótesis se añadieron cuatro nucleótidos N (A, C, G o U) al extremo 5' del adaptador 3' y también al extremo 3' del adaptador 5'. Estos adaptadores se denominan adaptadores de Alta Definición (HD) para distinguirlos de los adaptadores Illumina (que tienen secuencias fijas).

Se utilizaron los mismos oligonucleótidos N9 y N21 para la generación de la biblioteca con los adaptadores HD como anteriormente y se secuenciaron de nuevo las bibliotecas en la plataforma Illumina GAII. Casi el doble (un 78% vs. 44%) de secuencias diferentes se leían entre las lecturas que se obtenían de la biblioteca N9 probando que los adaptadores HD eran incluso mucho más sensibles que los adaptadores Illumina. Las frecuencias de nucleótidos de las lecturas que se obtenían por los adaptadores HD eran muchos más similares entre ellas en todas las posibles secuencias diferentes (compárese la Figura 1A con 1B y 2B con 2C). La mayoría de las lecturas para las bibliotecas N21 que se obtenían con los adaptadores HD estaban alrededor de las 1 a 2 esperadas y solamente muy pocas lecturas estaban presentes en un número más alto (Figura 2A). Estos resultados demuestran que añadiendo nucleótidos degenerados al extremo de los adaptadores aumenta drásticamente su sensibilidad y reduce la tendencia de secuencia. Si se observara una profunda sobre-representación de algunas secuencias en los datos biológicos se reduciría la representación media de otras secuencias. Por lo tanto, el potencial para secuenciar ARN pequeños de baja abundancia se reduciría y aparece la posibilidad de que algunos ARN pequeños sean 'inclonables' utilizando protocolos de referencia y por tanto de ser identificados. Los actuales adaptadores HD muestran aún alguna tendencia de secuencia pero una vigilancia sistemática puede establecer la posición óptima y el número de nucleótidos degenerados en los adaptadores HD que mostrarán tendencias mínimas y una eficacia máxima utilizando el ensayo descrito en el presente documento.

Los adaptadores HD también se pueden apreciar como un grupo complejo de 256 adaptadores con 65.536 pares posibles. Los pares adaptadores individuales tienen preferencias particulares para clonar un grupo de secuencias. Por ejemplo, el adaptador HD 3' con la secuencia **ATTG**TCGTATGCCGTCTTCTGCTTG (SEC ID Nº 2) tiene una tendencia muy fuerte por secuencias con nucleótidos Guanina en el extremo 3' cuando se comparan con secuencias clonadas por el adaptador 3' **GAGA**TCGTATGCCGTCTTCTGCTTG (SEC ID Nº 1). El 75% de todas las secuencias clonadas por el adaptador **ATT**GTCGTATGCCGTCTTCTGCTTG (SEC ID Nº 2) tiene una Guanina en la posición 9 (Figura 4). Esto es un hallazgo importante y permite el diseño de adaptadores sin tendencia por multiplexación. Quizá una aplicación más útil en el futuro será la manipulación de tendencias utilizando adaptadores, por ejemplo, para secuenciar preferentemente miARN de baja abundancia asociados con una enfermedad o para excluir secuencias altamente abundantes que dominen en los datos.

Los adaptadores HD también se ensayaron con muestras biológicas. Se generaron bibliotecas utilizando adaptadores Illumina o HD a partir del ARN de la línea celular MCF7 de cáncer de mama y se compararon los resultados con la línea celular MCF10a no cancerosa. Estos experimentos se diseñaron para ensayar la eficacia de los adaptadores HD para identificar miARN expresados diferencialmente y para demostrar su cuantificación precisa. Estos experimentos pueden demostrar que el uso de adaptadores HD aumentan la capacidad de identificar más miARN. Se descubrió que una biblioteca preparada con adaptadores HD identificaba más del doble de secuencias distintas que mapeaban el genoma (Tabla 1) en comparación con una biblioteca preparada con adaptadores Illumina. Por ejemplo en la muestra MCF7, 23.228 lecturas por millón eran distintas utilizando los adaptadores HD mientras que solo 10.903 secuencias por millón serán distintas utilizando adaptadores Illumina.

**Tabla 1:** Número de secuencias distintas y redundantes y las que mapean el genoma. Los datos normalizados de lecturas por millón demuestran que los adaptadores HD tienen una eficacia de más de dos veces en clonar pequeños ARN.

| Línea Celular | Total de distintos | Total de redundantes | Mapeo distinto | Mapeo redundante | mapeo distinto por millón |
|---|---|---|---|---|---|
| mcf7r1fa | 627509 | 27054847 | 24755 | 22481015 | 11011,69 |
| mcf7r2fa | 542078 | 26769495 | 249639 | 23126674 | 10794,42 |
| mcf7r1va | 705058 | 15459296 | 211847 | 9211267 | 22998,68 |
| mcf7r2va | 759603 | 16391091 | 217605 | 9276366 | 23458 |
| mcf10r1fa | 698069 | 24782969 | 334196 | 19769573 | 16904,56 |
| mcf10r2fa | 466299 | 18008481 | 238092 | 14370373 | 16568,25 |
| mcf10r1va | 989636 | 12728537 | 257233 | 6545624 | 39298,47 |
| mcf10r2va | 1002778 | 12330132 | 243861 | 6134363 | 39753,27 |

El aumento en complejidad de las lecturas de secuencias produce una reducción concomitante en la tendencia de secuencia en el grupo de datos HD. Esto se puede observar en los gráficos de frecuencia de nucleótidos (Figura 5; compárese A con B y C con D). La tendencia de secuencia en las bibliotecas generadas a partir de líneas celulares se espera que presenten una tendencia considerable debido a que el repertorio de microARN es limitado. Las bibliotecas generadas a partir de muchos tipos de tejido se espera que capture un grupo más diverso de pequeños ARN y tienen tendencias más reducidas. Las bibliotecas se generaron a partir de ratón entero (P3 C57 post-natal). Como se esperaba muchas más secuencias que las mapeadas para el genoma del ratón eran distintas (47.674 de ratón vs. 23.228 MCF7 vs. 39.525 MCF10a por millón) y en esta grupo de datos la tendencia de secuencia está más reducida en comparación con las muestras MCF7 y MCF10a (Figura 5 E).

A menudo es valioso identificar cuales miARN se expresan en un grupo de datos. Para demostrar que los adaptadores HD producen datos con más miARN identificables, los inventores buscaron miARN conocidos en el grupo de datos. La Figura 6 muestra el número de secuencias que alcanzan un umbral de números de lecturas. A un nivel de umbral de 5 lecturas por millón, solamente se identifican ~ 250 miARN conocidos a partir de adaptadores Illumina pero los adaptadores HD identificaban ~ 350. Según aumenta el umbral se identifican menos miARN en los grupos de datos tanto de Illumina como HD. Sin embargo se identificaron más miARN conocidos utilizando adaptadores HD que adaptadores Illumina a cualquier umbral en particular.

La cuantificación absoluta de algunos miARN era más similar al número de lecturas generadas por los adaptadores HD (Figura 7). Por ejemplo, utilizando análisis de Northern se encontró que la cuantificación absoluta de mir25 (0,508 nM) era 9,5 veces mayor que mir-103 (0,0537). Sin embargo, el grupo de datos generados con adaptadores Illumina sugieren una relación opuesta; que la secuencia mir-103 (15.877 lecturas) era 15 veces más prevalente que la secuencia mir-25 (1035 lecturas). El grupo de datos del adaptador HD era capaz de predecir correctamente las cantidades relativas de esos dos miARN; mir-25 (23.269 lecturas) tenía lecturas 6,4 veces mayores que mir-103 (3610 lecturas).

Los inventores ensayaron a continuación los adaptadores HD sobre la línea celular de cáncer de colon DLD-1 y la línea celular DLD-1 Dicer KO mutante parcial de exón 5. Dado que se espera que las tendencias sean específicas de secuencia, las mismas secuencias en diferentes muestras se someterán a tendencias similares. Los análisis de veces de cambio de expresión por lo tanto no se afectan mucho por estas tendencias. Se confirmó que las veces de cambio de la expresión de miARN entre DLD-1 WT y DLD-1 Dicer KO era similar en las bibliotecas que utilizan adaptadores HD e Illumina (Figura 11 a). Por lo tanto, los adaptadores HD e Illumina son valiosos para identificar los sARN que se expresan diferencialmente.

La cuantificación precisa de sARN es crucial debido a que los investigadores se centran en miARN con alto número de lecturas. Los inventores descubrieron que los miARN con altos recuentos leídos en miRBase era significativamente más probable que se citaran por la comunidad investigadora. ($R2 = 0,25$, $p = 10\text{-}15$). Esto no es sorprendente debido a que habitualmente los miARN que se expresan altamente (es decir que tienen un número de lecturas alto) y muestran la expresión diferencial más fuerte en comparación con otra muestra (de control u otro tratamiento u otro tejido, etc.) se seleccionan por análisis funcional en profundidad. Los miARN se clasificaron en base a su número de lecturas normalizado en células DLD-1 utilizando adaptadores HD e Illumina. El miARN más abundante en las bibliotecas generadas con el adaptador HD era el miR-29b con más de 150.000 lecturas pro millón, lo que es casi dos veces más que el siguiente miARN. Por lo tanto sería razonable escoger el miR-29b para posteriores análisis si se estuviera interesado en el papel de los miARN en la biología del cáncer de colon. Sin embargo, utilizando los adaptadores Illumina, el miR-29b era solo el 29 de la lista de clasificación con 3.336 lecturas normalizadas, mientras que los cuatro miARN superiores tenían más de 100.000 lecturas normalizadas en esa biblioteca. Está claro que el miR-29b no se debería escoger para análisis posteriores basándose en los resultados de secuenciación con Illumina. Además, solo cinco de los diez miARN superiores más secuenciados utilizando adaptadores HD estaban también en los diez miARN superiores más secuenciados utilizando adaptadores Illumina. Por lo tanto, la priorización de los miARN por análisis en profundidad podría ser altamente dependiente de los adaptadores utilizados, al menos para algunas muestras. Se utilizó el análisis de transferencia de Northern cuantitativo para demostrar que las bibliotecas hechas con adaptadores HD reflejaban la abundancia celular de los sARN pero las bibliotecas hechas con adaptadores Illumina no (Figura 11c, Figura 12). No todos los miARN muestran una diferencia drástica en las dos listas de clasificación (por ejemplo, miR-93 y miR-10a se clasificaron segundo y tercero en la lista de adaptador HD, y cuarto y segundo en la lista de adaptador Illumina, respectivamente) pero el ejemplo de miR-29b ilustra que potencialmente muchos miARN no se eligieron para análisis funcionales en los estudios previos. A continuación los inventores investigaron la cobertura de secuencia de los adaptadores HD. Se descubrió que el protocolo HD identificaba más del doble de secuencias distintas que mapeaban el genoma en comparación con una biblioteca preparada con adaptadores Illumina. Los adaptadores HD también capturaban aproximadamente un 25% más de miARN conocidos en cualquier umbral de recuento particular en comparación con los adaptadores Illumina (Figura 11b). Los adaptadores HD también eran capaces de capturar miARN previamente no identificados. El algoritmo MIRCat (Moxon et al 2008) se utilizó para identificar 32 miARN candidatos utilizando datos HD o Illumina. Además para identificar 309 miARN conocidos en esta línea celular los adaptadores HD eran capaces para capturar 26 nuevos miARN. Cinco de estos se secuenciaron también por adaptadores Illumina, pero solo había tres nuevos miARN, que solo eran capturados por los adaptadores Illumina. El número de lecturas normalizado de estos 29 nuevos miARN era al menos 1,5 veces menor en la línea celular DLD-1

KO Dicer, apoyando que se generaban por Dicer. Además, los inventores buscaron datos de secuenciación profundos en miRBase y encontraron lecturas que hacían coincidir las supuestas secuencias de miARN* para todos los nuevos genes miARN. Diecisiete de estos nuevos miARN (13 capturados solo por los adaptadores HD) no se había podido encontrar previamente ya que no estaban incluidas en ninguna de las secuencias básicas depositadas en la miRBase a partir de más de 100 experimentos d secuenciación profunda diferentes. Por lo tanto es razonable sugerir que los nuevos miARN se identificarán en otros tejidos, especialmente en el tejido cerebral, que muestra la población de miARN más diversa.

Otra consecuencia de la tendencia de unión es la potencial mala anotación de las dos cadenas de un miARN doble. El 'miARN maduro' activo habitualmente se determina por el número de lecturas más alto en comparación con la secuencia 'estrella' y estas frecuencias se pueden estimar o la relación de recuentos de las dos cadenas. Sin embargo, estas estimaciones también son proclives a distorsionarse por las tendencias de unión que dan lugar potencialmente a una anotación incorrecta de maduros y estrellas. Las relaciones de recuento se compararon en todas las parejas anotadas de miARN derivadas del mismo precursor que se expresaban con un nivel de moderado a alto (> 10 lecturas por millón), utilizando los grupos de datos HD e Illumina en DLD-1. Aunque la correlación entre las relaciones obtenidas con los dos protocolos era relativamente fuerte (R2 = 0,69, datos no mostrados), los inventores descubrieron 14 pares fuera de los 122 pares miARN/miARN* para los que la cadena de miARN con un número de lecturas más alto era diferente de los datos obtenidos con adaptadores Illumina y HD.

Los adaptadores HD generaban además algunas tendencias de secuencia. Las secuencias que se preveían por RNAfold (Hofacker 2003) que tenían estructuras secundarias fuertes con las secuencias de adaptador se secuenciaron preferentemente. No era posible alterar esta secuencia central, pero eliminando los efectos de esta secuencia central se debería reducir significativamente la tendencia de secuenciación. Esto se podía hacer de dos maneras. Un adaptador con una secuencia degenerada y una secuencia a medida se podría unir al pequeño ARN seguido por una PCR con un oligo para incorporar la secuencia de adaptador Illumina. De manera alternativa la secuencia central del adaptador se podría bloquear de que forme estructuras secundarias utilizando un oligonucleótido complementario, por ejemplo:

ADAPTADOR 5' (r = ARN)

```
5' GTTCAGAGTTCTACAGTCCGACGATCrNrNrNrN 3'   (SEC ID Nº 3)

   ||||||||||||||||||||||||||

3' CAAGTCTCAAGATGTCAGGCTGCTAG 5'           (SEC ID Nº 4)
```

ADAPTADOR 3' (r = ARN)

```
5' rNrNrNrNATCTCGTATGCCGTCTTCTGCTTG 3'     (SEC ID Nº 5)

           ||||||||||||||||||||||||||

3'         TAGAGCATACGGCAGAAGACGAAC 5'     (SEC ID Nº 6)
```

La tendencia de la PCR también es un factor que contribuye cuando se genera una biblioteca de pequeños ARN. Utilizando una secuencia de ADN degenerada flanqueada por los sitios de imprimación Illumina PCR, los presentes inventores también han mostrado que la tendencia de la PCR era mínima para la muestra N9 pero era más alta para la muestra N21 más larga (Figura 1C y 2D). Esto sugiere que según es más larga la secuencia clonada aumenta la tendencia PCR . Las secuencias que tienen un alto contenido en G/A están favorecidas particularmente durante la PCR y la secuenciación. Los inventores también descubrieron una relación entre la estructura secundaria del producto PCR y los recuentos de lectura (Figura 8). En la muestra N9 la correlación entre el número de lecturas para cada secuencia y las energías mínimas libres, aproximadamente igual a 0,1, se encontró que era estadísticamente significativa (valor p de t-test < $10^{-15}$).

**Sec ARN**

Este método es una alternativa al análisis de micromatriz por análisis de transcriptoma. Tiene la ventaja de identificar genes desconocidos previamente y está previsto para micromatrices supersede. La tendencia de secuencia se ha identificado para genes pequeños y genes que son ricos en AT (Oshlack and Wakefield 2009), y por el método de generación de biblioteca de imprimación de hexámero aleatorio (Hansen, Brenner et al. 2010). Los adaptadores HD podían reducir la tendencia para algunos métodos de generación de bibliotecas. Los adaptadores HD se podían utilizar en protocolos en los que se cortan los mARN y luego se unen los adaptadores. La transcripción inversa se continúa entonces tras la unión. La cobertura de secuenciación cuando se utilizan los adaptadores HD podrían ser más incluso a través de genes de interés.

