(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 776 673**

(51) Int. CI.:

C40B 50/10 (2006.01)
C12Q 1/68 (2008.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(86) Fecha de presentación y número de la solicitud internacional: **27.02.2013 PCT/US2013/027891**

(87) Fecha y número de publicación internacional: **06.09.2013 WO13130512**

(96) Fecha de presentación y número de la solicitud europea: **27.02.2013 E 13754428 (4)**

(97) Fecha y número de publicación de la concesión europea: **25.12.2019 EP 2820174**

(54) Título: **Métodos y usos para etiquetas moleculares**

(30) Prioridad:

27.02.2012 US 201261603909 P

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
31.07.2020

(73) Titular/es:

THE UNIVERSITY OF NORTH CAROLINA AT CHAPEL HILL (100.0%)
Office of Commercialization & Economic Development, 109 Church Street
Chapel Hill, NC 27516, US

(72) Inventor/es:

JABARA, CASSANDRA, B.;
ANDERSON, JEFFREY, A. y
SWANSTROM, RONALD, I.

(74) Agente/Representante:

ARIAS SANZ, Juan

DESCRIPCIÓN

Métodos y usos para etiquetas moleculares

Campo de la invención

La invención se refiere a un método para analizar una pluralidad de moléculas de ácido nucleico que comprende:

(a) unir una pluralidad de cebadores que comprenden un ID de cebador a una pluralidad de moléculas de ácido nucleico en una muestra para generar moldes de ácido nucleico etiquetados, en donde

(i) la pluralidad de moléculas de ácido nucleico comprende 10 o más moldes de ácido nucleico, y
(ii) cada molde de ácido nucleico etiquetado se une a un único ID de cebador;

(b) amplificar los moldes de ácido nucleico etiquetados para producir amplicones etiquetados;
(c) detectar los amplicones etiquetados, analizando de esta manera la pluralidad de moléculas de ácido nucleico; y
(d) determinar un sesgo de amplificación de la reacción de amplificación basado en la detección de las moléculas de ácido nucleico etiquetadas, en donde determinar el sesgo de amplificación se basa en la comparación de dos o más proporciones, en donde la comparación de dos o más proporciones comprende comparar una primera proporción de la cuantificación de diferentes ID de cebador asociados con dos o más tipos de moléculas de ácido nucleico a una segunda proporción de la cuantificación del número total de amplicones de dos o más tipos de moléculas de ácido nucleico,
en donde la primera proporción se basa en una cantidad de diferentes ID de cebador que se asocian con un primer tipo de molécula de ácido nucleico y una cantidad de diferentes ID de cebador asociados con un segundo tipo de molécula de ácido nucleico,
en donde la segunda proporción se basa en una cantidad de amplicones totales que se asocian con el primer tipo de molécula de ácido nucleico y una cantidad de amplicones que se asocian con el segundo tipo de molécula de ácido nucleico y
en donde el sesgo de amplificación se revela por la diferencia en la primera proporción y la segunda proporción.

La presente divulgación se refiere en general a etiquetas moleculares y más específicamente a composiciones que comprenden etiquetas moleculares y métodos para usar las etiquetas moleculares en análisis genético. Las etiquetas moleculares también se pueden usar en la identificación de variantes resistentes a fármacos. También se divulgan métodos para usar etiquetas moleculares para detectar y corregir errores de amplificación por PCR y errores de secuenciación.

Antecedentes de la invención

Las tecnologías de secuenciación profunda permiten el muestreo extenso de poblaciones genéticas. Las limitaciones de estas tecnologías, sin embargo, predisponen potencialmente este muestreo, en particular cuando una etapa de PCR precede al protocolo de secuenciación. Típicamente, se usa un número desconocido de moldes en iniciar la amplificación por PCR, y esto puede llevar a remuestreo de secuencia no reconocido creando homogeneidad aparente; además, la recombinación mediada por PCR puede perturbar el ligamiento, y la eficacia de amplificación diferencial o moldes que entran en diferentes ciclos de PCR pueden distorsionar la frecuencia de alelos. Por último, la mala incorporación de nucleótidos durante la PCR y los errores durante el protocolo de secuenciación pueden inflar la diversidad.

Jabara y col. divulga un método para revelar remuestreo de secuencia usando un único ID de cebador, que comprende etiquetar moléculas de ácido nucleico con un ID de cebador, amplificar, secuenciar, formar secuencias consenso y revelar el remuestro (Jabara et al, 2011, Proceedings of the National Academy of sciences, vol. 108, no. 50, páginas 20166-20171). Kivioja y col. divulga el uso de identificadores moleculares únicos (UMI) para recuento molecular absoluto (Kivioja et al, 2012, Nature Methods, vol. 9, no. 1,S páginas 72-74). Shiroguchi y col divulga el uso de códigos de barras para recuento de expresión donde una mezcla de adaptadores de códigos de barras se liga a una genoteca de ADNc ya formada (Shiroguchi et al, 2012, Proceedings of the National Academy of Sciences, vol 109, no. 4, páginas 1347-1352).

Compendio de la invención

Las limitaciones a las actuales técnicas se pueden superar al incluir una etiqueta de secuencia única en el cebador inicial de modo que cada molde recibe un ID de cebador único. Después de secuenciar, la identificación repetida de un ID de cebador revela remuestreo de secuencia. Estas secuencias remuestreadas se usan entonces para crear una secuencia consenso precisa para cada molde, corrigiendo para recombinación, distorsión alélica, errores de mala incorporación y errores de secuenciación. La población resultante de secuencias consenso representa directamente los moldes muestreados iniciales. El uso de estas etiquetas moleculares puede detectar y corregir error de PCR y/o error de secuenciación.

Este enfoque también se puede usar en análisis genético. El uso de etiquetas moleculares, tal como el ID de cebador, permite el análisis de la distribución de variación de secuencia de un gen en una población genética compleja. Con este enfoque, se han identificado polimorfismos principales y secundarios en posiciones codificantes y no codificantes. Además, se pueden observar cambios genéticos dinámicos en la población durante la exposición intermitente a fármaco, incluyendo la aparición de múltiples alelos resistentes. Los métodos divulgados en el presente documento proporcionan una visión sin precedentes de una población genética compleja en ausencia de remuestreo de PCR, sesgos de PCR y error de secuenciación.

Se divulgan métodos y usos para etiquetas moleculares. Cada copia de una molécula de ácido nucleico elige aleatoriamente de una reserva no agotable de diversos ID de cebador. La unión del ID de cebador a la molécula de ácido nucleico antes de la amplificación y secuenciación permite el recuento directo de moléculas de ácido nucleico y precisión aumentada en detectar variantes genéticas. Los ID de cebador también se pueden usar para la detección de variantes resistentes a fármacos. Por último, los ID de cebador también se pueden usar para reducir y/o corregir errores de PCR y/o errores de secuenciación.

En el presente documento se divulga un método para determinar diversidad genética de una muestra que comprende: (a) proporcionar una muestra que comprende una molécula molde de ácido nucleico; (b) unir un cebador que comprende un ID de cebador a cada molécula molde de ácido nucleico que se va a analizar para generar un molde de ácido nucleico etiquetado, en donde cada molde de ácido nucleico etiquetado está unido a un ID de cebador único; (c) amplificar el molde de ácido nucleico etiquetado para producir amplicones etiquetados; y (d) detectar los amplicones etiquetados, determinando mediante ello la diversidad genética de una muestra.