**Secuenciación del genoma**

Se generaron bibliotecas genómicas del ADN cortado de doble cadena. Al ADN cortado se le truncan los extremos y luego se generan protuberancias 'A'. Los adaptadores tienen protuberancias 'T' y se unen al ADN con una ADN ligasa. Las tecnologías de secuenciación Illumina y Sólida se predisponen contra las regiones ricas en AT (Dohm, Lottaz et al. 2008; Harismendy, Ng et al. 2009) pero no tienen tendencias significativas en los extremos de secuencia (Hansen, Brenner et al. 2010). Los adaptadores HD pueden ayudar a reducir tendencias si la ligasa ADN tiene cualquier preferencia de secuencia. Además, los nucleótidos degenerados pueden ayudar a reducir la composición AT de algunas secuencias que podrían dar como resultado un aumento de las lecturas. Puede ser deseable también aumentar la degeneración de nucleótidos en los sitios de unión. Más que utilizar solo protuberancias T, se podrían utilizar protuberancias G/T en los adaptadores en conjunción con protuberancias C/A en las inserciones.

**QPCR**

La QPCR es particularmente desafiante para cuantificar pequeños ARN por al menos las siguientes razones:

(i) los miARN maduros son cortos (~ 22 nucleótidos; nts);
(ii) los miARN son heterogéneos en su contenido en GC, que resulta en un intervalo relativamente grande de temperatura de fusión ($T_m$) de dúplex de ácido nucleico para la población de miARN;
(iii) los miARN maduros carecen de una característica de secuencia común que facilitaría su purificación selectiva [por ejemplo, poli(A)];
(iv) la secuencia diana está presente en la transcripción primaria (pri-miARN) y el precursor (pre-miARN), en adición al miARN maduro;
(v) los miARN con la misma familia pueden diferenciarse por un único nucleótido (por ejemplo, familia Let-7).

Las fuertes tendencias pueden producirse por las diferentes preparaciones de biblioteca (Linsen, de Wit et al. 2009; Benes and Castoldi 2010). Utilizando adaptadores HD modificados se puede ayudar a reducir estas tendencias. Un adaptador HD sugerido para QPCR puede tener una región fija (de forma que la $T_m$ del cebador PCR se pueda ajustar) seguida por 20 nucleótidos degenerados y una secuencia fija 3' para la transcripción inversa y PCR.

Por ejemplo:

5' CAAANNNNNNNNNNNNNNNNNNNNSECUENCIAFIJA 3' (SEC ID Nº 7)

La secuencia fija podría bloquear la hibridación or un oligo complementario para reducir adicionalmente la tendencia.

**Métodos**

1. Aislar pequeños ARN utilizando un kit tal como el kit de aislamiento de pequeños ARN Mirvana (Ambion).
2. Mezclar lo siguiente:

   0,5 μl de adaptador adenilado de alta definición 3' (5 μM) 2,5 μl de pequeños ARN enriquecidos en ARN

3. Calentar la muestra a 70 ºC durante 2 minutos, y luego colocar inmediatamente la muestra en hielo.
4. Añadir lo siguiente:

   0,5 μl de tampón de ligasa truncada (NEB)
   0,4 μl $MgCl_2$ 100 mM
   0,75 μl ligasa truncada (NEB)
   0,25 μl RNasaOUT (Invitrogen)

5. Incubar la muestra a 22 ºC durante 1-2 h.
6. Calentar el adaptador de alta definición 5' a 70 ºC durante 2 minutos, después colocarla en hielo.
7. Añadir lo siguiente:

   0,5 μl ATP (10 mM)
   0,5 μl adaptador de alta definición 5'
   0,5 μl ssARN ligasa (NEB)

8. Incubar la muestra durante 1-2 h a 20 ºC.
9. Añadir lo siguiente al ARN unido (6,9 μl):

   1,5 μl cebador RT diluida (5 μM)

10. Incubar a 70 ºC durante 2 min.

11. Añadir lo siguiente:

3 μl de tampón de transcriptasa inversa de 1ª cadena (Invitrogen)
0,75 μl de dNTPs (10 mM)
1,5 μl DTT (Invitrogen)
0,75 μl de RNasaOUT (Invitrogen)

12. Incubar a 48 ºC durante 3 minutos.
13. Añadir 1,5 μl de Superscript II (Invitrogen).
14. Incubar a 44 ºC durante 1 h.
15. Llevar a cabo la PCR. Añadir lo siguiente:

12,9 μl de agua
4 μl de 5x Tampón (Finnzyme)
0,2 μl de cebador GX1 (50 μM)
0,2 μl de cebador GX2 (50 μM)
0,5 μl de dNTPs (10 nM)
0,2 μl Phusion polimerasa (Finnzyme)
2 μl de reacción de transcripción inversa

Llevar a cabo los ciclos de PCR

98 ºC 30 segundos
luego 8-13 ciclos
98 ºC 10 segundos
60 ºC 30 segundos
72 ºC 15 segundos
Luego 72 ºC durante 10 minutos

16. Procesar la muestra en gel de poliacrilamida (8%) a 150 V durante 1 h.
17. Aislar el fragmento de gel correspondiente a ~ 100 pb.

Secuencias adaptadoras (r = ARN; App = difosfato de adenosina)

Adaptadores Illumina v1.5 'Fijos':

5'rGrTrTrCrArGrArGrTrTrCrTrArCrArGrTrCrCrGrArCrGrArTrC 3' (SEC ID Nº 8)
5'ApprATCTCGTATGCCGTCTTCTGCTTG 3' (SEC ID Nº 9)

Adaptadores de 'Alta Definición':

5'GTTCAGAGTTCTACAGTCCGACGATCNrNrNrN 3' (SEC ID Nº 10)
5'ApprNrNrNNTCGTATGCCGTCTTCTGCTTG 3' (SEC ID Nº 11)

**Investigación de la función de ADN polimerasa y ARN ligasa**

El método de la invención se utilizó también como una forma eficaz para investigar la función de enzimas que modifican los ácidos nucleicos tales como ADN polimerasa y ARN ligasas.

Asumiendo que la biblioteca degenerada que se presenta a la enzima es equimolar, si la eficacia de la reacción catalizada por la enzima es la misma para cada secuencia de nucleótidos, la equimolaridad se debería preservar. Si este es el caso, debido al procedimiento de secuenciación es esencialmente un proceso de muestreo donde el tamaño de la muestra es muy grande (> $10^7$) y las frecuencias son muy bajas, el número de recuentos que se observa debería estar muy próximo a la distribución de Poisson. Es decir, el número de secuencias de nucleótidos distintas que se secuencian *k* veces debería ser aproximadamente igual a:

$$\frac{\lambda^k}{k!} \cdot e^{-\lambda}$$

donde $\lambda$ es igual a la relación entre el número total de secuencias leídas y el número de secuencias posibles. Por medio de un ensayo $\chi^2$ es posible ensayar esta hipótesis. Para todas las bibliotecas de tamaño 9 el valor de p de este ensayo estaba por debajo de $10^{-15}$, para bibliotecas de tamaño 21 que se preparaba con ligasa y adaptadores de referencia el valor de p estaba por debajo de $10^{-15}$, para bibliotecas preparadas solo con PCR y las bibliotecas preparadas con ligasa y adaptadores HD el valor de p era mayor de 0,2. Por lo tanto, el método demuestra que la

enzima no tiene la misma eficacia para cada secuencia de nucleótidos.

La ADN polimerasa es probablemente la enzima más importante que se utiliza en biología molecular ya que es esencial para la PCR. La comprensión de las preferencias de secuencia de la ADN polimerasa ayudará a mejorar su eficacia. Se utilizó un oligonucleótido degenerado para estudiar las preferencias de secuencia de la ADN polimerasa Phusion®. Se diseñó un oligonucleótido ADN que incorporaba una región degenerada (9 nt o 21 nt) flanqueada por las secuencias adaptadoras necesarias para la secuenciación Illumina. Esta secuencia se amplificó utilizando la ADN polimerasa Phusion® durante 15 ciclos y se extrajo del gel la correspondiente banda de 100 nt y se secuenció utilizando técnicas Illumina de referencia. Se encontró que para el 9 mero aleatorio, el 99,5% de todas las secuencias posibles se identificaban pero muchas estaban o sobre- o bajo-representadas. De manera similar, muchas secuencias se sobre-representaban en la muestra de 21mero (Figura 2A, no unión).

Había una fuerte correlación entre la estructura secundaria del producto PCR de cadena sencilla y se leyeron los recuentos para las muestras de N9 y N21. En comparación con las muestras de control (las 10.000 secuencias inferiores leídas para N9 y generadas aleatoriamente para las muestras de 21 N) la polimerasa Phusion® prefería una estructura en horquilla en el extremo del extremo 3' del cebador inverso (Figura 8). Esta preferencia se confirmó utilizando matrices de oligonucleótidos sintetizados con horquillas en varias posiciones. Se utiliza QPCR para medir la actividad de la Taq polimerasa. Como se esperaba la presencia de una horquilla en el sitio del cebador para los cebadores directo e inverso reducía significativamente la actividad de la Taq polimerasa. Sin embargo, la Taq polimerasa era más eficaz para la secuencia con una horquilla en el sitio del cebador del extremo 5' del cebador 3' (Figura 9).

La reducción de la eficacia de la PCR para las estructuras secundarias en la región del cebador se ha descrito como un fenómeno, pero faltan pruebas experimentales (Hoebeeck, van der Luijt et al. 2005). Este trabajo proporciona la primera evidencia completa de que las estructuras secundarias en el sitio del cebador son perjudiciales para la PCR. Basándose en estas observaciones se prevé que las estructuras tallo-lazo modificadas en los extremos 5' de los sitios de cebadores 3' se podrían utilizar para optimizar la eficacia de la PCR.

Los métodos descritos en el presente documento se utilizaron también para investigar las necesidades funcionales para la ARN ligasa 1 y 2. Las ARN ligasas son dependientes del contexto de estructura secundaria en el sitio de unión. La T4 ARN ligasa 1 favorece el ARN de cadena sencilla. La T4 ARN ligasa 2 truncada se puede unir al ARN de cadena doble o sencilla pero se cree que prefiere la doble cadena (Yin, Ho et al. 2003; Nandakumar, Ho et al. 2004). Se analizó la preferencia por la estructura secundaria de la T4 ARN ligasa 1 y la T4 ARN ligasa 2 truncada. Se unió un oligonucleótido 21 degenerado al adaptador HD 3' seguido por la unión al adaptador HD 5'. La mayoría de las secuencias clonadas representan la estructura secundaria preferida para la actividad de la ARN ligasa.

Para analizar la preferencia de estructura secundaria de la segunda unión se generó un grupo de datos de control por plegado computacional aleatorizado de 10.000 oligonucleótidos 29meros junto con los adaptadores 5' y 3' utilizando plegamiento ARN (Hofacker 2003). Se descubrió que los sitios de unión de la ARN ligasa 1 o las regiones flanqueantes no tenían una preferencia distintiva por ARN de cadena sencilla o de doble cadena (Figura 10A). Se generó la estructura secundaria de las 5.000 secuencias superiores leídas que incluían las secuencias de adaptador HD 3' y 5'. Se encontró un cambio en las estructuras secundarias de las regiones flanqueantes del sitio de unión. Las secuencias que generaron estructuras lazo de cadena sencilla y tallos dúplex 5' y 3' se secuenciaron preferentemente. Esto sugiere que la actividad de la ARN ligasa 1 prefiere los lazos de cadena sencilla, que es similar a su papel in vivo para la reparación del bucle escindido de ARNt-lys y por lo tanto apoya la validez de esta estrategia (Amitsur, Levitz et al. 1987).

Para analizar la preferencia de estructura secundaria de la primera unión que utiliza la T4 ARN ligasa 2 truncada se generó un grupo de datos de control por plegado computacional aleatorio de 10.000 oligonucleótidos 25meros junto con los adaptadores HD 3' utilizando plegamiento de ARN. Los sitios de unión de la T4 ARN ligasa 2 truncada no tenían una preferencia distintiva pro ARN de cadena sencilla o de doble cadena (Figura 10B). La estructura secundaria de las 5.000 secuencias superiores leídas incluyendo las secuencias adaptadoras 3' y 5' se generó entonces. Se encontró que las estructuras secundarias que flanqueaban la unión estaban distorsionadas a partir del grupo de datos aleatorizadas. Las secuencias que generaban estructuras dúplex 5' y las estructuras de cadena sencilla 30 con respecto al sitio de unión, y donde el sitio de unión se situaba en un lazo, se secuenciaban preferentemente. Esto sugiere que la actividad de la T4 ARN ligasa 2 truncada prefiere las secuencias dúplex 3' en vez de las regiones de cadena sencilla 5'. La creencia actual es que la ARN ligasa 2 sella las muescas en el ARN de doble cadena donde las regiones 5' y 3' del sitio de unión serían de doble cadena. Es ciertamente verdad que la ARN ligasa 2 puede unir eficazmente muescas en el ARN de doble cadena sin embargo se demostró inicialmente que puede unir secuencias enlazadas sencillas (Yin, Ho et al. 2003). Además, los análisis de los inventores están más en la línea de la función *in vivo* propuesta del complejo de edición de la ARN ligasa 2 de Tripanosoma que prefiere los restos de cadena sencilla en el sitio de unión (Cruz-Reyes, Zhelonkina et al. 2001).

El uso de oligonucleótidos degenerados para estudiar la función de las proteínas no es nuevo. Los oligonucleótidos aleatorios se utilizan en las estrategias SELEX para identificar ligandos para proteínas. Sin embargo esta estrategia necesita varias rondas de selección y enriquecimiento y no identifica determinantes de actividad (Tuerk and Gold

1990). Este es el primer trabajo que utiliza el método del Análisis Funcional por Secuenciación de la Próxima Generación (FANGS) para investigar directamente la función de una proteína. Se prevé que el método FANGS se puede utilizar para estudiar muchas otras proteínas que modifican ácidos nucleicos tales como transcriptasa inversa, cinasas de ácido nucleico y fosfatasas y quizá girasas.

La presente invención no se limita en el ámbito por los aspectos específicos y realizaciones descritos en el presente documento. Además, serán aparentes varias modificaciones de la invención además de las descritas en el presente documento para los expertos en la técnica a partir de la descripción anterior y las figuras adjuntas. Tales modificaciones se pretende que se encuentran en el alcance de las reivindicaciones adjuntas. Además, todos los aspectos y realizaciones descritas en el presente documento se consideran que son ampliamente aplicables y se pueden combinar con todos y cada uno de otros aspectos y realizaciones consistentes, si es apropiado.

**Referencias**


Amitsur, M., R. Levitz, et al. (1987). "Bacteriophage T4 anticodon nuclease, polynucleotide kinase and RNA ligase reprocess the host lysine tRNA." EMBO J 6(8): 2499-2503.

Benes, V. and M. Castoldi (2010). "Expression profiling of microRNA using real-time quantitative PCR, how to use it and what is available." Methods 50(4): 244-249.

Cruz-Reyes, J., A. Zhelonkina, et al. (2001). "Trypanosome RNA editing: simple guide RNA features enhance U deletion 100-fold." Mol Cell Biol 21(3): 884-892.

Dalmay, T. (2008). "Identification of genes targeted by microRNAs." Biochemical Society Transactions 36(part 6): 1194-1196.

Dohm, J. C., C. Lottaz, et al. (2008). "Substantial biases in ultra-short read data sets from high-throughput DNA sequencing." Nucleic acids research 36(16): e105.

Fernandez-Valverde, S. L., R. J. Taft, et al. (2010). "Dynamic isomiR regulation in Drosophila development." RNA 16(10): 1881.

Fire, A., S. Xu, et al. (1998). "Potent and specific genetic interference by doublestranded RNA in Caenorhabditis elegans." Nature 391 (6669): 806-811.

Guo, L. and Z. Lu (2010). "Global expression analysis of miRNA gene cluster and family based on isomiRs from deep sequencing data." Computational Biology and Chemistry 34(3): 165-171.

Hafner, M., et al. (2011). "RNA-ligase-dependent biases in miRNA representation in deep-sequenced small RNA cDNA libraries." RNA Jul. 20 [Epub ahead of print].

Hansen, K. D., S. E. Brenner, et al. (2010). "Biases in Illumina transcriptome sequencing caused by random hexamer priming." Nucleic acids research 38(12): e131.

Harismendy, O., P. C. Ng, et al. (2009). "Evaluation of next generation sequencing platforms for population targeted sequencing studies." Genome Biol 10(3): R32.

Hoebeeck, J., R. van der Luijt, et al. (2005). "Rapid detection of VHL exon deletions using real-time quantitative PCR." Lab Invest 85(1): 24-33.

Hofacker, I. L. (2003). "Vienna RNA secondary structure server." Nucleic Acids Res 31 (13): 3429-3431.

Linsen, S. E., E. de Wit, et al. (2009). "Limitations and possibilities of small RNA digital gene expression profiling." Nat Methods 6(7): 474-476.

McCormick, K. P., M. R. Willmann, et al (2011). "Experimental design, preprocessing, normalization and differential expression analysis of small RNA sequencing experiments." Silence 2(1): 2.

Moxon, S. et al (2008). "A toolkit for analysing large-scale plant small RNA datasets." Bioinformatics 24(19): 2252-2253

Nandakumar, J., C. K. Ho, et al. (2004). "RNA substrate specificity and structure-guided mutational analysis of bacteriophage T4 RNA ligase 2." J Biol Chem 279(30): 31337-31347.

Nandakumar, J. and Shuman S. (2004). "How an RNA ligase discriminates RNA versus DNA damage." Mol Cell. 16(2): 211-21

Oshlack, A. and M. J. Wakefield (2009). "Transcript length bias in RNA-seq data confounds systems biology." Biology Direct 4(1): 14.

Starega-Roslan, J., J. Krol, et al. (2011). "Structural basis of microRNA length variety." Nucleic acids research 39(1): 257.

Tian, G., X. Yin, et al. (2010) "Sequencing bias: comparison of different protocols of microRNA library construction." BMC Biotechnol 10: 64.

Tuerk, C. and L. Gold (1990). "Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase." Science 249(4968): 505-510.

Voinnet, O. (2002). "RNA silencing: small RNAs as ubiquitous regulators of gene expression." Curr Opin Plant Biol 5(5): 444-451.

Willenbrock, H., J. Salomon, et al. (2009). "Quantitative miRNA expression analysis: comparing microarrays with next-generation sequencing." RNA 15(11): 2028-2034.

Yin, S., C. K. Ho, et al. (2003). "Structure-function analysis of T4 RNA ligase 2." J Biol Chem 278(20): 17601-17608.

LISTADO DE SECUENCIAS

<110> University of East Anglia

<120> ANÁLISIS DE TENDENCIA EN SECUENCIACIÓN

<130> SCB/P115491WO00

<150> GB1113214.9
<151> 29-07-2011

<160> 11

<170> PatentIn versión 3.5

<210> 1
<211> 25
<212> ADN
<213> Secuencia artificial

<220>
<223> Adaptador HD

<400> 1
gdgatcgtat gccgtcttct gcttg 25

<210> 2
<211> 25
<212> ADN
<213> Secuencia artificial

<220>
<223> Adaptador HD

<400> 2
attgtcgtat gccgtcttct gcttg 25

<210> 3
<211> 30
<212> ADN
<213> Secuencia artificial

<220>
<223> Adaptador 5'

<220>
<221> misc_feature
<222> (27)..(30)
<223> n es cualquier base de ARN

<400> 3
gttcagagtt ctacagtccg acgatcnnnn 30

<210> 4
<211> 26
<212> ADN
<213> Secuencia artificial

<220>
<223> Oligonucleótido complementario

<400> 4
gatcgtcgga ctgtagaact ctgaac 26

<210> 5
<211> 28

<212> ADN
<213> Secuencia artificial

<220>
<223> Adaptador 3'

<220>
<221> misc_feature
<222> (1)..(4)
<223> n es cualquier base de ARN.

<400> 5
nnnnatctcg tatgccgtct tctgcttg 28

<210> 6
<211> 24
<212> ADN
<213> Secuencia artificial

<220>
<223> Oligonucleótido complementario

<400> 6
caagcagaag acggcatacg agat 24

<210> 7
<211> 25
<212> ADN
<213> Secuencia artificial

<220>
<223> Oligonucleótido sintético

<220>
<221> misc_feature
<223> n es cualquier base

<220>
<221> misc_feature
<222> (5)..(24)
<223> n es a, c, g o t

<220>
<221> misc_feature
<222> (25)..(25)
<223> n es una secuencia fija

<400> 7
caaannnnnn nnnnnnnnnn nnnnn 25

<210> 8
<211> 26
<212> ARN
<213> Secuencia artificial

<220>
<223> Adaptador de ligamiento

<400> 8
guucagaguu cuacaguccg acgauc 26

<210> 9
<211> 25
<212> ADN
<213> Secuencia artificial

<220>
<223> Adaptador de ligamiento

<220>
<221> misc_feature
<222> (1)..(1)
<223> Difosfato de adenosina

<220>
<221> misc_feature
<222> (2)..(2)
<223> Ribonucleótido

<400> 9
aatctcgtat gccgtcttct gcttg 25

<210> 10
<211> 30
<212> ADN
<213> Secuencia artificial

<220>
<223> Adaptador HD

<220>
<221> misc_feature
<222> (27)..(27)
<223> n es cualquier base

<220>
<221> misc_feature
<222> (28)..(30)
<223> n es cualquier base de ARN

<400> 10
gttcagagtt ctacagtccg acgatcnnnn 30

<210> 11
<211> 26
<212> ADN
<213> Secuencia artificial

<220>
<223> Adaptador HD

<220>
<221> misc_feature
<222> (1)..(1)
<223> Difosfato de adenosina

<220>
<221> misc_feature
<222> (2)..(4)
<223> n es cualquier base de ARN

<220>
<221> misc_feature
<222> (5)..(5)
<223> n es cualquier base

<400> 11
annnntcgta tgccgtcttc tgcttg 26

**REIVINDICACIONES**

1. Un método para reducir la tendencia de secuencia de una técnica de secuenciación que implica la unión con un adaptador, comprendiendo el método:

(a) proporcionar un grupo de oligonucleótidos de cadena sencilla de secuencia conocida con extremos 3' bloqueados (moléculas adaptadoras 3') y un grupo de oligonucleótidos de cadena sencilla de secuencia conocida con extremos 5' bloqueados (moléculas adaptadoras 5'), donde las moléculas 3' y 5' comprenden uno o más nucleótidos degenerados;
(b) unir las moléculas adaptadoras 3' a los extremos 3' de las moléculas de ácido nucleico diana utilizando una ligasa;
(c) unir las moléculas adaptadoras 5' a los extremos 5' de las moléculas de ácido nucleico diana de la etapa (b) utilizando una ligasa; y
(d) determinar la secuencia de las moléculas de ácido nucleico diana que se obtienen en la etapa (c) utilizando un cebador capaz de hibridarse con el complemento de la molécula adaptadora 5' y un cebador capaz de hibridarse con la molécula adaptadora 3'.

2. Un método para detectar una molécula de ácido nucleico diana en un biblioteca de moléculas de ácidos nucleicos, comprendiendo el método:

(a) proporcionar un grupo de oligonucleótidos de cadena sencilla de secuencia conocida con extremos 3' bloqueados (moléculas adaptadoras 3') y un grupo de oligonucleótidos de cadena sencilla de secuencia conocida con extremos 5' bloqueados (moléculas adaptadoras 5'), donde las moléculas 3' y 5' comprenden uno o más nucleótidos degenerados y se pueden unir a la molécula de ácido nucleico diana;
(b) unir la molécula adaptadora 3' a los extremos 3' de las moléculas de ácido nucleico de la biblioteca de moléculas de ácido nucleico utilizando una ligasa;
(c) unir la molécula adaptadora 5' a los extremos 5' de moléculas de ácido nucleico de la etapa (b) utilizando una ligasa;
(d) crear una biblioteca amplificada de moléculas de ácido nucleico por PCR de las moléculas de ácido nucleico que se obtienen en la etapa (c) utilizando un cebador capaz de hibridarse con el complemento de la molécula adaptadora 5' y un cebador capaz de hibridarse con la molécula adaptadora 3';
(e) secuenciar la biblioteca amplificada resultante de moléculas de ácido nucleico; y
(f) analizar los resultados de la secuenciación para determinar si la molécula de ácido nucleico diana está presente en la biblioteca de moléculas de ácido nucleico.

3. El método de acuerdo con la reivindicación 2, donde el ácido nucleico diana se asocia con una enfermedad o estado de pre-enfermedad.

4. El método de acuerdo con la reivindicación 2, donde el ácido nucleico diana se asocia con un organismo en particular.

5. El método de acuerdo con la reivindicación 2, donde el ácido nucleico diana se asocia con un tipo de tejido en particular.

6. El método de acuerdo con la reivindicación 2, donde el ácido nucleico diana se asocia con un estado de desarrollo en particular.

7. El método de acuerdo con una cualquiera de las reivindicaciones 1-6, donde las moléculas de ácido nucleico son moléculas de ARN.

8. El método de acuerdo con una cualquiera de las reivindicaciones 1-6, donde las moléculas de ácido nucleico son moléculas de ADN.

9. Un método para generar una biblioteca de ADNc a partir de una biblioteca de moléculas de ARN, comprendiendo el método:

(a) proporcionar un grupo de oligonucleótidos de cadena sencilla de secuencia conocida con extremos 3' bloqueados (moléculas adaptadoras 3') y un grupo de oligonucleótidos de cadena sencilla de secuencia conocida con extremos 5' bloqueados (moléculas adaptadoras 5'), donde las moléculas 3' y 5' comprenden uno o más nucleótidos degenerados;
(b) unir las moléculas adaptadoras 3' a los extremos 3' de las moléculas de ácido nucleico diana utilizando una ligasa;
(c) unir las moléculas adaptadoras 5' a los extremos 5' de las moléculas de ácido nucleico diana de la etapa (b) utilizando una ligasa;
(d) crear moléculas de estructura híbrida ARN/ADN a partir de moléculas de ARN que se obtienen en la etapa (c) utilizando una enzima transcriptasa inversa y un cebador capaz de hibridarse con la molécula adaptadora 3'; y

(e) crear una biblioteca de ADNc por PCR de las moléculas de estructura híbrida ARN/ADN que se obtienen en la etapa (d) utilizando un cebador capaz de hibridarse con el complemento de la molécula adaptadora 5' y un cebador capaz de hibridarse con la molécula adaptadora 3'.

10. Un grupo de oligonucleótidos de cadena sencilla de secuencia conocida con extremos 3' bloqueados (moléculas adaptadoras 3') y un grupo de oligonucleótidos de cadena sencilla de secuencia conocida con extremos 5' bloqueados (moléculas adaptadoras 5'), donde las moléculas adaptadoras 3' y 5' comprenden uno o más nucleótidos degenerados, para su uso en un método como se define en una cualquiera de las reivindicaciones 1-9.

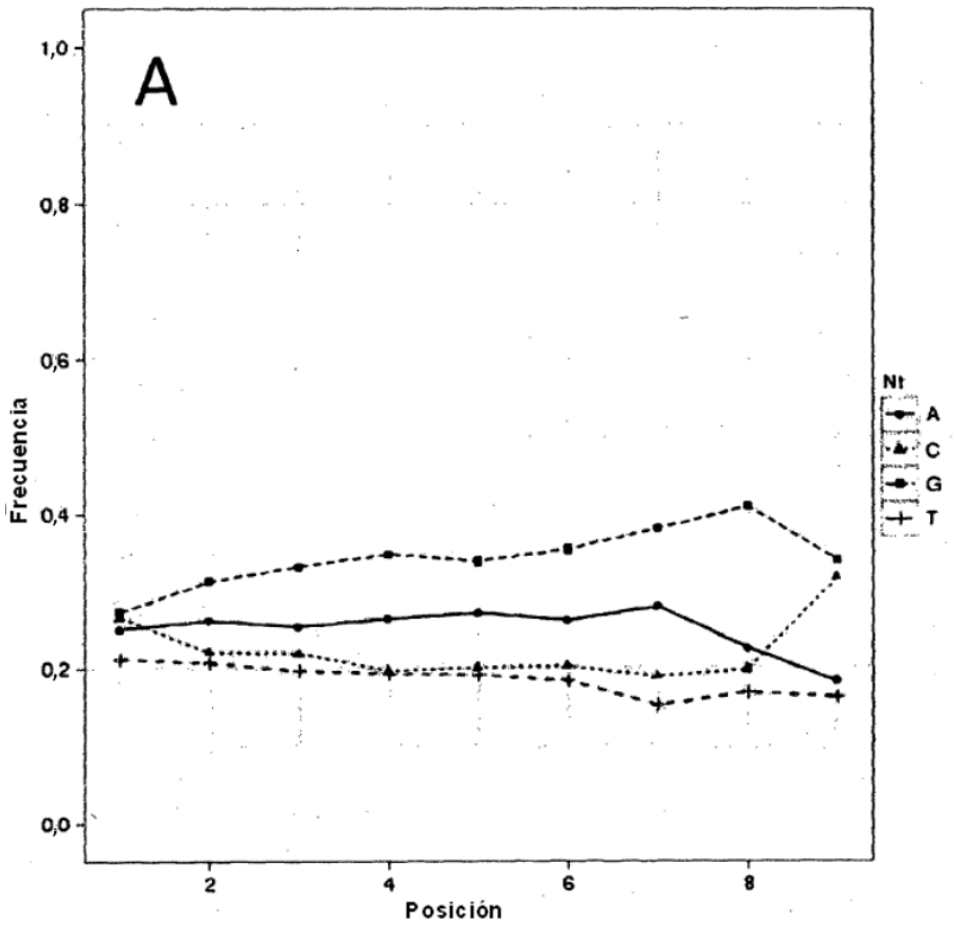



Figura 1 A.

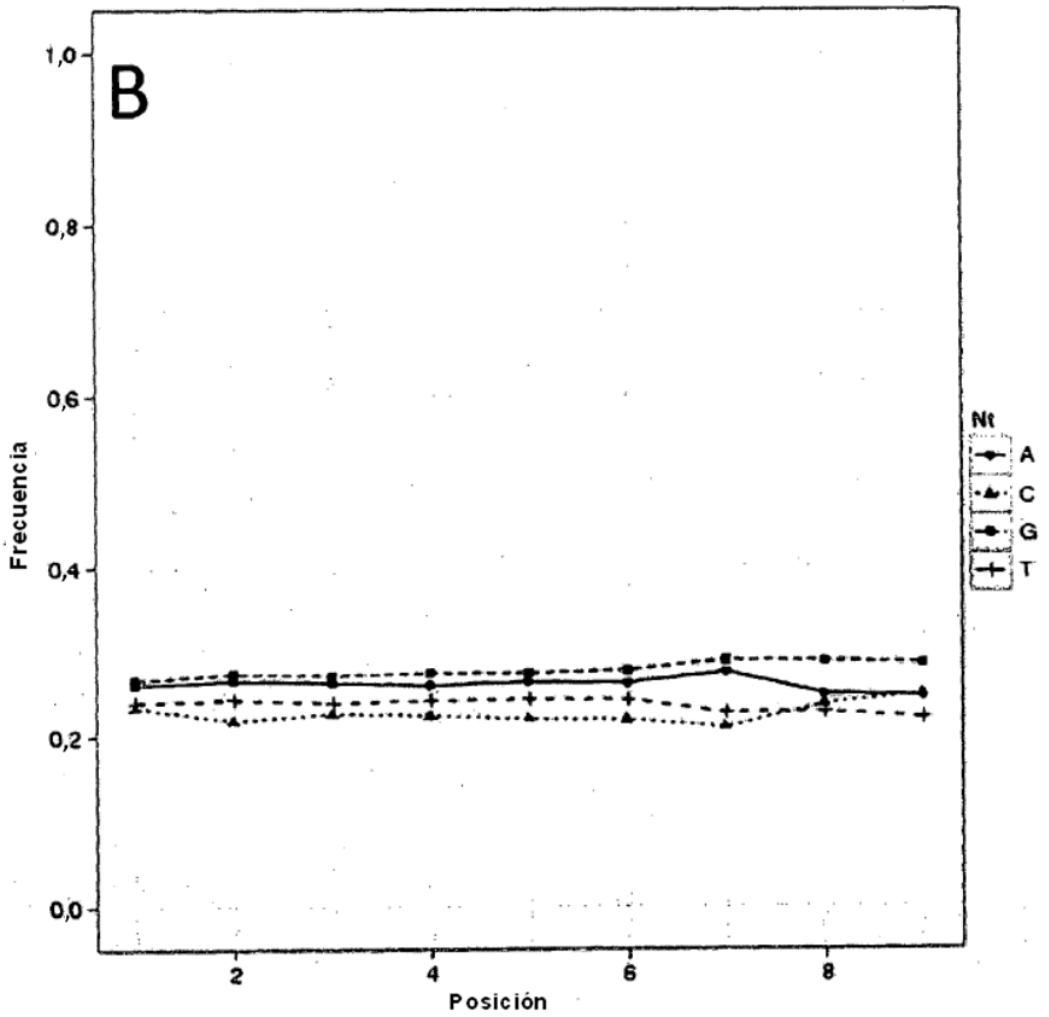



Figura 1 B.