Se proporciona además un método para detectar variantes genéticas que comprende: (a) proporcionar una muestra que comprende una molécula molde de ácido nucleico; (b) unir un cebador que comprende un ID de cebador a cada molécula molde de ácido nucleico que se va a analizar para generar un molde de ácido nucleico etiquetado, en donde cada molde de ácido nucleico etiquetado está unido a un ID de cebador único; (c) amplificar el molde de ácido nucleico etiquetado para producir amplicones etiquetados; y (d) detectar los amplicones etiquetados, determinando mediante ello las variantes genéticas.

También se proporciona en el presente documento un método para determinar o cribar variantes resistentes a fármacos que comprende: (a) proporcionar una muestra que comprende una molécula molde de ácido nucleico; (b) unir un cebador que comprende un ID de cebador a cada molécula molde de ácido nucleico que se va a analizar para generar un molde de ácido nucleico etiquetado, en donde cada molde de ácido nucleico etiquetado está unido a un ID de cebador único; (c) amplificar el molde de ácido nucleico etiquetado para producir amplicones etiquetados; y (d) detectar los amplicones etiquetados, determinando o cribando mediante ello variantes resistentes a fármacos.

Se divulga además en el presente documento un método para determinar remuestreo de PCR en una reacción de amplificación que comprende: (a) proporcionar una muestra que comprende una molécula molde de ácido nucleico; (b) unir un cebador que comprende un ID de cebador a cada molécula molde de ácido nucleico que se va a analizar para generar un molde de ácido nucleico etiquetado, en donde cada molde de ácido nucleico etiquetado está unido a un ID de cebador único; (c) amplificar el molde de ácido nucleico etiquetado para producir amplicones etiquetados; y (d) detectar los amplicones etiquetados, determinando mediante ello el remuestreo de PCR en una reacción de amplificación.

Se divulga además en el presente documento un método para determinar error de PCR y/o error de secuenciación que comprende: (a) proporcionar una muestra que comprende una molécula molde de ácido nucleico; (b) unir un cebador que comprende un ID de cebador a cada molécula molde de ácido nucleico que se va a analizar para generar un molde de ácido nucleico etiquetado, en donde cada molde de ácido nucleico etiquetado está unido a un ID de cebador único; (c) amplificar el molde de ácido nucleico etiquetado para producir amplicones etiquetados; y (d) detectar los amplicones etiquetados, determinando mediante ello error de PCR y/o error de secuenciación. En algunas formas de realización, determinar el error de PCR y/o error de secuenciación comprende determinar la fidelidad de una polimerasa. En algunas formas de realización, determinar el error de PCR y/o error de secuenciación comprende determinar la precisión de los oligonucleótidos sintetizados in vitro.

Se divulga además en el presente documento un método para corregir error de PCR y/o error de secuenciación que comprende: (a) proporcionar una muestra que comprende una molécula molde de ácido nucleico; (b) unir un cebador que comprende un ID de cebador a cada molécula molde de ácido nucleico que se va a analizar para generar un molde de ácido nucleico etiquetado, en donde cada molde de ácido nucleico etiquetado está unido a un ID de cebador único; (c) amplificar el molde de ácido nucleico etiquetado para producir amplicones etiquetados; y (d) detectar los amplicones etiquetados, corrigiendo mediante ello error de PCR y/o error de secuenciación.

En algunas formas de realización, el ID de cebador comprende una secuencia degenerada. En algunas formas de realización, el ID de cebador comprende una secuencia semidegenerada. En algunas formas de realización, el ID de cebador comprende una secuencia mixta. En algunas formas de realización, el ID de cebador comprende una secuencia ambigua. En algunas formas de realización, el ID de cebador comprende una secuencia titubeante. En

algunas formas de realización, el ID de cebador comprende una secuencia aleatoria. En algunas formas de realización, el ID de cebador comprende una secuencia predeterminada.

En algunas formas de realización, el ID de cebador está unido al molde por ligación. En algunas formas de realización, el ID de cebador está unido al molde por hibridación. En algunas formas de realización, el ID de cebador está unido al molde a través de PCR. En algunas formas de realización, se analiza al menos una molécula molde. En algunas formas de realización, se analizan al menos dos moléculas molde diferentes. En algunas formas de realización, detectar los amplicones etiquetados comprende secuenciar los amplicones etiquetados. La secuenciación de los amplicones etiquetados se puede producir por una variedad de métodos, incluyendo, pero no limitados al método de secuenciación de Maxam-Gilbert, el método de secuenciación del dideoxi de Sanger, el método de secuenciación de terminador colorante, pirosecuenciación, secuenciación de ADN con cebador múltiple, secuenciación aleatoria, y desplazamiento sobre el cebador. En algunas formas de realización, la secuenciación comprende pirosecuenciación.

En algunas formas de realización, detectar los amplicones etiquetados comprende además contar un número de diferentes ID de cebador asociados con los amplicones etiquetados, en donde el número de diferentes ID de cebador asociados con los amplicones etiquetados refleja el número de moldes muestreados. En algunas formas de realización el método comprende además formar una secuencia consenso para amplicones etiquetados que comprenden el mismo ID de cebador.

En algunas formas de realización, el molde de ácido nucleico comprende un molde de ADN. En algunas formas de realización, el molde de ácido nucleico comprende un molde de ARN.

En algunas formas de realización, amplificar comprende un método basado en PCR. En algunas formas de realización, el método basado en PCR comprende PCR. En algunas formas de realización, el método basado en PCR comprende PCR cuantitativa. En algunas formas de realización, el método basado en PCR comprende PCR en emulsión. En algunas formas de realización, el método basado en PCR comprende PCR en gota. En algunas formas de realización, el método basado en PCR comprende PCR de inicio en caliente. En algunas formas de realización, el método basado en PCR comprende PCR in situ. En algunas formas de realización, el método basado en PCR comprende PCR inversa. En algunas formas de realización, el método basado en PCR comprende PCR multiplex. En algunas formas de realización, el método basado en PCR comprende PCR de número variable de repeticiones en tándem (VNTR). En algunas formas de realización, el método basado en PCR comprende PCR asimétrica. En algunas formas de realización, el método basado en PCR comprende PCR larga. En algunas formas de realización, el método basado en PCR comprende PCR anidada. En algunas formas de realización, el método basado en PCR comprende PCR hemianidada. En algunas formas de realización, el método basado en PCR comprende PCR touchdown. En algunas formas de realización, el método basado en PCR comprende PCR de ensamblaje. En algunas formas de realización, el método basado en PCR comprende PCR de colonia.

En algunas formas de realización, amplificar comprende un método no basado en PCR. En algunas formas de realización, el método no basado en PCR comprende amplificación por desplazamiento múltiple (MDA). En algunas formas de realización, el método no basado en PCR comprende amplificación mediada por transcripción (TMA). En algunas formas de realización, el método no basado en PCR comprende amplificación basada en secuencia de ácido nucleico (NASBA). En algunas formas de realización, el método no basado en PCR comprende amplificación por desplazamiento de la hebra (SDA). En algunas formas de realización, el método no basado en PCR comprende SDA en tiempo real. En algunas formas de realización, el método no basado en PCR comprende amplificación por círculo rodante. En algunas formas de realización, el método no basado en PCR comprende amplificación de círculo a círculo.

Breve descripción de los dibujos

El experto en la materia entenderá que los dibujos descritos a continuación son para fines de ilustración solo. No se pretende que los dibujos limiten el ámbito de las presentes enseñanzas en modo alguno.

La **figura 1A** muestra un ejemplo de cebador que comprende un ID de cebador y un código de barras.