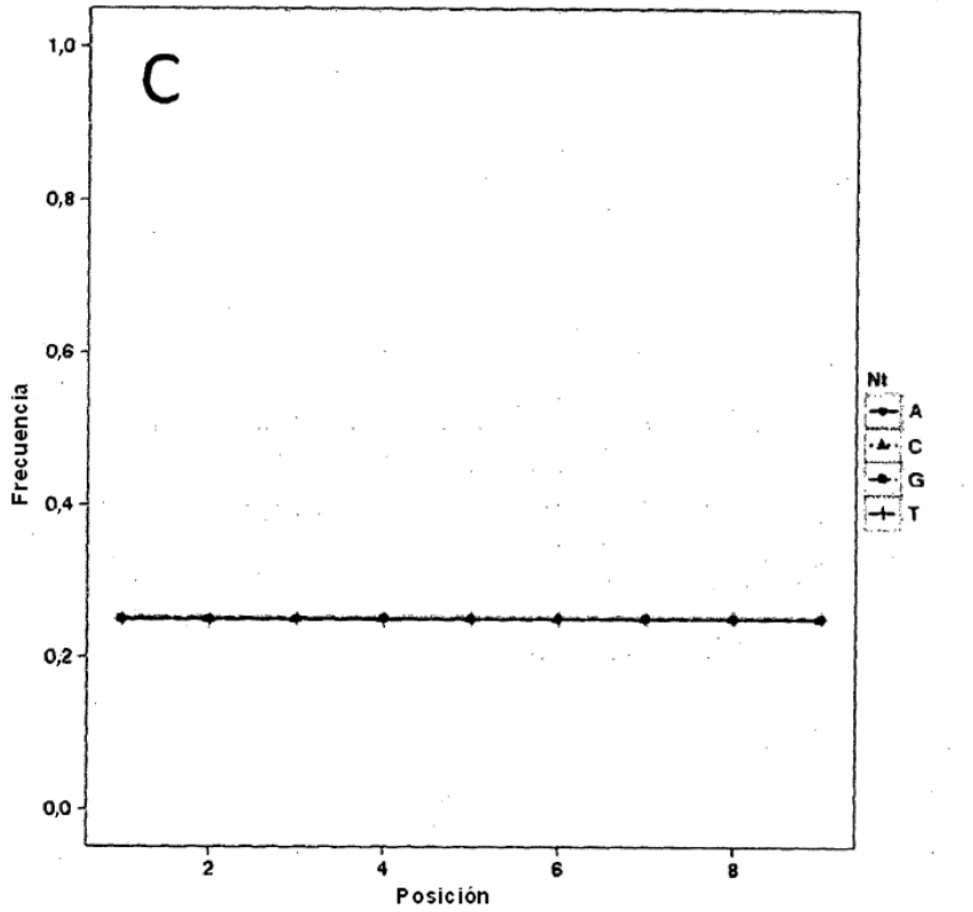



Figura 1 C.

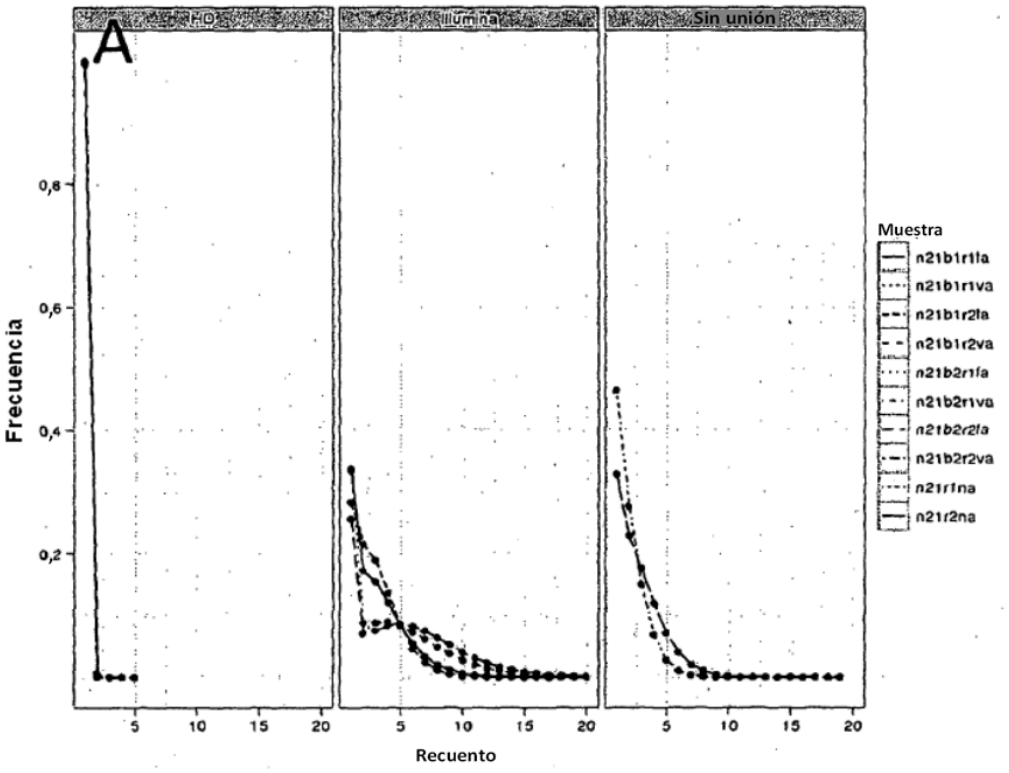

A
HQ
Illumina
Sin unión
Frecuencia
0,8
0,6
0,4
0,2
5
10
15
20
Recuento
Muestra
n21b1r1fa
n21b1r1va
n21b1r2fa
n21b1r2va
n21b2r1fa
n21b2r1va
n21b2r2fa
n21b2r2va
n21r1na
n21r2na

Figura 2 A.

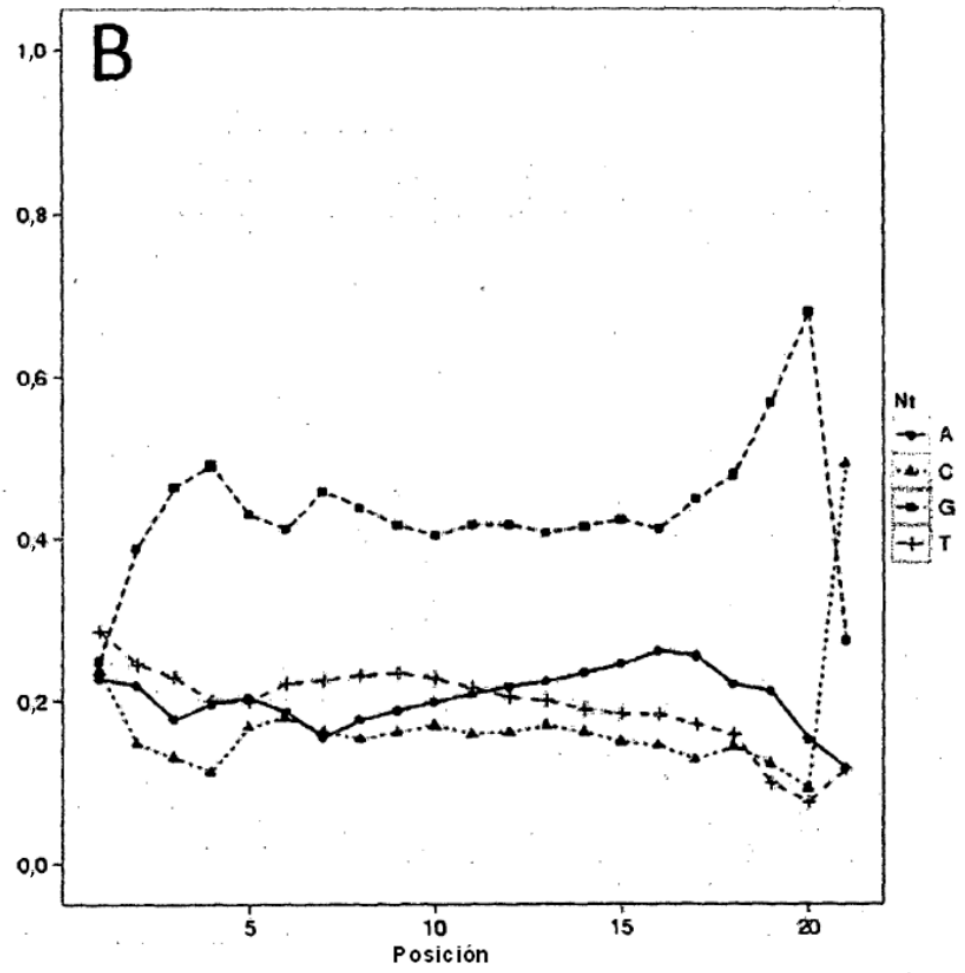



Figura 2 B.

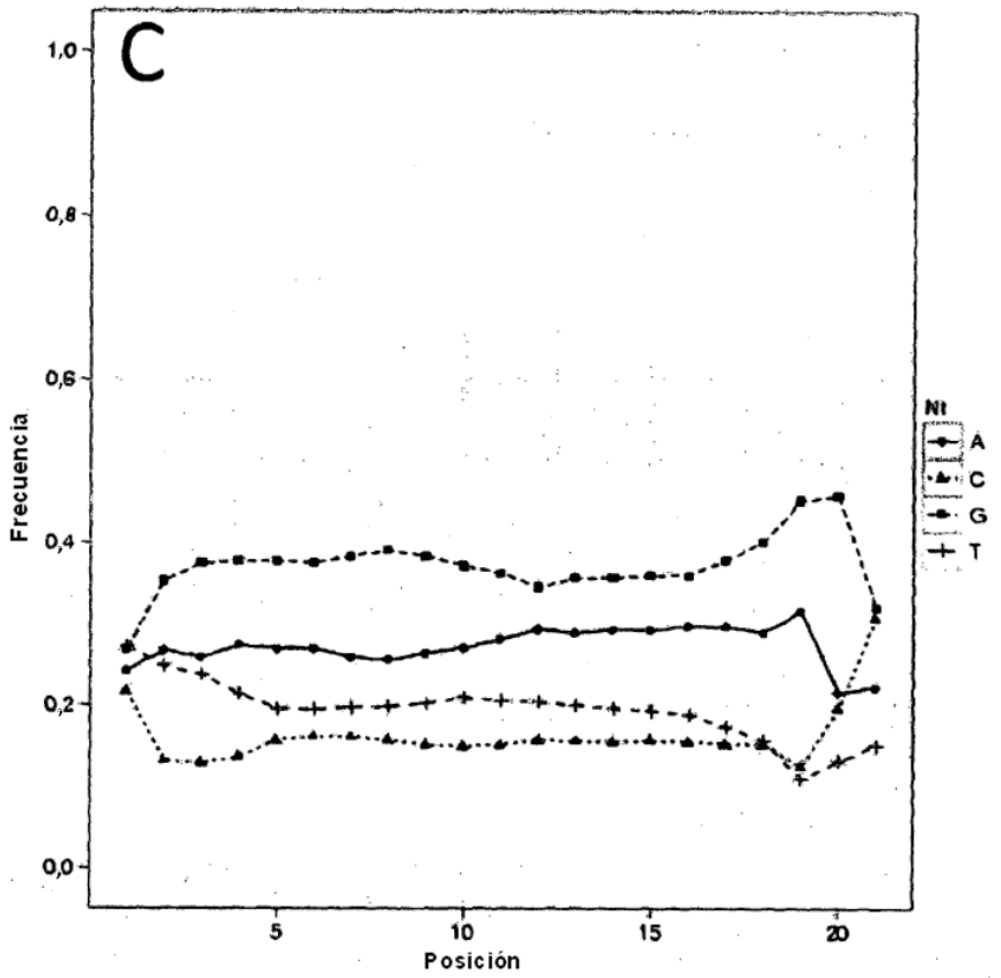



Figura 2 C.

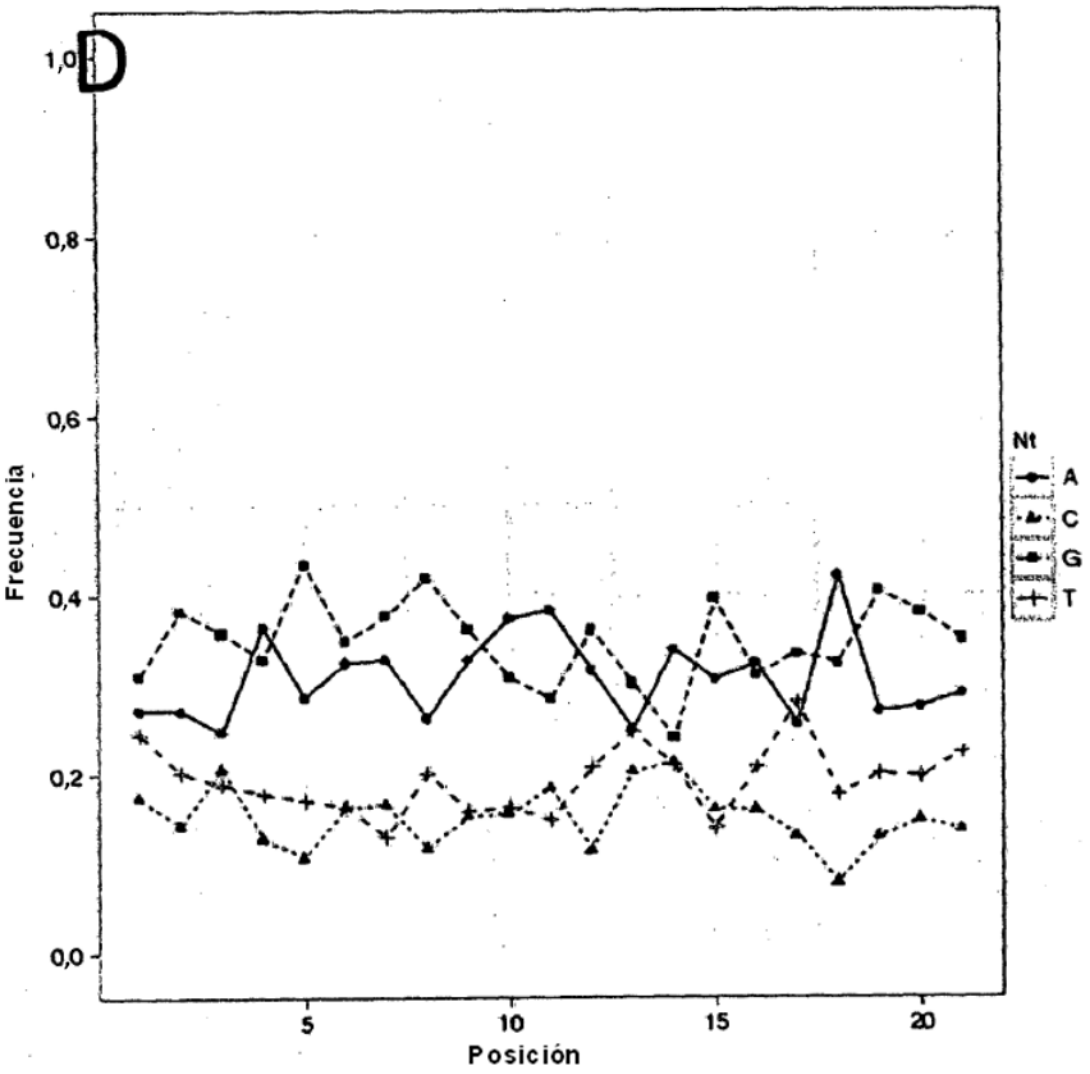



Figura 2 D.