La **figura 1B** muestra el uso de un cebador que comprende un ID de cebador y un código de barras para detectar y corregir sesgos de PCR y error de secuenciación (SEQ ID No. 1).

La **figura 1C** muestra la creación de una secuencia consenso. En particular, la figura 1A-1C muestra que etiquetar moldes de ARN vírico con un ID de cebador antes de la amplificación por PCR y secuenciación permite la eliminación directa de errores artefactuales e identifica remuestreo. La figura 1A muestra un cebador que se diseñó para unirse después del dominio codificante de proteasa. En la cola 5' del cebador, una cadena degenerada de ocho nucleótidos creó un ID de cebador, que permite 65.536 combinaciones únicas. Se diseñó un código de barras de tres nucleótidos seleccionado *a priori* para el ID de la muestra. Por último, una cadena heteróloga de nucleótidos con baja afinidad al genoma de VIH-1 se incluyó en el extremo 5' lejano para uso como el sitio cebador en la amplificación por PCR. (Figura 1B) Se introducen sesgos de PRC y error de secuenciación durante la amplificación y secuenciación de moldes víricos. La identificación repetitiva del código de barras y el ID de cebador permite el seguimiento de cada suceso de molde a partir de un ADNc etiquetado individual. Como los errores son componentes minoritarios en la población de ID de

cebador, formar una secuencia consenso directamente los elimina, y corrige el remuestreo de PCR. (Figura 1C). Moldes de ARN de VIH-1 aislados de muestras de plasma de dos pre terapia de fármaco ritonavir y una posterior intermitente se etiquetaron, amplificaron y se sometieron a secuenciación profunda. Se usaron secuencias etiquetadas que contenían la proteasa de longitud completa para crear una población de secuencias consenso cuando al menos tres secuencias contenían un código de barras y un ID de cebador idénticos.

La **figura 2A-2B** muestra la frecuencia de variación de codón a través de las 99 posiciones en la proteasa a lo largo de tres puntos de tiempo. En una posición de codón, las primeras dos barras representan puntos de tiempo sin tratar 1 y 2, respectivamente. Las barras 3 y 4 son el tercer punto de tiempo separado basado en la presencia o ausencia de las mutaciones de resistencia a ritonavir. La barra 3 es la población de genotipos susceptibles (definidos como no V82A, I84V o L90M), y la barra 4 es la población variante resistente principal, V82A. Las barras hacia arriba son cambios no sinónimos (escala en letra regular), y las barras hacia abajo son cambios sinónimos (escala en letra negrita). En una posición de codón, el sombreado diferente representa diferentes SNP.

La **figura 3** muestra la representación filogenética de población de proteasa derivada de secuenciación profunda con un ID de cebador. Se construyó un árbol de unión de vecinos a partir de secuencias derivadas de los tres puntos de tiempo y se coloreó basado en la susceptibilidad a ritonavir. Los taxones V82 representan variantes susceptibles (definidos como no V82A/I/L/F, I84V o L90M). Los taxones V82A representan variantes que contienen la variante resistente a ritonavir principal, V82A. Otros taxones representan las variantes resistentes minoritarias V82I/L/F y los alelos resistentes minoritarios L90M e I84V, respectivamente. En un grupo el brillo se correlaciona con el tiempo de muestra. Las flechas oscuras señalan a secuencias pre-RTV de baja abundancia que se amplificaron clonalmente a sus respectivos clados.

La **figura 4** muestra un muestreo longitudinal de plasma sanguíneo de un único individuo infectado con VIH-1 de subtipo B antes y después de una pauta de monoterapia con ritonavir fallida. Se muestrearon dos puntos de tiempo separados ~6 meses antes de la terapia de ritonavir (T1 y T2). Se muestreó un punto de tiempo después de la monoterapia de ritonavir intermitente fallida (T3). Las áreas sombreadas representan tiempos de cumplimiento de terapia basado en un autoinforme.

La **figura 5** muestra el flujo lógico de la tubería bioinformática que procesó lecturas de secuencia en bruto a secuencias consenso. Primero, cuando es aplicable, las lecturas se convirtieron a orientación directa. A continuación, las lecturas se evaluaron para cebador de etiquetado de síntesis de ADNc que contiene información que identifica la muestra y el cebador correctamente (código de barras e ID de cebador, respectivamente). Las secuencias se agrupan después basado en un código de barras, y dentro de cada código de barras, se agrupan por ID de cebador, después se recortan a solo el dominio codificante de proteasa. Para las secuencias de proteasa de longitud completa, cuando al menos 3 secuencias dentro de un archivo de código de barras contenían un ID de cebador idéntico, se hizo una secuencia consenso basada en la regla de la mayoría y el uso de designaciones de nucleótidos ambiguas para vínculos. Las secuencias se filtraron después además basado en estimaciones de antecedentes de error para la síntesis de ADNc por RT in vitro y la primera ronda de síntesis de ADN polimerasa Taq.

La **figura 6A** muestra la distribución del número de lecturas por ID de cebador o secuencia consenso. Las barras grises izquierdas representan la distribución del remuestreo de la población de secuencias filtradas inmediatamente antes de la generación de la secuencia consenso. Dentro de un único ID de cebador, cuando estaban presentes tres o más secuencias, se formó una secuencia consenso. Las barras grises derechas representan la distribución del número de lecturas que fueron a cada secuencia consenso. Los valores mostrados representan la media para los datos de los tres puntos de tiempo con las barras de error que representan la DE entre las tres muestras. Las barras con estrellas se incluyen para marcar posiciones donde una única secuencia tenía alta aparición de remuestreo.

La **figura 6B** muestra el número de secuencias consenso que contienen una ambigüedad como función de la extensión de remuestreo. Se combinaron los tres puntos de tiempo. Las barras grises oscuras representan las secuencias consenso sin ambigüedad, y las barras grises claras representan las secuencias consenso con ambigüedad. Hay un patrón discernible de un número aumentado de ambigüedades que va hasta 22 lecturas/secuencia consenso para esas secuencias consenso creadas de un número par de lecturas, el resultado de tener un vínculo entre dos secuencias diferentes en una posición. Sin embargo, esto representa solo una pequeña fracción de las lecturas totales (5,4%). La posición de aminoácidos en la mayor ambigüedad total se usó por población de ID de cebador.

La **figura 7** muestra un análisis de variantes de baja abundancia para la distribución de distorsión alélica. Usamos secuencias descartadas (es decir, secuencias únicas representadas por un único ID de cebador) y genomas transitorios definidos como que tienen un SNP de baja abundancia en la población preconsenso por punto de tiempo sin tratar. Se definieron secuencias transitorias como que tienen al menos dos secuencias en solo uno de los puntos de tiempo sin tratar, o una copia en uno de los puntos de tiempo sin tratar y después otra vez en el tercer punto de tiempo. Estas secuencias se usaron para definir un conjunto de secuencias que se podría comparar para abundancia de baja frecuencia en el conjunto de datos total frente a las secuencias consenso. Las barras horizontales representan la frecuencia medida de secuencias de una copia única en la población consenso en T1 y T2. Los puntos oscuros representan genomas descartados, y los puntos claros representan genomas transitorios su posición indica su abundancia en la población de secuencias total antes de la construcción de las secuencias consenso. Los puntos

grises claros representan secuencias presentes en T1, los puntos grises más oscuros representan secuencias en T2. Estos datos muestran que la distorsión alélica de 2 veces hacia arriba y de 10 hacia abajo es común antes de la formación de la secuencia consenso.