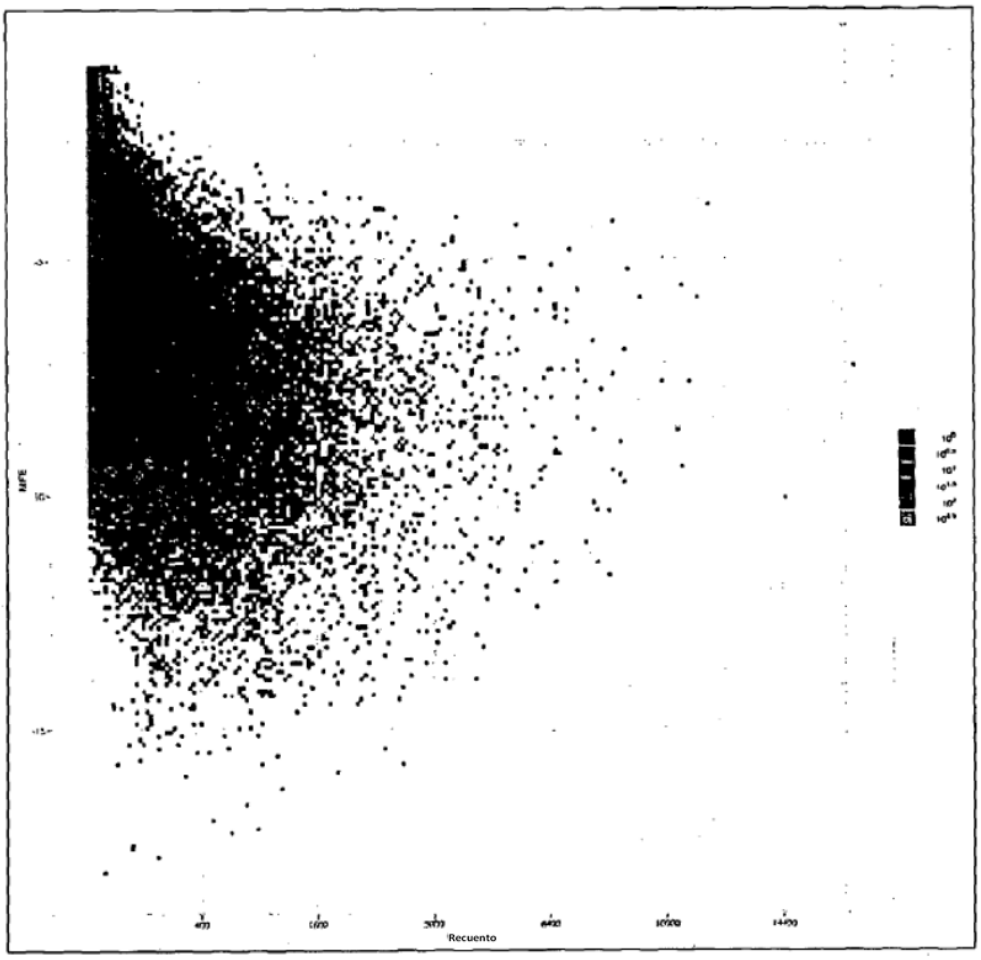



Figura 3.

**GAGA**TCGTATGCCGTCTTCTGCTTG (SEC ID Nº 1)

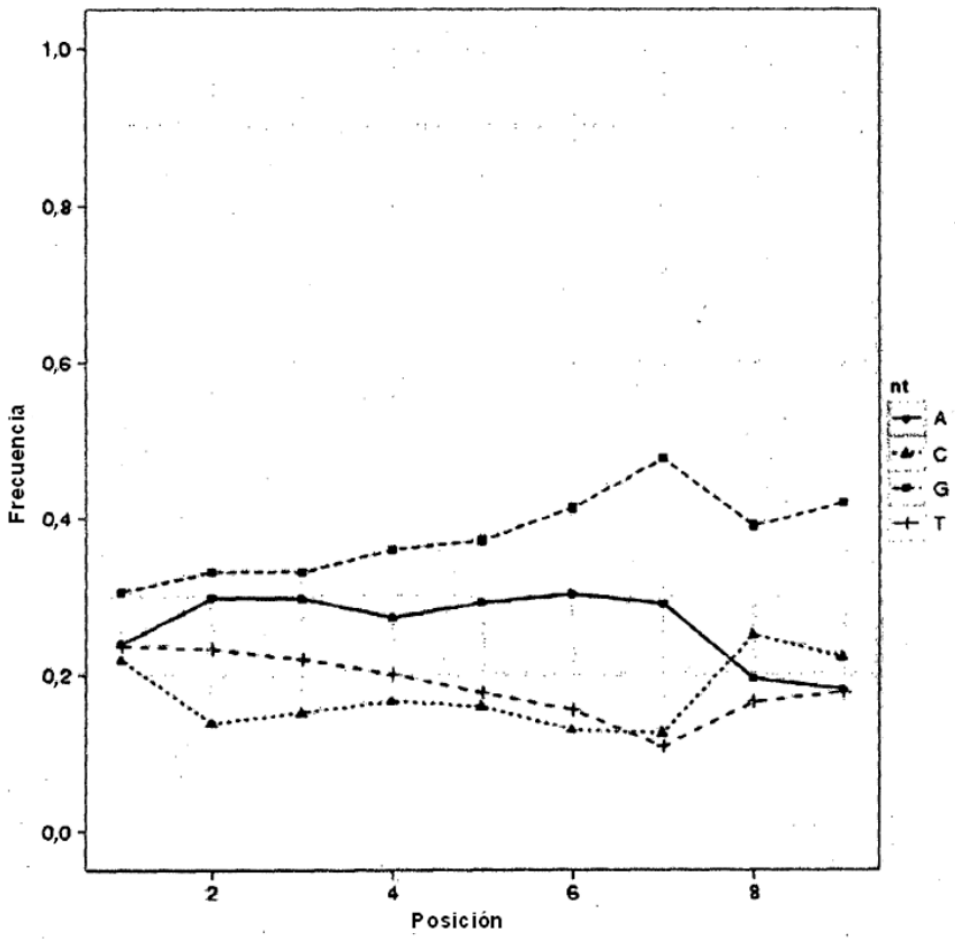



Figura 4.

**ATTG**TCGTATGCCGTCTTCTGCTTG (SEC ID Nº 2)

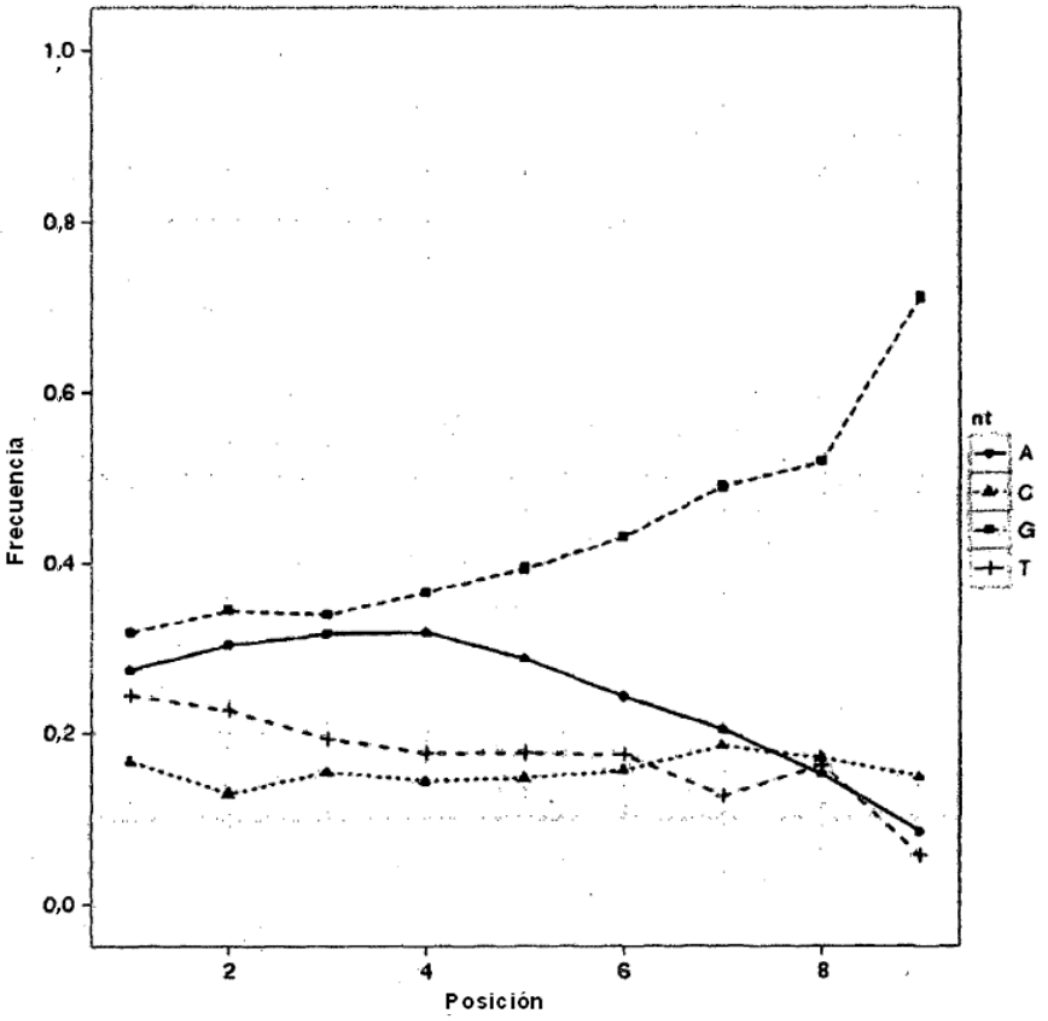



Figura 4.

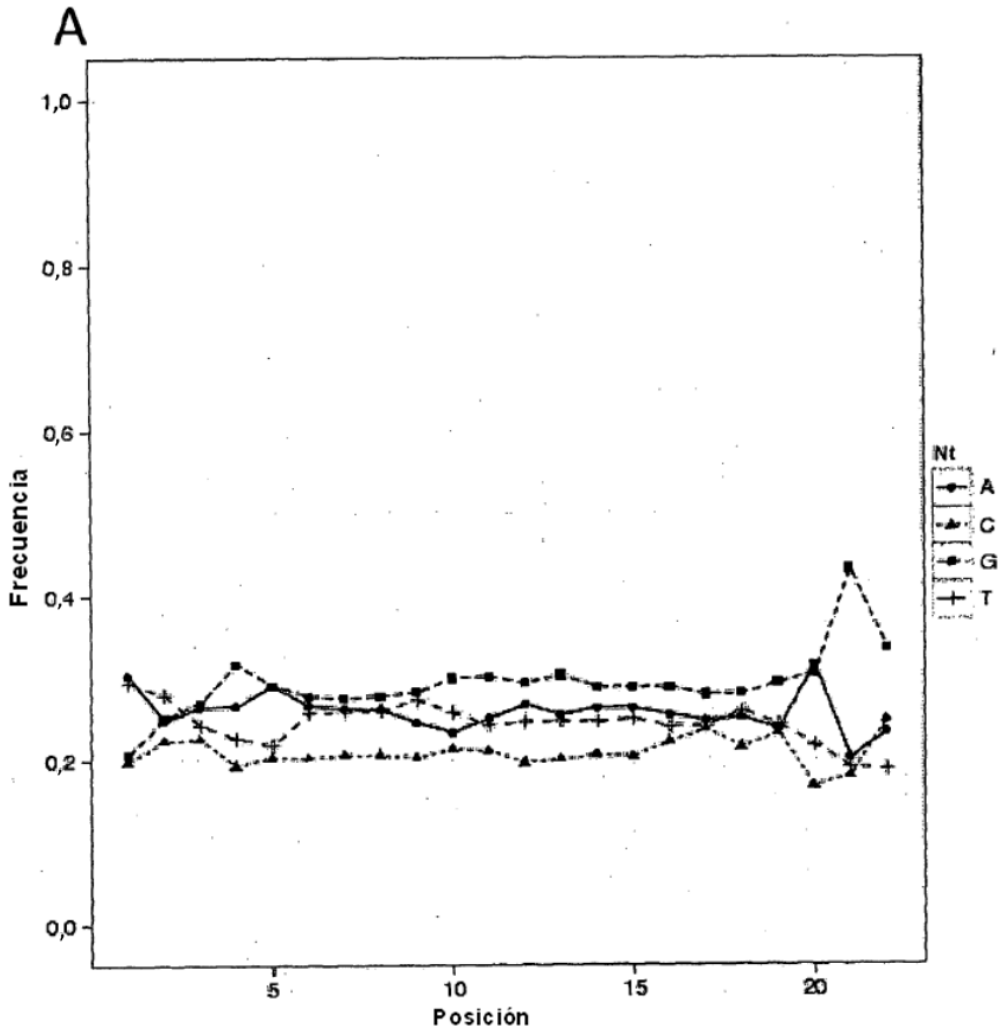



Figura 5 A.

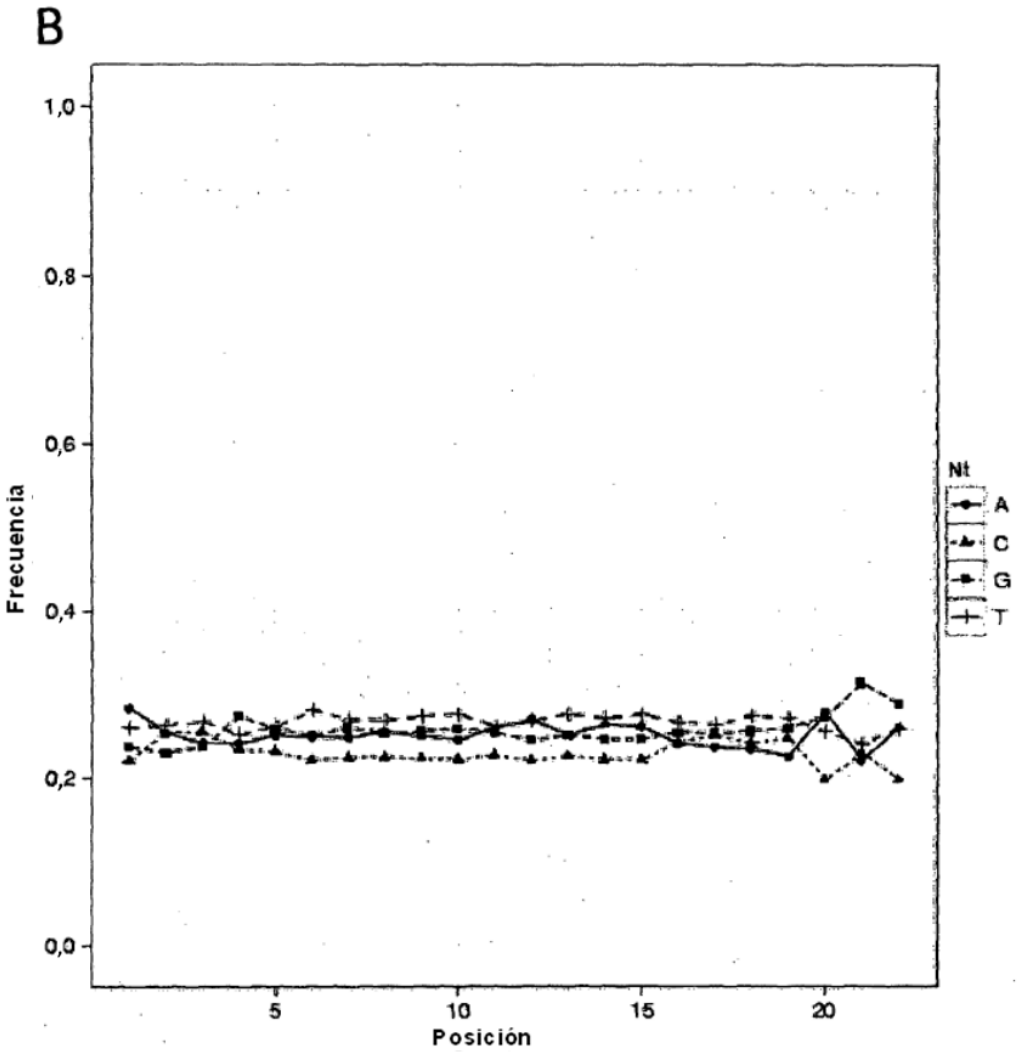



Figura 5 B.