La **figura 8A-8C** muestra las variantes alélicas principales y secundarias en las poblaciones sin tratar. La figura 8A muestra la frecuencia de las secuencias pro gen principales y secundarias únicas. Los colores grises representan secuencias pro gen presentes entre el 2,5 y el 0,5% en frecuencia. El negro representa la suma de todas las secuencias pro gen individualmente presentes a <0,5%. La figura 8B muestra la distribución de SNP de las secuencias pro gen más abundantes (>2,5%), los puntos sombreados a la derecha indican las secuencias correspondientes identificadas en el gráfico de sectores (Fig. 8A). La figura 8C muestra la distribución de SNP de variantes presentes entre el 2,5 y el 0,5%, las mismas secuencias indicadas en el panel figura 8A con la barra gris. En la línea en la parte inferior indicada por el círculo negro representa la suma de todas las variantes <0,5% en frecuencia para las secuencias mostradas en negro en el gráfico de sectores (Fig. 8A).

La **figura 9A-9F** muestra las secuencias pro gen principales y secundarias únicas en las poblaciones resistentes principales V82A, L90M, e I84V. (Fig. 9A) Frecuencia de diferentes secuencias pro gen únicas que portan la mutación V82A a alta frecuencia (coloreado >2,5%) y baja frecuencia (<2,5%, negro con la abundancia reunida). (Fig. 9B) Gráfico Highlither que muestra los cambios de secuencia de la secuencia consenso para las variantes pro gen principales (>2,5%) que portan la mutación V82A. La sustitución V82A está indicada por el cambio de nucleótido en la posición 245 mostrada en gris claro. (Fig. 9C) Frecuencia de diferentes secuencias pro gen únicas que portan la mutación L90M a alta frecuencia (coloreado >2,5%) y baja frecuencia (<2,5%, negro con la abundancia reunida). (Fig. 9D) Gráfico Highlighter que muestra los cambios de secuencia de la secuencia consenso para las variantes pro gen principales (>2,5%) que portan la mutación L90M. La sustitución L90M está indicada por el cambio de nucleótido en la posición 268 mostrada en gris. (Fig. 9E) Frecuencia de diferentes secuencias pro gen únicas que portan la mutación I84V a alta frecuencia (coloreado >2,5%) y baja frecuencia (<2,5%, negro con la abundancia reunida). (Fig. 9F) Gráfico Highlither que muestra los cambios de secuencia de la secuencia consenso para las variantes pro gen principales (>2,5%) que portan la mutación I84V. La sustitución I84V está indicada por el cambio de nucleótido en la posición 250 mostrada en gris.

La **figura 10** muestra las frecuencias de aparición de nucleótidos individuales en cada posición del ID de cebador (marcado 1-8) en secuencias consenso resueltas. El sombreado representa dA, dT, dC y dG, respectivamente. En el eje horizontal, cada posición del ID de cebador está subdividida por punto de tiempo (T1, T2 y T3).

La **figura 11** muestra la frecuencia de deleciones en secuencias totales frente a consenso. El porcentaje y posición de nucleótido de deleciones de nucleótidos únicos se representan en secuencias totales (barras hacia arriba) y consenso (barras hacia abajo). La sombra corresponde al punto de tiempo para T1, T2 y T3.

La **figura 12A** muestra un esquema de un cebador de etiqueta. La secuencia degenerada etiqueta moldes de ARNv individuales con un ID único. Un código de barras seleccionado *a priori* sirve como ID de muestra. Juntos, las muestras y moldes individuales se pueden seguir tras la amplificación por PCR y secuenciación. La **figura 12B** muestra el muestreo de pacientes y antecedentes clínicos. Se extrajo el ARN de VIH-1 de plasma sanguíneo de un único individuo infectado. La proteasa de dos pre-terapias de ritonavir y una post-terapia de ritonavir intermitente se etiquetó y secuenció.

La **figura 13A** muestra las frecuencias alélicas por posición de aminoácido. Las barras hacia arriba son cambios codificantes, las barras hacia abajo son cambios silenciosos. El cambio en color en una posición de aminoácido se correlaciona con cambio en codón. 1 = T1, 2 = T2, 3S = T3 V82, T3R = T3 V92A. La **figura 13B** muestra el resumen de las secuencias resueltas. Las secuencias totales son el número de secuencias que contienen proteasa de longitud completa con el cebador de etiqueta. En una muestra, cuando tres o más secuencias contenían un ID de cebador idéntico, se generó una secuencia consenso.

La **figura 14** muestra una tubería bioinformática. Las lecturas de secuencia en crudo se cribaron para cebadores de etiqueta incorruptos y proteasa de longitud completa. Se usaron un mínimo de tres ID de muestras y cebadores idénticos para crear secuencias consenso individuales.

La **figura 15** muestra la generación de una secuencia consenso. Malas incorporaciones de la polimerasa, recombinación artificial, amplificación diferencial y errores de secuenciación introducen diversidad de secuencia y distorsionan las frecuencias alélicas. Crear una secuencia consenso corrige directamente esto (SEQ ID No. 1).

La **figura 16** muestra un árbol filogenético y un gráfico Highlighter demuestra la aparición de la cepa V82A. Se construyó un árbol de unión de vecinos a partir de las secuencias consenso para los tres puntos de tiempo. Los clados se colorearon basados en el aminoácido V82 (gris claro) o V82A (gris más oscuro). El gráfico Highlighter representa los SNP (definidos a partir de la secuencia consenso de T1 y T2) en la población V82A. Indicada al lado de cada secuencia está la frecuencia en la población y el número de secuencias en la construcción. A; gris, T; negro, G; gris más oscuro, G; gris más claro.

Tabla 1. Frecuencia de codones no consenso por posición.
Tabla 2. Resumen de variación de nucleótidos en puntos de tiempo muestreados.

Descripción detallada de la invención

En el presente documento se divulgan métodos, kits, y sistemas para analizar una o más moléculas de ácido nucleico en una muestra. En general, el método comprende (a) unir un ID de cebador a una molécula de ácido nucleico o fragmento de la misma para producir un ácido nucleico etiquetado; y (b) detectar la molécula de ácido nucleico etiquetada o un derivado o un producto de la misma.

La unión del ID de cebador a la molécula de ácido nucleico se puede producir por cualquier método conocido en la técnica. Por ejemplo, la unión del ID de cebador puede comprender ligación. La ligación puede comprender ligación de extremos romos. Alternativamente, la ligación comprende ligación de extremos cohesivos. Alternativamente, o además, la unión del ID de cebador pueden comprender extensión de cebador. La unión del ID de cebador a la molécula de ácido nucleico puede comprender transcripción o transcripción inversa. La unión del ID de cebador a la molécula de ácido nucleico puede comprender una o más técnicas de reparación de extremos de los extremos de la molécula de ácido nucleico.

El método puede además comprender, antes de la etapa de detección, amplificar la molécula de ácido nucleico etiquetada para producir uno o más amplicones etiquetados, en donde detectar comprende detectar los amplicones etiquetados. La amplificación de los amplicones etiquetados puede comprender cualquier método conocido en la técnica. Por ejemplo, la amplificación puede comprender un método de amplificación basado en PCR. Alternativamente, o además, la amplificación puede comprender un método de amplificación no basado en PCR.

El método puede además comprender, antes de la etapa de unión, fragmentar una molécula de ácido nucleico para producir fragmentos de ácido nucleico, en donde los ID de cebador se unen a los fragmentos de ácido nucleico. Fragmentar la molécula de ácido nucleico se puede producir por cualquier método conocido en la técnica. Por ejemplo, fragmentar la molécula de ácido nucleico puede comprender corte. Cortar puede comprender corte mecánico. Fragmentar puede comprender sonicar la muestra. Alternativamente, fragmentar puede comprender una o más enzimas de restricción. La una o más o enzimas de restricción puede ser una endonucleasa de restricción.