C

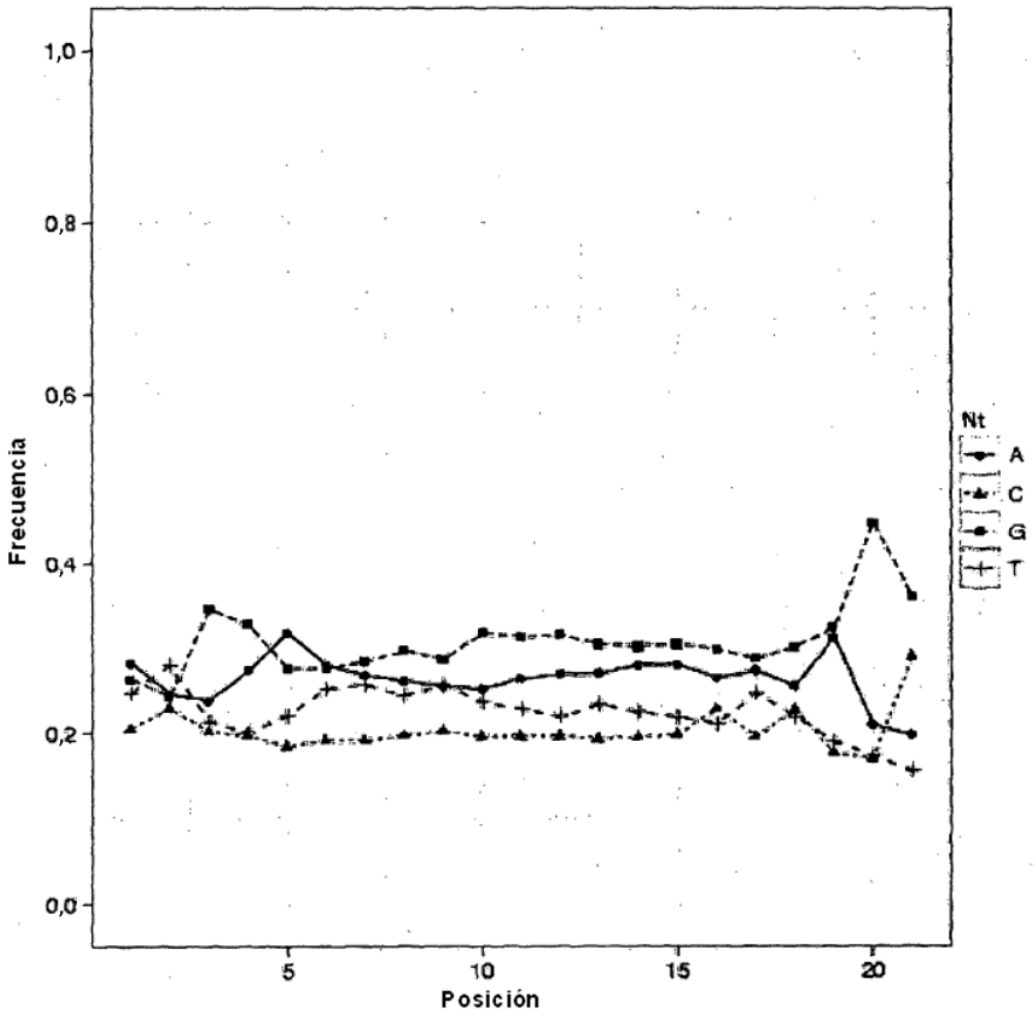



Figura 5 C.

D

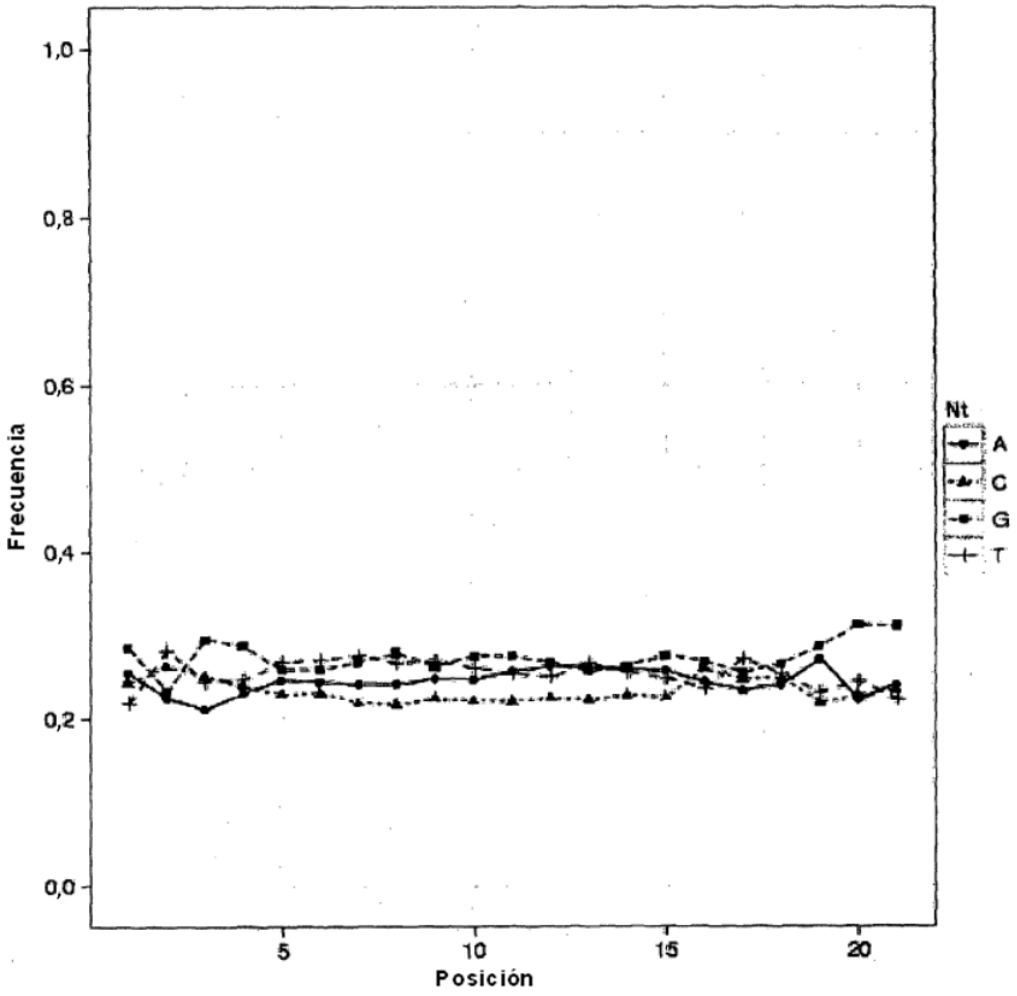



Figura 5 D.

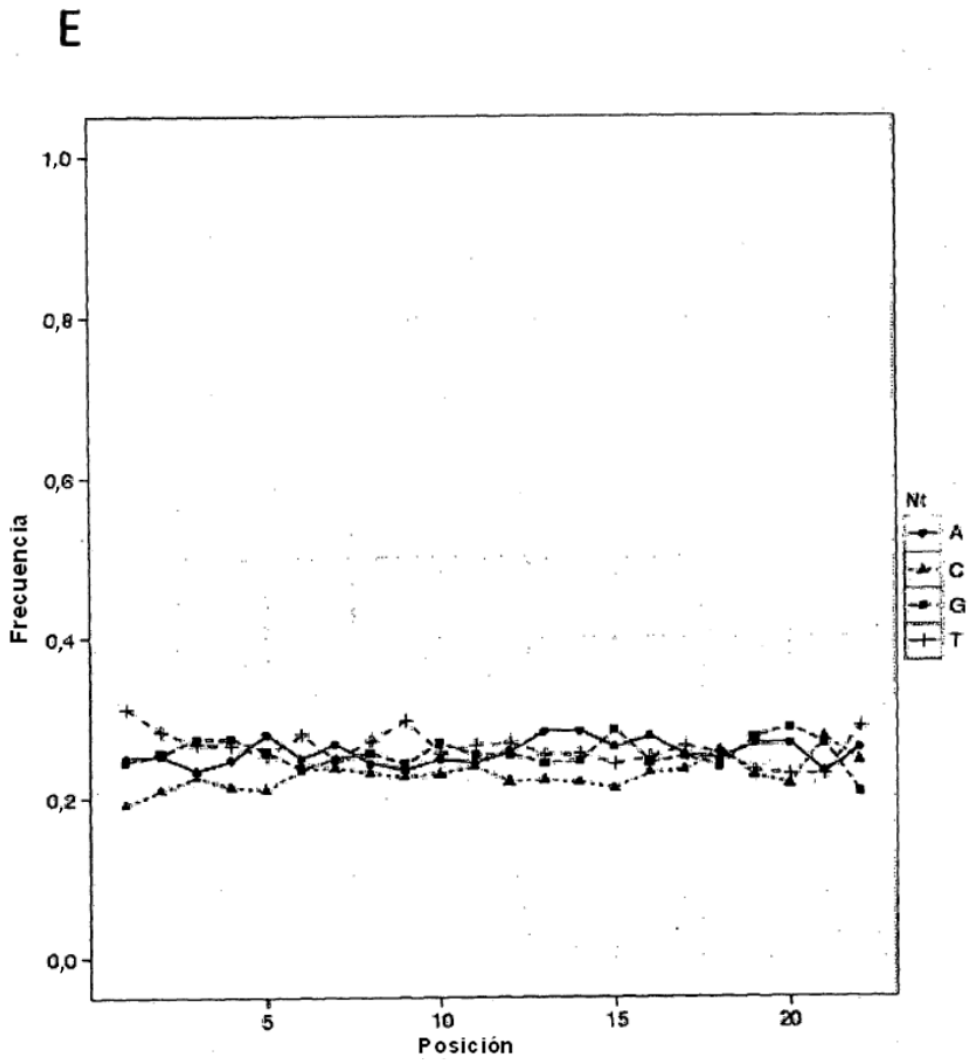



Figura 5 E.

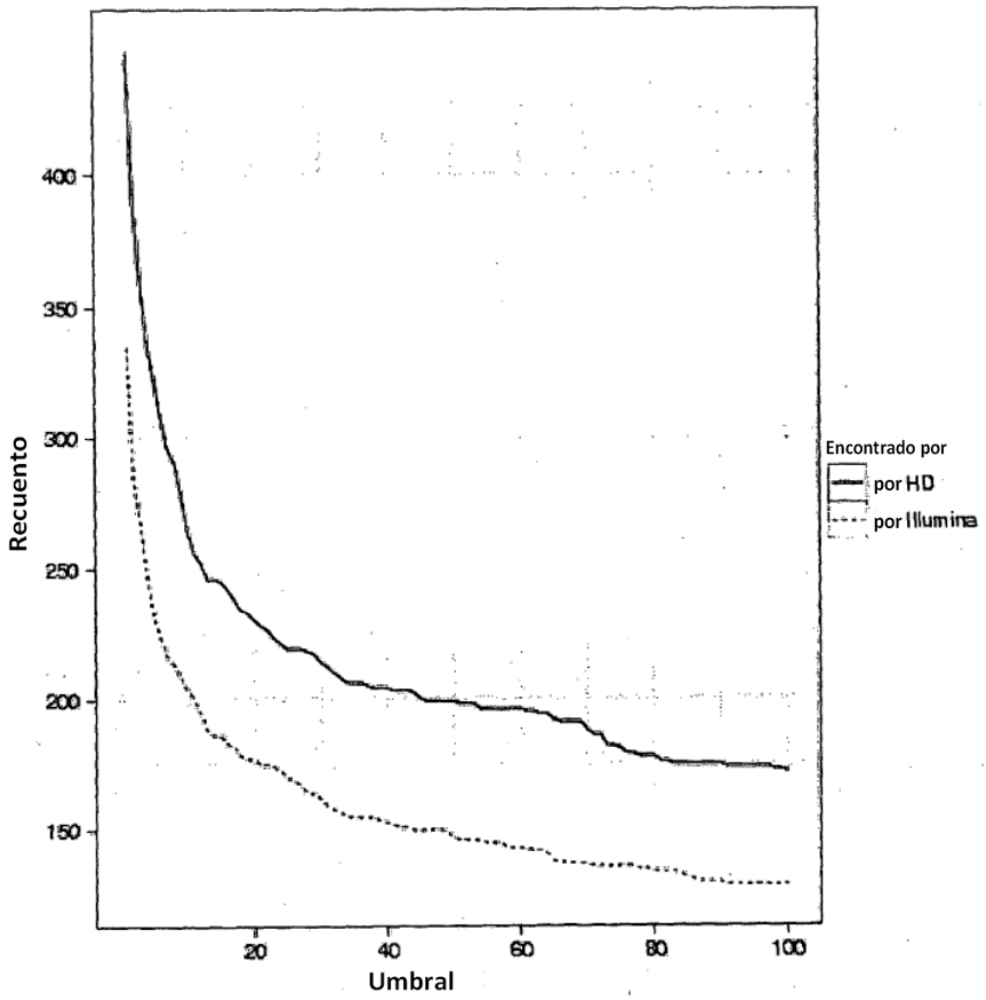
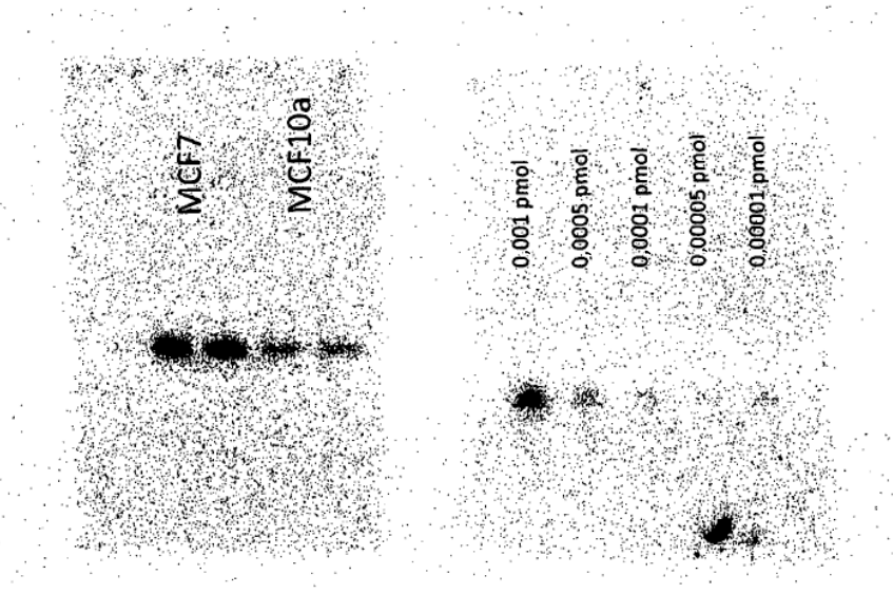
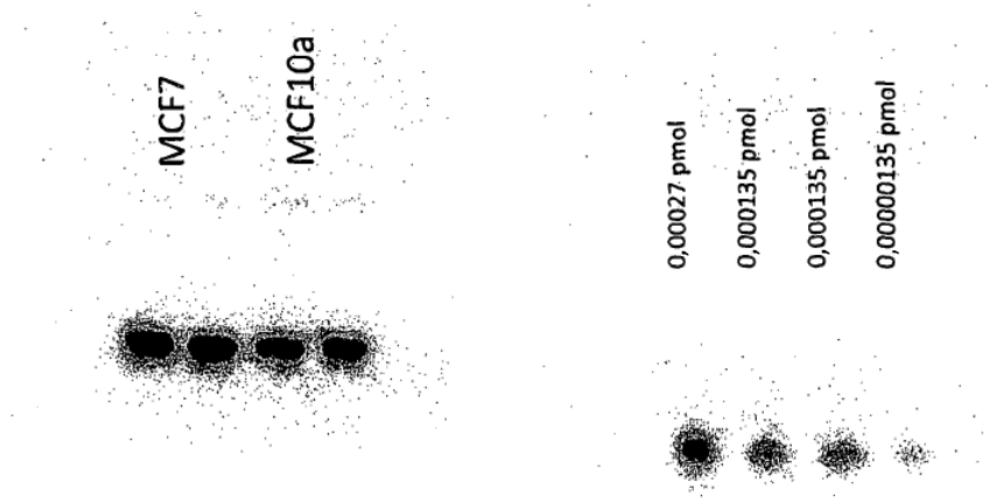

Recuento
400
350
300
250
200
150
20
40
60
80
100
Umbral
Encontrado por
por HD
por Illumina

Figura 6.




A. Mir-25

Figura 7A.




B. Mir-103

Figura 7B.

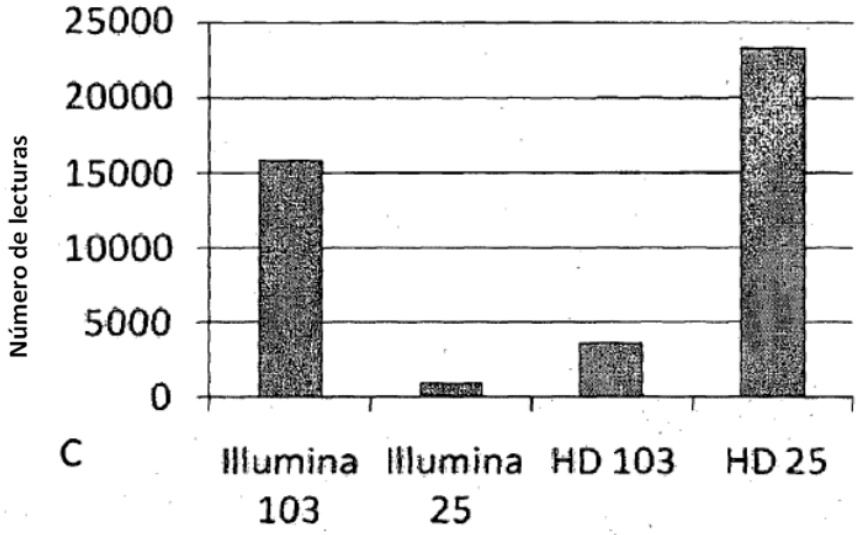



Figura 7C.