La detección de la molécula de ácido nucleico etiquetada puede comprender cualquier método conocido en la técnica. La detección de la molécula de ácido nucleico etiquetada puede comprender hibridación, secuenciación, captura de la molécula de ácido nucleico etiquetada, electroforesis, luminiscencia, quimioluminiscencia, o cualquier combinación de las mismas. La detección puede comprender detección de la parte ID de cebador de la molécula de ácido nucleico etiquetada. El ID de cebador puede comprender un marcador detectable (por ejemplo, fluoróforo, colorante, bola, antígeno, anticuerpo, péptido, etc.). La detección de la molécula de ácido nucleico etiquetado puede comprender hibridación de la molécula de ácido nucleico etiquetado a un soporte sólido (por ejemplo, matriz, bola, placa).

El método puede comprender además detectar una o más variantes genéticas basado en la detección de la molécula de ácido nucleico etiquetado. Por ejemplo, las variantes genéticas se pueden detectar secuenciando la molécula de ácido nucleico etiquetado. Las secuencias con el mismo ID de cebador se pueden agrupar para formar una familia de ID de cebador. Se puede detectar una variante genética cuando al menos el 50% de las moléculas de ácido nucleico en la familia del ID de cebador contienen la misma variación de secuencia de nucleótidos. Cuando menos de 105 de las moléculas de ácido nucleico en la familia de ID de cebador contiene la misma variación de secuencia de nucleótidos, entonces la variación de la secuencia de nucleótidos puede ser debida a error de secuenciación y/o amplificación.

El método puede comprender además determinar el sesgo de amplificación de una reacción de amplificación basado en la detección de las moléculas de ácido nucleico etiquetado. Sesgo de amplificación o remuestreo de PCR se pueden usar de forma intercambiable y se puede referir a la amplificación desigual de moldes de ácido nucleico. El sesgo de amplificación puede producir una distorsión de la distribución de productos de PCR (por ejemplo, amplicones). El sesgo de amplificación puede ser debido a diferencias en la eficacia de amplificación de dos o más moldes de ácido nucleico. Alternativamente, o además, el sesgo de amplificación puede ser debido a la inhibición de la amplificación de un molde de ácido nucleico. Determinar el sesgo de amplificación se puede basar en la comparación de dos o más proporciones, en donde la comparación de las dos o más proporciones comprende comparar una primera proporción de la cuantificación de los diferentes ID de cebador asociados con dos o más tipos de moléculas de ácido nucleico a una segunda proporción de la cuantificación del número total de amplicones de dos o más tipos de moléculas de ácido nucleico. La primera proporción se puede basar en la cantidad de diferentes ID de cebador asociados con un primer tipo de molécula de ácido nucleico y la cantidad de diferentes ID de cebador asociados con un segundo tipo de molécula de ácido nucleico. La segunda proporción se puede basar en el número de amplicones totales asociados con el primer tipo de molécula de ácido nucleico y el número de amplicones totales asociados con el segundo tipo de molécula de ácido nucleico. En algunos casos, la diferencia en la primera proporción y la segunda proporción puede revelar sesgo de amplificación.

El método puede comprender además determinar la eficacia de amplificación de una molécula de ácido nucleico basado en la detección de la molécula de ácido nucleico etiquetado. Determinar la eficacia de amplificación puede comprender cuantificar el número de diferentes ID de cebador asociados con la molécula de ácido nucleico. El método puede comprender además comparar el número de diferentes ID de cebador asociados con la molécula de ácido nucleico con el número de diferentes ID de cebador asociados con un control de ácido nucleico.

Se hará referencia ahora en detalle a formas de realización ejemplares de la divulgación. Mientras que la invención se describirá junto con las formas de realización ejemplares, se entenderá que no se pretende que limiten la divulgación a estas formas de realización. Por el contrario, se pretende que la divulgación cubra alternativas, modificaciones y equivalentes.

La divulgación tiene muchas formas de realización preferidas y se basa en muchas patentes, solicitudes y otras referencias para detalles conocidos por los expertos en la materia.

Como se usa en esta solicitud, la forma singular “un”, “una”, “el” y “la incluye referencias plurales a menos que el contexto claramente indique otra cosa. Por ejemplo, el término “un agente” incluye una pluralidad de agentes, incluyendo mezclas de los mismos.

Un individuo no está limitado a un ser humano, sino que también puede ser otros organismos incluyendo, pero no limitado a, mamíferos, plantas, bacterias, células derivadas de cualquiera de los anteriores, virus o células infectadas con virus.

A lo largo de esta divulgación, varios aspectos de esta divulgación se pueden presentar en un formato de intervalo. Se debe entender que la descripción en formato de intervalo es solamente por conveniencia y brevedad y no se debe interpretar como una limitación inflexible en el ámbito de la divulgación. Según esto, la descripción de un intervalo se debe considerar que tiene específicamente divulgados todos los posibles subintervalos, así como los valores numéricos individuales en ese intervalo. Por ejemplo, la descripción de un intervalo tal como de 1 a 6 se debe considerar que divulga específicamente subintervalos tal como de 1 a 3, de 1 a 4, de 1 a 5, de 2 a 4, de 2 a 6, de 3 a 6, etc., así como los números individuales en ese intervalo, por ejemplo, 1, 2, 3, 4, 5, y 6. Esto se aplica independientemente de la amplitud del intervalo.

Las muestras adecuadas para análisis pueden derivar de una variedad de fuentes. Las muestras biológicas pueden ser de cualquier tejido o fluido biológico o células de cualquier organismo. Con frecuencia la muestra será una “muestra clínica” que es una muestra derivada de un paciente. Las muestras clínicas proporcionan una fuente rica de información respecto a los varios estados de expresión génica y número de copia. Las muestras clínicas típicas incluyen, pero no están limitadas a, esputo, sangre, muestras de tejido o biopsia por aguja fina, orina, líquido peritoneal, y derrame pleural, o células de las mismas. Las muestras biológicas también pueden incluir secciones de tejidos, tal como secciones congeladas o secciones fijadas en formalina tomadas para fines histológicos, que pueden incluir muestras fijadas en formalina, embebidas en parafina (FFPE) y muestras derivadas de las mismas. Las muestras FFPE son una fuente particularmente importante para estudio de tejido archivado ya que los ácidos nucleicos se pueden recuperar de estas muestras incluso después de almacenamiento a largo plazo de las muestras a temperatura ambiente. Véase, por ejemplo, Specht *et al. Am J Path.* (2001), 158(2):419-429. Los ácidos nucleicos aislados de muestras frescas congeladas también se pueden analizar usando los métodos divulgados.

La práctica de la presente divulgación puede emplear, a menos que se indique otra cosa, técnicas y descripciones convencionales de química orgánica, tecnología de polímeros, biología molecular (incluyendo técnicas recombinantes), biología celular, bioquímica, e inmunología, que están dentro de las capacidades de la técnica. Tales técnicas convencionales incluyen síntesis de matrices de polímeros, hibridación, ligación, y detección de hibridación usando un marcador. Las ilustraciones específicas de las técnicas adecuadas se pueden tener mediante referencia al ejemplo en el presente documento posteriormente. Sin embargo, por supuesto, también se pueden usar otros procedimientos convencionales equivalentes. Tales técnicas y descripciones convencionales se pueden encontrar en manuales de laboratorio estándar tal como *Genome Analysis: A Laboratory Manual Series (Vols. I-IV), Using Antibodies: A Laboratory Manual, Cells: A Laboratory Manual, PCR Primer: A Laboratory Manual,* y Molecular *Cloning: A Laboratory Manual* (todos de Cold Spring Harbor Laboratory Press), *Gait, "Oligonucleotide Synthesis: A Practical Approach" 1984, IRL Press, Londres,* Nelson y Cox (2000), Lehninger *et* al.,(2008) *Principles of Biochemistry* 5ª Ed., W.H. Freeman Pub., Nueva York, NY y Berg *et al.* (2006) *Biochemistry,* 6ª Ed., W.H. Freeman Pub., Nueva York, NY.