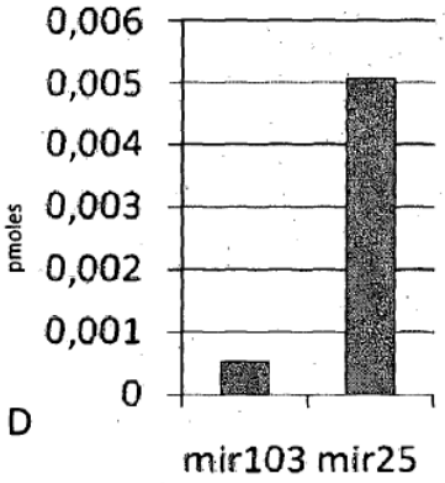



Figura 7D.

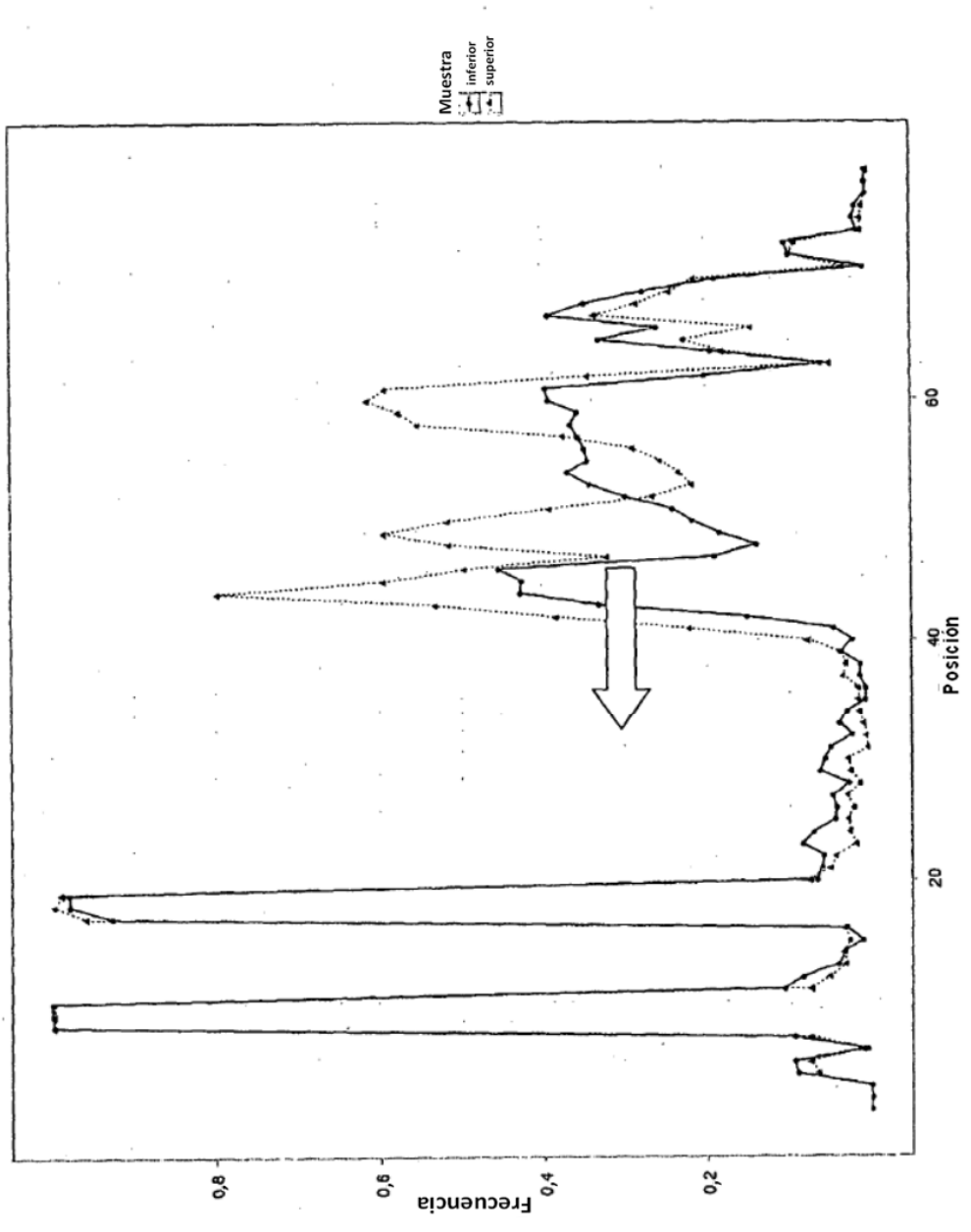

Muestra
inferior
superior
Frecuencia
0,8
0,6
0,4
0,2
20
40
60
Posición

Figura 8 A

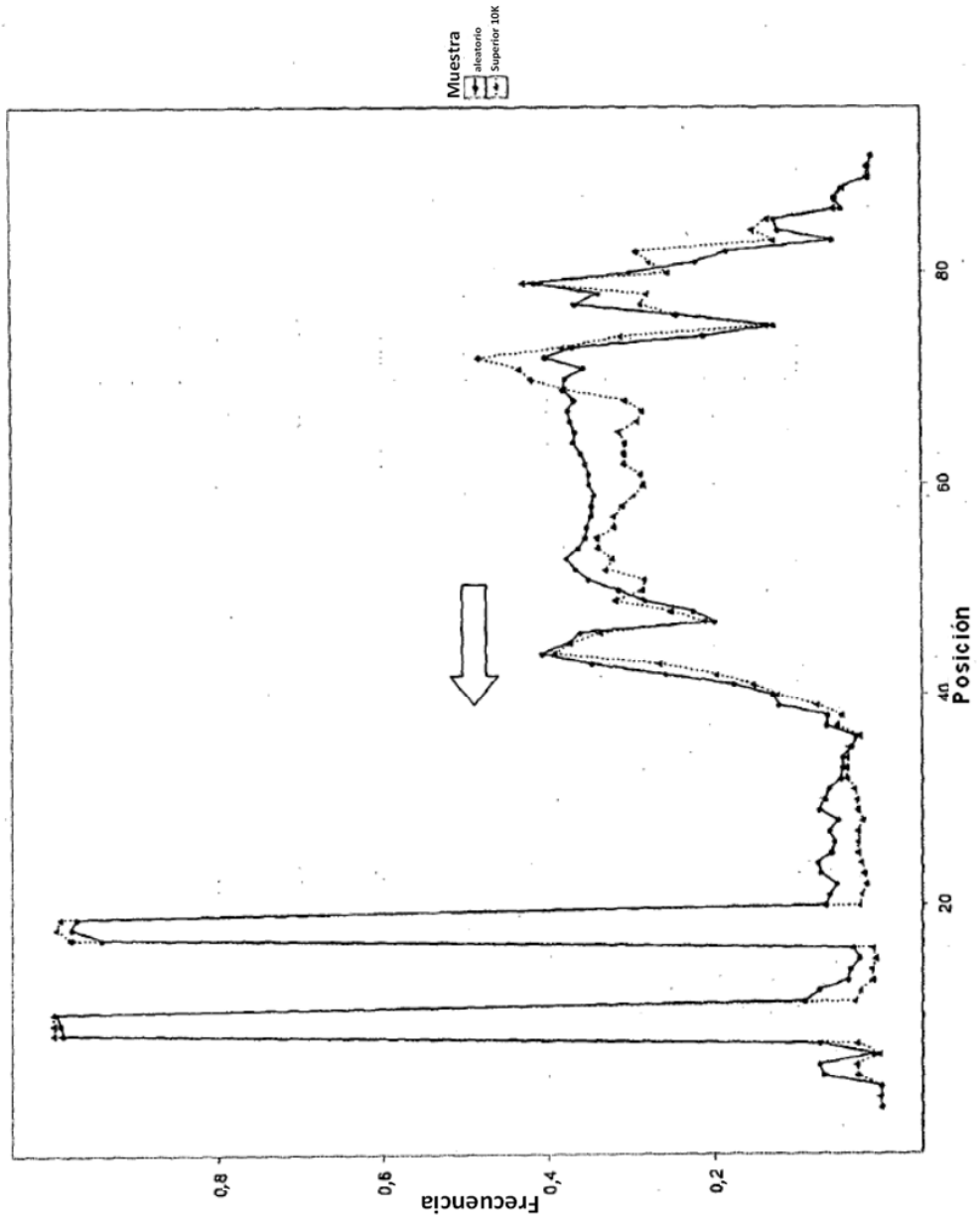

Muestra
aleatorio
Superior 10K
Frecuencia
0,8
0,6
0,4
0,2
20
40
60
80
Posición

Figura 8 B

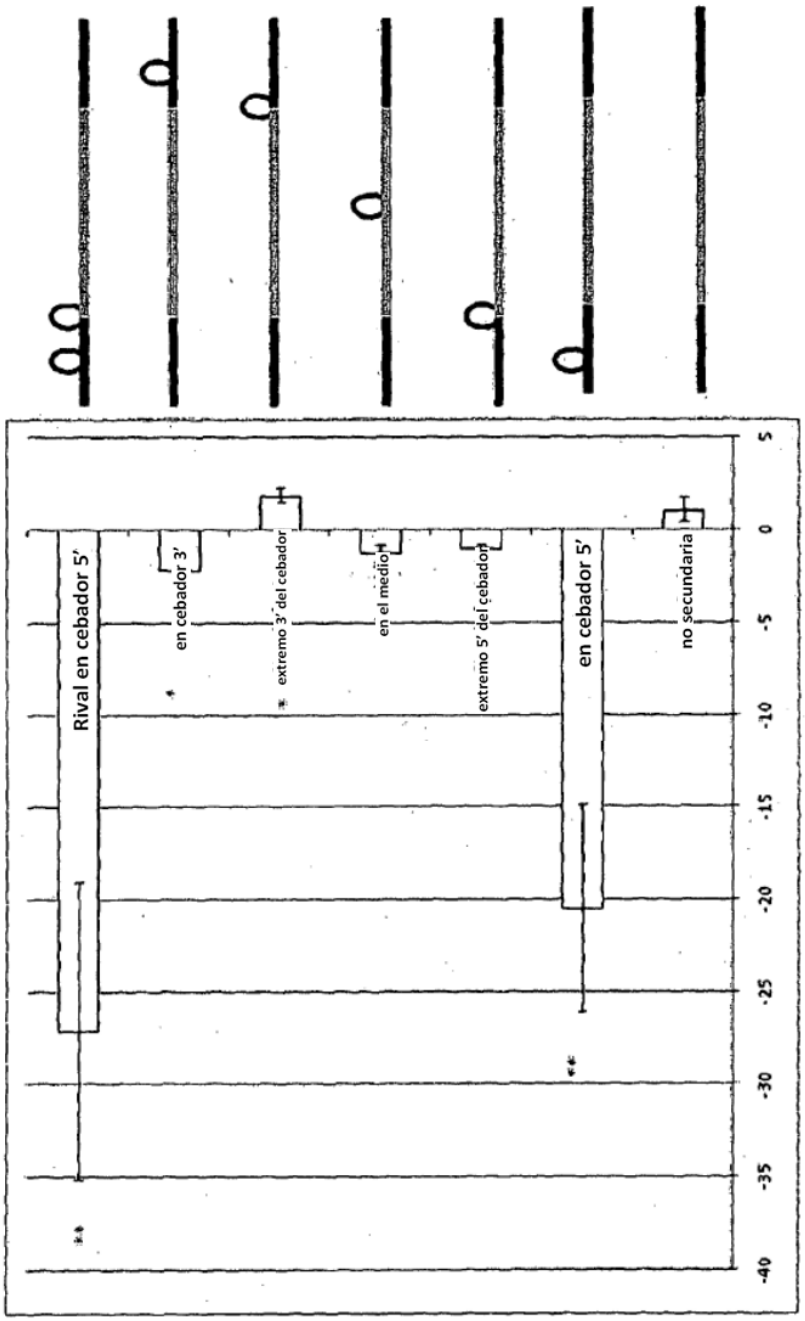

Rival en cebador 5'
en cebador 3'
extremo 3' del cebador
en el medio
extremo 5' del cebador
en cebador 5'
no secundaria
5
0
-5
-10
-15
-20
-25
-30
-35
-40
*
**
*=P<0,05
**=p<0,005