La presente divulgación puede emplear sustratos sólidos, incluyendo matrices en algunas formas de realización preferidas. Se han descrito métodos y técnicas aplicables a la síntesis de matrices de polímeros (incluyendo proteínas) en la publicación de patente en EE UU No. 20050074787, documento WO 00/58516, patentes en EE UU Nos. 5.143.854, 5.242.974, 5.252.743, 5.324.633, 5.384.261, 5.405.783, 5.424.186, 5.451.683, 5.482.867, 5.491.074, 5.527.681, 5.550.215, 5.571.639, 5.578.832, 5.593.839, 5.599.695, 5.624.711, 5.631.734, 5.795.716, 5.831.070, 5.837.832, 5.856.101, 5.858.659, 5.936.324, 5.968.740, 5.974.164, 5.981.85, 5.981.956, 6.025.601, 6.033.860, 6.040.193, 6.090.555, 6.136.269, 6.269.846 y 6.428.752, en las publicaciones PCT No. WO 99/36760 y WO 01/58593. Las patentes que describen técnicas de síntesis en formas de realización específicas incluyen las patentes en EE UU No. 5.412.087, 6.147.205, 6.262.216, 6.310.189, 5.889.165 y 5.959.098. Las matrices de ácidos nucleicos se

describen en muchas patentes anteriores, pero muchas de las mismas técnicas se pueden aplicar a matrices de polipéptidos.

La presente divulgación también contempla muchos usos para polímeros unidos a sustratos sólidos. Estos usos incluyen seguimiento de la expresión génica, perfil de transcripción, cribado de genotecas, genotipado, análisis epigenético, análisis del patrón de metilación, tipado de tumores, farmacogenómica, agrogenética, perfil de patógeno y detección y diagnóstico. El seguimiento de la expresión génica y métodos de perfiles se han mostrado en las patentes en EE UU No. 5.800.992, 6.013.449, 6.020.135, 6.033.860, 6.040.138, 6.177.248 y 6.309.822. El genotipado y usos para el mismo se muestran en las publicaciones de patente en EE UU No. 20030036069 y 20070065816 y las patentes en EE UU No. 5.856.092, 6.300.063, 5.858.659, 6.284.460, 6.361.947, 6.368.799 y 6.333.179. Otros usos están representados en las patentes en EE UU No. 5.871.928, 5.902.723, 6.045.996, 5.541.061 y 6.197.506.

La presente divulgación también contempla métodos de preparación de muestras en ciertas formas de realización. Antes de o al mismo tiempo que el análisis, la muestra se puede amplificar por una variedad de mecanismo. En algunos aspectos los métodos de amplificación de ácido nucleico tal como PCR se pueden combinar con los métodos y sistemas divulgados. Véase, por ejemplo, *PCR Technology: Principles and Applications for DNA Amplification* (Ed. H.A. Erlich, Freeman Press, NY, NY, 1992); *PCR Protocols: A Guide to Methods and Applications* (Eds. Innis, et al., Academic Press, San Diego, CA, 1990); Mattila et al., *Nucleic Acids Res.* 19, 4967 (1991); Eckert et al., *PCR Methods and Applications 1,* 17 (1991); *PCR* (Eds. McPherson et al., IRL Press, Oxford); y las patentes en EE UU Nos. 4.683.202, 4.683.195, 4.800.159, 4.965.188, y 5.333.675. Se describen métodos adicionales de preparación de muestras y técnicas para reducir la complejidad de una muestra de ácido nucleico en Dong et al., *Genome Research 11,* 1418 (2001), en las patentes en EE UU Nos. 6.300.070 (amplificación en una matriz), 6.361.947, 6.391.592, 6.872.529 y 6.458.530 y publicaciones de patente en EE UU. Nos. 20030096235, 20030082543, 20030039069, 20050079536, 20040072217, 20050142577, 20050233354, 20050227244, 20050208555, 20050074799, 20050042654, y 20040067493.

Muchos de los métodos y sistemas divulgados en el presente documento utilizan actividades enzimáticas. Se revisan enzimas y métodos relacionados para uso en biología molecular que se pueden usar en combinación con los métodos divulgados, por ejemplo, en Rittie and Perbal, *J. Cell Commun. Signal.* (2008) 2:25-45. Una variedad de enzimas se conoce bien, se han caracterizado y muchas están comercialmente disponibles de uno o más suministradores. Las enzimas ejemplares incluyen ADN polimerasas dependientes de ADN (tal como las mostradas en la tabla 1 de Rittie y Perbal), ADN polimerasa dependiente de ARN (véase la tabla 2 de Rittie y Perbal), ARN polimerasas (tal como T7 y SP6), ligasas (véase la tabla 3 de Rittie y Perbal), enzimas para transferencia y eliminación de fosfato (véase la tabla 4 de Rittie y Perbal), nucleasas (véase la tabla 5 de Rittie y Perbal), y metilasas.

Otros métodos de análisis y reducción de complejidad del genoma incluyen, por ejemplo, AFLP, véase la patente en EE UU 6.045.994, y PCR arbitrariamente cebada (AP-PCR) véase, McClelland y Welsh, in *PCR Primer: A laboratory Manual,* (1995) eds. C. Dieffenbach and G. Dveksler, Cold Spring Harbor Lab Press, por ejemplo, en la p 203.

Otros métodos de amplificación adecuados incluyen la reacción en cadena de la ligasa (LCR) *(por ejemplo,* Wu y Wallace, *Genomics* 4, 560 (1989), Landegren et al., *Science* 241, 1077 (1988) y Barringer et al. *Gene* 89:117 (1990)), amplificación por transcripción (Kwoh et al., *Proc. Natl. Acad. Sci. USA* 86, 1173 (1989) y documento WO88/10315), replicación de secuencia autosostenida (Guatelli et al., *Proc. Nat. Acad. Sci. USA,* 87, 1874 (1990) y documento WO90/06995), amplificación selectiva de moléculas de polinucleótido diana (patente en EE UU No. 6.410.276), reacción en cadena de la polimerasa cebada con secuencia consenso (CP-PCR) (patente en EE UU No. 4.437.975), reacción en cadena de la polimerasa arbitrariamente cebada (AP-PCR) (patentes en EE UU Nos. 5.413.909, 5.861.245), amplificación por círculo rodante (RCA) por ejemplo, Fire y Xu, PNAS 92:4641 (1995) y Liu et al., J. Am. Chem. Soc. 118:1587 (1996)) y patente en EE UU No. 5.648.245, amplificación por desplazamiento de hebra (véase Lasken y Egholm, *Trends Biotechnol.* 2003 21(12):531-5; Barker et al. *Genome Res.* Mayo 2004;14(5):901-7; Dean *et al. Proc Natl Acad Sci* U S A. 2002; 99(8):5261-6; Walker et al. 1992, Nucleic Acids Res. 20(7):1691-6, 1992 y Paez, *et al. Nucleic Acids Res.* 2004; 32(9):e71), replicasa Qbeta, descrita en la solicitud de patente PCT No. PCT/US87/00880 y amplificación de secuencia basada en ácido nucleico (NABSA). (Véase, patentes en EE UU Nos. 5.409.818, 5.554.517, y 6.063.603), Otros métodos de amplificación que se pueden usar se describen en, las patentes en EE UU Nos. 6.582.938, 5.242.794, 5.494.810, 4.988.617, y publicación en EE UU. No. 20030143599. El ADN también se puede amplificar por PCR múltiple específica de locus o usando ligación de ID de cebador y PCR de cebador único (Véase Kinzler y Vogelstein, NAR (1989) 17:3645-53. Otros métodos disponibles de amplificación, tal como PCR equilibrada (Makrigiorgos, et al. (2002), *Nat Biotechnol,* Vol. 20, pp.936-9), también se pueden usar.