Figura 9

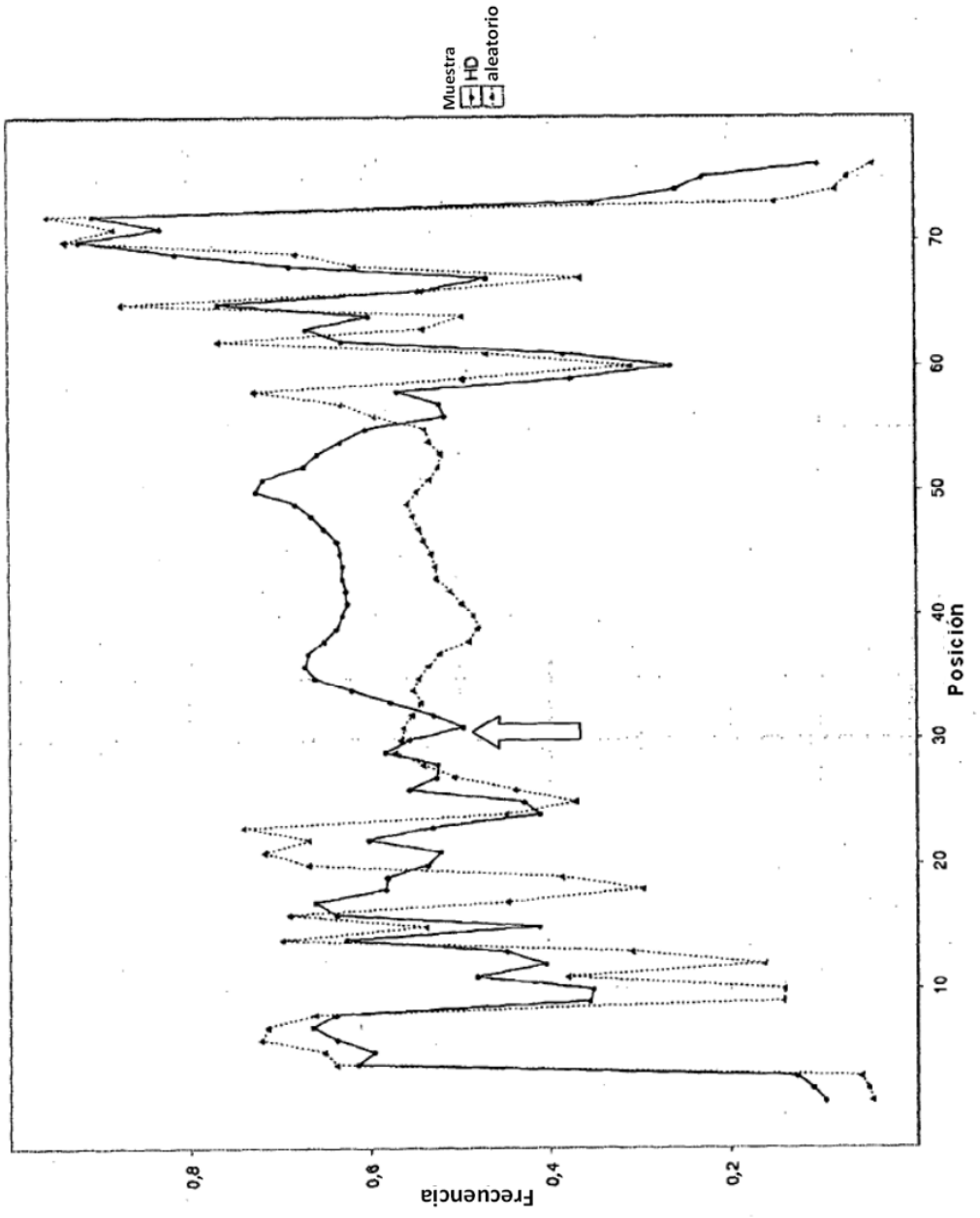

Muestra
HD
aleatorio
Frecuencia
0,2
0,4
0,6
0,8
Posición
10
20
30
40
50
60
70

Figura 10 A

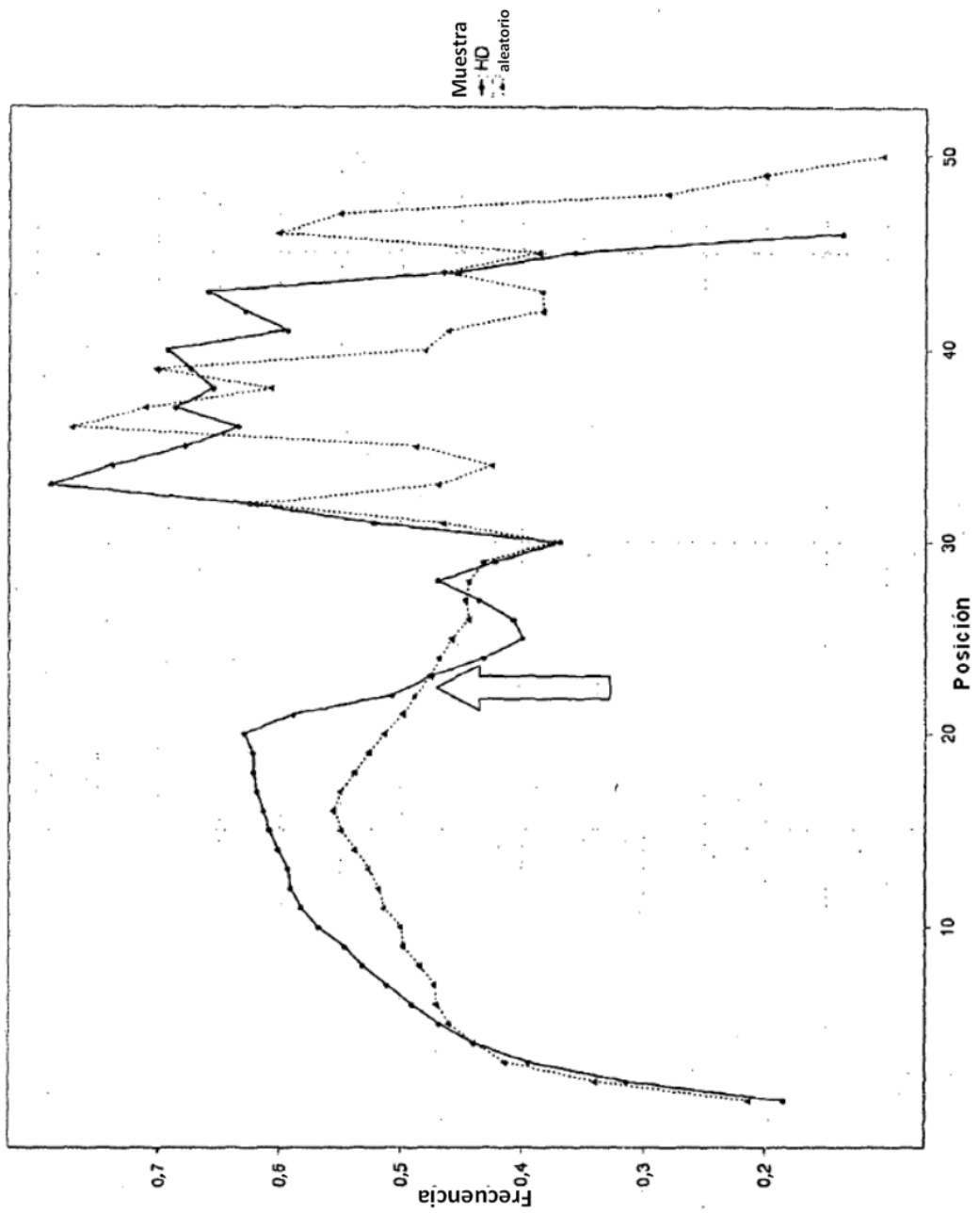

Muestra
HD
aleatorio
Frecuencia
0,7
0,6
0,5
0,4
0,3
0,2
Posición
10
20
30
40
50

Figura 10 B

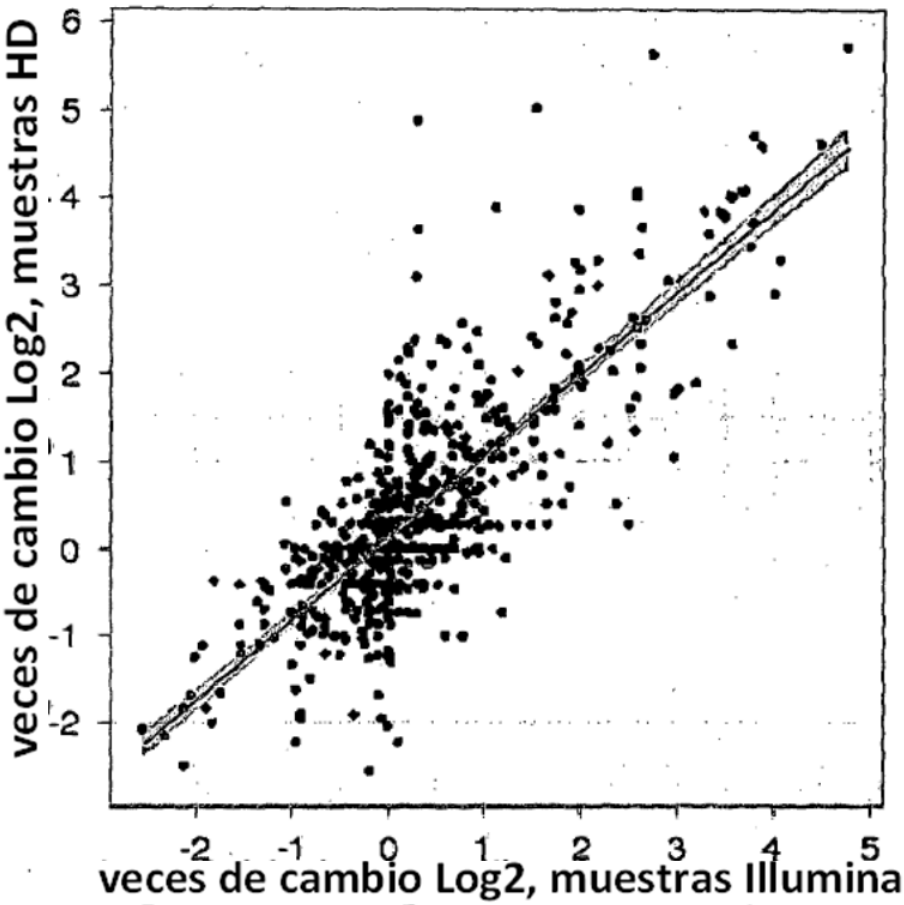

veces de cambio Log2, muestras HD
6
5
4
3
2
1
0
-1
-2
-2
-1
0
1
2
3
4
5
veces de cambio Log2, muestras Illumina

Figura 11a

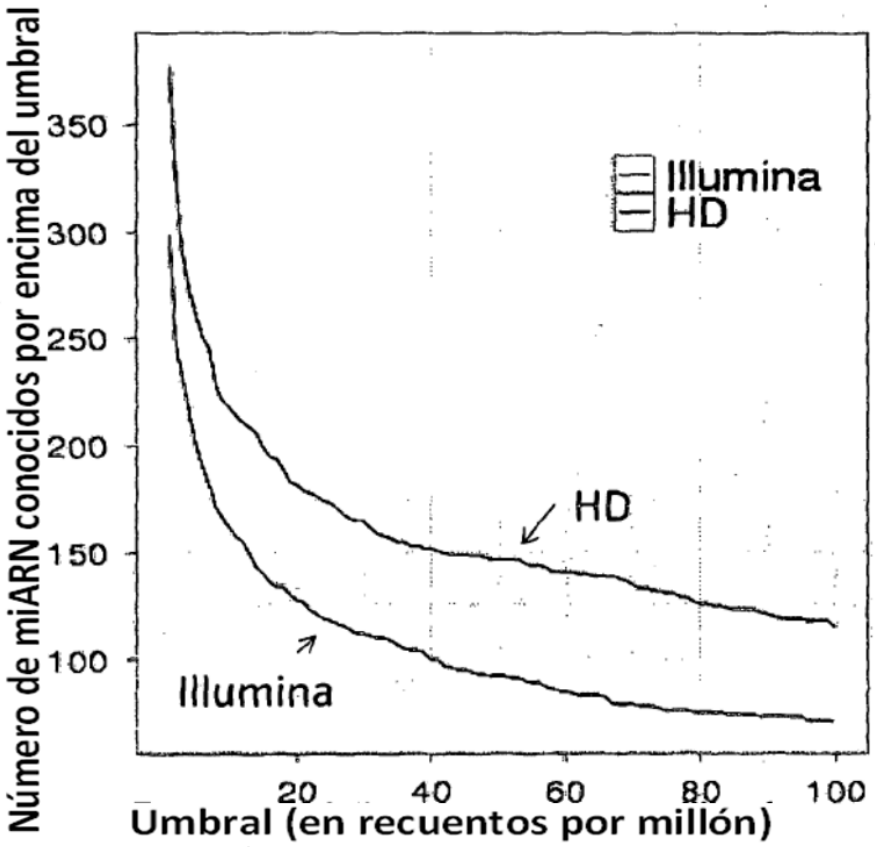

Número de miARN conocidos por encima del umbral
350
300
250
200
150
100
20
40
60
80
100
Umbral (en recuentos por millón)
Illumina
HD
HD
Illumina

Figura 11b

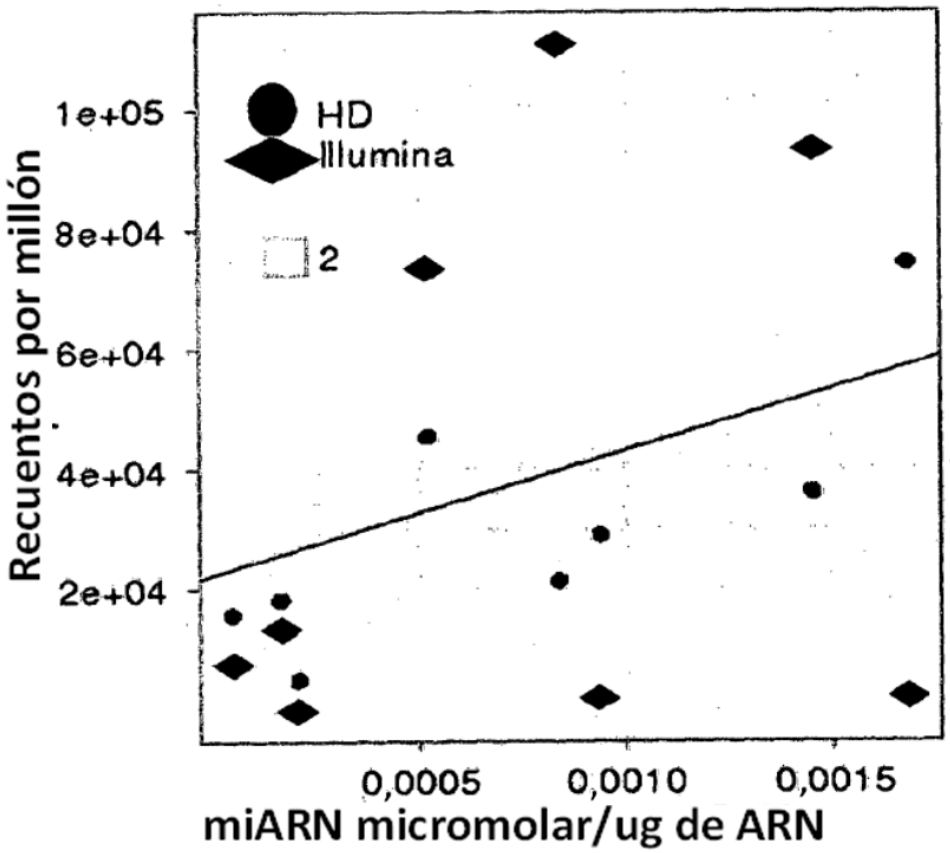

Recuentos por millón
1e+05
8e+04
6e+04
4e+04
2e+04
HD
Illumina
2
0,0005
0,0010
0,0015
miARN micromolar/ug de ARN

Figura 11c

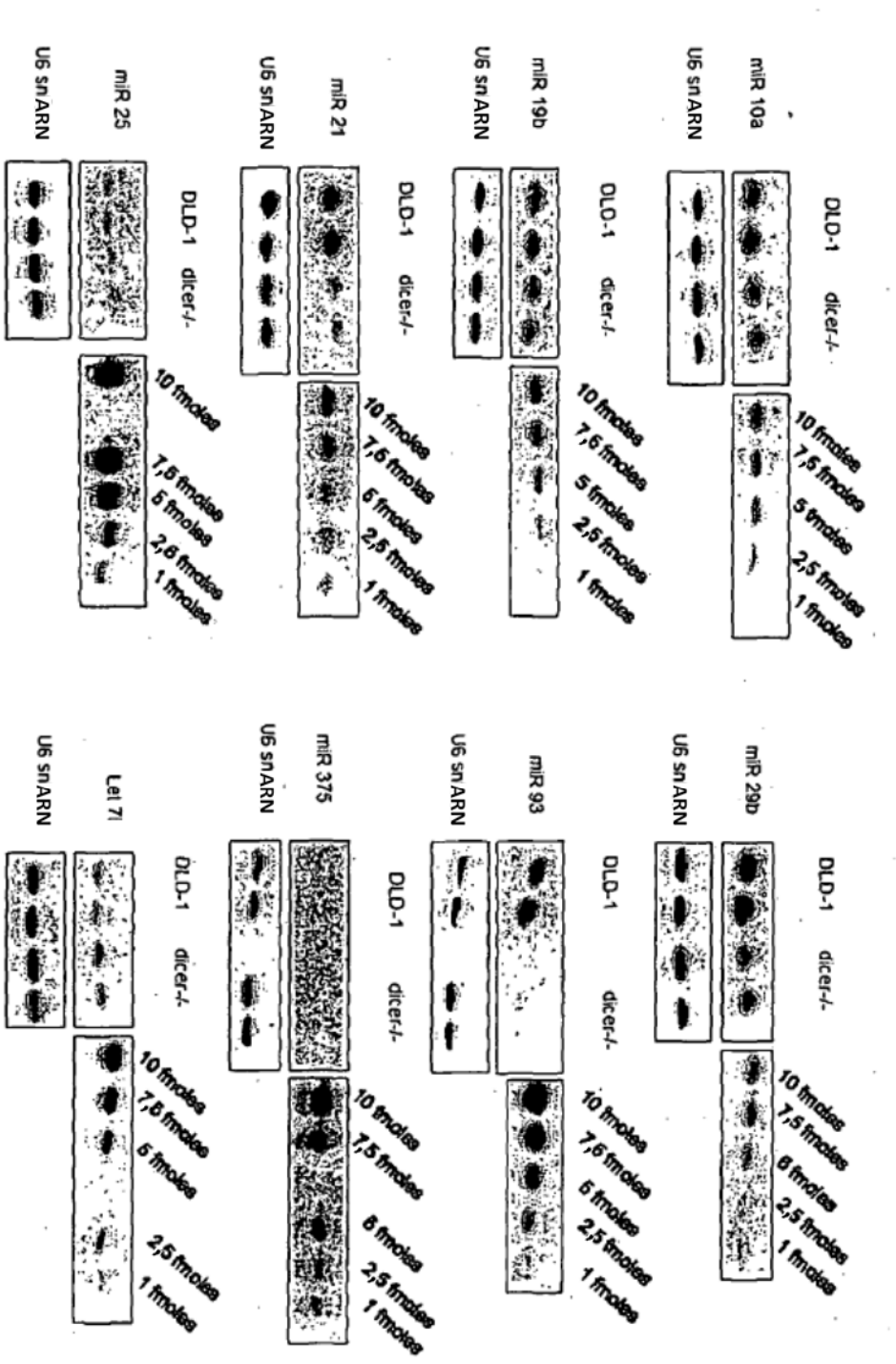

DLD-1 dicer-/-
miR 10a
U6 snARN
10 fmoles
7,5 fmoles
5 fmoles
2,5 fmoles
1 fmoles
DLD-1 dicer-/-
miR 19b
U6 snARN
10 fmoles
7,5 fmoles
5 fmoles
2,5 fmoles
1 fmoles
DLD-1 dicer-/-
miR 21
U6 snARN
10 fmoles
7,5 fmoles
5 fmoles
2,5 fmoles
1 fmoles
DLD-1 dicer-/-
miR 25
U6 snARN
10 fmoles
7,5 fmoles
5 fmoles
2,5 fmoles
1 fmoles
DLD-1 dicer-/-
miR 29b
U6 snARN
10 fmoles
7,5 fmoles
5 fmoles
2,5 fmoles
1 fmoles
DLD-1 dicer-/-
miR 93
U6 snARN
10 fmoles
7,5 fmoles
5 fmoles
2,5 fmoles
1 fmoles
DLD-1 dicer-/-
miR 375
U6 snARN
10 fmoles
7,5 fmoles
5 fmoles
2,5 fmoles
1 fmoles
DLD-1 dicer-/-
Let 7i
U6 snARN
10 fmoles
7,5 fmoles
5 fmoles
2,5 fmoles
1 fmoles

Figura 12