También se pueden usar sondas de inversión molecular ("MIP") para la amplificación de dianas seleccionadas. Las MIP se pueden generar de modo que los extremos de la sonda pre-círculo sean complementarios a regiones que flanquean la región que se va a amplificar. El hueco se puede cerrar por extensión del extremo de la sonda de modo que el complemento de la diana se incorpora a la MIP antes de la ligación de los extremos para formar un círculo cerrado. El círculo cerrado se puede amplificar y detectar por secuenciación o hibridación como se ha divulgado previamente en Hardenbol *et al., Genome Res.* 15:269-275 (2005) y en la patente en EE UU No. 6.858.412.

Los métodos de ligación los conocerán los expertos en la materia, y se describen, por ejemplo, en Sambrook et al. (2001) y el catálogo de New England Biolabs. Los métodos incluyen usar ADN ligasa T4 que cataliza la formación de un enlace fosfodiéster entre extremos 5' fosfato y 3' hidroxilo yuxtapuestos en ADN o ARN dúplex con extremos romos y cohesivos; ADN ligasa Taq que cataliza la formación de un enlace fosfodiéster entre extremos 5' fosfato y 3' hidroxilo yuxtapuestos de dos oligonucleótidos adyacentes que están hibridados a ADN diana complementario; ADN ligasa de E. coli que cataliza la formación de un enlace fosfodiéster entre extremos 5' fosfato y 3' hidroxilo yuxtapuestos en ADN dúplex que contiene extremos cohesivos; y ARN ligasa T4 que cataliza la ligación de un donante de ácido nucleico terminado en 5' fosforilo a un aceptor de ácido nucleico terminado en 3' hidroxilo mediante la formación de un enlace fosfodiéster 3'→5', los sustratos incluyen ARN y ADN monocatenario, así como dinucleósido pirofosfatos; o cualquier otro método descrito en la técnica. Se puede tratar ADN fragmentado con una o más enzimas, por ejemplo, una endonucleasa, antes de la ligación de los ID de cebador a uno o ambos extremos para facilitar la ligación al generar extremos que son compatibles con ligación.

Los métodos para ligar cebadores que comprenden los ID de cebador a fragmentos de ácido nucleico son bien conocidos. Los cebadores pueden ser bicatenarios, monocatenarios, o parcialmente monocatenarios. En algunos aspectos, los cebadores están formados de dos oligonucleótidos que tienen una región de complementariedad, por ejemplo, aproximadamente de 10 a 30, o aproximadamente de 15 a 40 bases de complementariedad perfecta, de modo que cuando los dos oligonucleótidos están hibridados forman una región bicatenaria. Opcionalmente, cualquiera o ambos de los oligonucleótidos pueden tener una región que no sea complementaria al otro oligonucleótido y forma un saliente monocatenario en uno o ambos extremos del cebador. Los salientes monocatenarios pueden ser preferiblemente aproximadamente de 1 a aproximadamente 8 bases, y lo más preferiblemente de aproximadamente 2 a aproximadamente 4. El saliente puede ser complementario al saliente creado por corte con una enzima de restricción para facilitar la ligación de "extremos cohesivos". Los cebadores pueden incluir otras características, tal como sitios de unión a cebador y sitios de restricción. En algunos aspectos el sitio de restricción puede ser para una enzima de restricción de tipo IIS u otra enzima que corte fuera de su secuencia de reconocimiento, tal como EcoP151 (véase, Mucke *et al. J Mol Biol* 2001, 312(4):687-698 y documento US 5.710.000).

Los métodos para usar matrices de mapeo véase, por ejemplo, Aplicaciones de micromatrices para genotipado de SNP, se han descrito en, por ejemplo, las patentes en EE UU No. 6.300.063, 6.361.947, 6.368.799 y las publicaciones de patente en EE UU No. 20040067493, 20030232353, 20030186279, 20050260628, 20070065816 y 20030186280, y Kennedy *et al., Nat. Biotech.* 21:1233-1237 (2003), Matsuzaki *et al., Genome Res.* 14:414-425 (2004), Matsuzaki *et al., Nat. Meth. 1:109-111* (2004) y publicación de patente en EE UU Nos. 20040146890 y 20050042654. Las matrices de mapeo de contenido fijo están disponibles de Affymetrix, por ejemplo, la matriz SNP 6.0 y el sistema de matriz AXIOM®. Paneles seleccionados de SNP y marcadores (por ejemplo, marcadores de número de copia) también se pueden interrogar usando un panel de sondas específicas de locus en combinación con una matriz universal como se describe en Hardenbol *et al., Genome Res.* 15:269-275 (2005) y en la patente en EE UU No. 6.858.412. Matrices de etiquetas universales y kits de reactivos para realizar tal genotipado específico de locus usando paneles de sondas de inversión molecular (MIP) a medida están disponibles de Affymetrix.

Los métodos para analizar el número de copia de cromosomas usando matrices de mapeo se divulgan, por ejemplo, en Bignell et al., *Genome Res.* 14:287-95 (2004), Lieberfarb, *et al., Cancer Res.* 63:4781-4785 (2003), Zhao *et al., Cancer Res.* 64:3060-71 (2004), Huang *et al., Hum Genomics* 1:287-299 (2004), Nannya *et al., Cancer Res.* 65:6071-6079 (2005), Slater *et al., Am. J. Hum. Genet.* 77:709-726 (2005) e Ishikawa *et al., Biochem. and Biophys. Res. Comm.,* 333:1309-1314 (2005). Se divulgan métodos implementados en ordenador para la estimación del número de copia basados en la intensidad de hibridación en las publicaciones de patente en EE UU Nos. 20040157243, 20050064476, 20050130217, 20060035258, 20060134674 y 20060194243.

Los métodos para realizar ensayos de hibridación de polinucleótidos se han desarrollado bien en la técnica. Los procedimientos y condiciones de ensayo de hibridación variarán dependiendo de la aplicación y se seleccionan según métodos de unión generales conocidos, incluyendo los referenciados en: Maniatis et al. *Molecular Cloning: A Laboratory Manual* (2ª Ed. Cold Spring Harbor, N.Y, 1989); Berger y Kimmel *Methods in Enzymology,* Vol. 152, *Guide to Molecular Cloning Techniques* (Academic Press, Inc., San Diego, CA, 1987); Young y Davis, *P.N.A.S,* 80: 1194 (1983). Los métodos y aparatos para llevar a cabo reacciones de hibridación repetidas y controladas se han descrito en las patentes en EE UU Nos. 5.871.928, 5.874.219, 6.045.996 y 6.386.749, 6.391.623.

La presente divulgación también contempla la detección de señal de hibridación entre ligandos en ciertas formas de realización preferidas. Véase, las patentes en EE UU 5.143.854, 5.578.832, 5.631.734, 5.834.758, 5.936.324, 5.981.956, 6.025.601, 6.141.096, 6.185.030, 6.201.639, 6.218.803, y 6.225.625 en la publicación de patente en EE UU No. 20040012676 y en la solicitud PCT PCT/US99/06097 (publicada como W099/47964).

Los métodos y aparatos para la detección de señal y procesamiento de datos de intensidad se divulgan en, por ejemplo, las patentes en EE UU 5.143.854, 5.547.839, 5.578.832, 5.631.734, 5.800.992, 5.834.758, 5.856.092, 5.902.723, 5.936.324, 5.981.956, 6.025.601, 6.090.555, 6.141.096, 6.185.030, 6.201.639, 6.218.803, y 6.225.625 en las publicaciones de patente en EE UU No. 20040012676 y 20050059060 y en la solicitud PCT PCT/US99/06097 (publicada como W099/47964).

La práctica de la presente divulgación también puede emplear métodos de biología, software y sistemas convencionales. Los productos de software de ordenador de la divulgación típicamente incluyen medio legible por ordenador que tiene instrucciones ejecutables por ordenador para realizar las etapas lógicas del método de la divulgación. Los medios legibles por ordenador incluyen disquete, CD-ROM/DVD/DVD-ROM, unidad de disco duro, memoria flash, ROM/RAM, cintas magnéticas, etc. Las instrucciones ejecutables por ordenador pueden estar escritas en un lenguaje informático adecuado o combinaciones de varios lenguajes. Los métodos de biología computacional básicos se describen en, por ejemplo, Setubal y Meidanis et al., *Introduction to Computational Biology Methods* (PWS Publishing Company, Boston, 1997); Salzberg, Searles, Kasif, (Ed.), *Computational Methods in Molecular Biology,* (Elsevier, Ámsterdam, 1998); Rashidi y Buehler, *Bioinformatics Basics: Application in Biological Science and Medicine* (CRC Press, Londres, 2000) y Ouelette y Bzevanis *Bioinformatics: A Practical Guide for Analysis of Gene and Proteins* (Wiley & Sons, Inc., 2ª ed., 2001). Véase también el documento US 6.420.108.

La presente divulgación también puede hacer uso de varios productos y software de programas informáticos para una variedad de fines, tal como diseño de sondas, gestión de datos, análisis y operación de instrumentos. Véase, las patentes en EE UU No. 5.593.839, 5.795.716, 5.733.729, 5.974.164, 6.066.454, 6.090.555, 6.185.561, 6.188.783, 6.223.127, 6.229.911 y 6.308.170. También se pueden usar métodos informáticos relacionados con genotipado que usan análisis de micromatrices de alta densidad en los métodos presentes, véase, por ejemplo, las publicaciones de patente en EE UU No. 20050250151, 20050244883, 20050108197, 20050079536 y 20050042654. Además, la presente divulgación puede tener formas de realización preferidas que incluyen métodos para proporcionar información genética sobre redes tal como la Internet como se muestra en las publicaciones de patente en EE UU Nos. 20030097222, 20020183936, 20030100995, 20030120432, 20040002818, 20040126840, y 20040049354.

Un alelo se refiere a una forma específica de una secuencia genética (tal como un gen) en una célula, un individuo o en una población, la forma específica se diferencia de otras formas del mismo gen en la secuencia de al menos uno, y con frecuencia más de uno, sitios variantes en la secuencia del gen. Las secuencias en estos sitios variantes que diferencian entre diferentes alelos se denominan “varianzas”, “polimorfismos” o “mutaciones”. En cada localización cromosómica específica autosómica o “locus” un individuo posee dos alelos, uno heredado de un padre y uno heredado del otro padre, por ejemplo, uno de la madre y uno del padre. Un individuo es “heterocigoto” en un locus si tiene dos alelos diferentes en ese locus. Un individuo es “homocigoto” en un locus si tiene dos lelos idénticos en ese locus.

El término “polimorfismo” como se usa en el presente documento se refiere a la aparición de dos o más secuencias alternativas genéticamente determinadas o alelos en una población. Un marcador o sitio polimórfico es el locus en el que se produce divergencia. En algunos casos, los marcadores polimórficos se producen a una frecuencia de menos del 0,5%. En algunos casos, los marcadores polimórficos se producen a una frecuencia de menos del 1%. En algunos casos, los marcadores polimórficos se producen a una frecuencia de menos del 2%. En algunos casos, los marcadores polimórficos se producen a una frecuencia de menos del 5%. En algunos casos, los marcadores polimórficos se producen a una frecuencia de más del 1%. En algunos casos, los marcadores polimórficos se producen a una frecuencia de más del 5%. En algunos casos, los marcadores polimórficos se producen a una frecuencia de más del 10%. En algunos casos, los marcadores polimórficos se producen a una frecuencia de más del 20%. En algunos casos, los marcadores polimórficos se producen a una frecuencia de más del 30%. En algunos casos, los marcadores preferidos tienen al menos dos alelos, cada uno se produce a una frecuencia de más del 1% y más preferiblemente mayor del 10% o el 20% en una población seleccionada. En algunos casos, los marcadores polimórficos preferidos comprenden secuencias víricas o bacterianas y se producen a una frecuencia de menos del 5%, y más preferiblemente, menor del 1% en una población seleccionada. Un polimorfismo puede comprender uno o más cambios de bases, una inserción, una repetición, o una deleción de una o más bases. Variantes de número de copia (CNV), transversiones y otras reorganizaciones también son formas de variación genética. Los marcadores polimórficos incluyen polimorfismos de longitud de fragmento de restricción, número variable de repeticiones en tándem (VNTR), regiones hipervariables, minisatélites, repeticiones de dinucleótidos, repeticiones de trinucleótidos, repeticiones de tetranucleótidos, repeticiones de secuencias simples, y elementos de inserción tal como Alu. La forma alélica que se produce con mayor frecuencia en una población seleccionada algunas veces se denomina forma de tipo salvaje. Los organismos diploides pueden ser homocigotos o heterocigotos para formas alélicas. Un polimorfismo dialélico tiene dos formas. Un polimorfismo trialélico tiene tres formas. Los polimorfismos de nucleótido único (SNP) son una forma de polimorfismos. Los SNP son un tipo común de variación genética humana y son útiles en la realización de estudios de asociación de amplitud genómica (GWAS). Se puede usar GWAS, por ejemplo, para el análisis de rutas biológicas, véase, Wang y Hakonarson, Nat. Rev. Genet. 2010, 11:843-854.

El término genotipado se refiere a la determinación de la información genética que porta un individuo en una o más posiciones en el genoma. Por ejemplo, el genotipado puede comprender la determinación de qué alelo o alelos porta un individuo para un único SNP o la determinación de qué alelo o alelos porta un individuo para una pluralidad de SNP o CNV. Un individuo diploide puede ser homocigoto para cada uno de los dos alelos posibles (por ejemplo, AA o BB) o heterocigoto (por ejemplo, AB). Para información adicional respecto al genotipado y la estructura del genoma véase, *Color Atlas of Genetics,* Ed. Passarge, Thieme, Nueva York, NY (2001).

Las células normales que son heterocigotas en uno o más loci pueden dar lugar a células tumorales que son homocigotas en esos loci. Esta pérdida de heterocigosidad (LOH) puede resultar de deleción estructural de genes

normales o pérdida del cromosoma que porta el gen normal, recombinación mitótica entre genes normal y mutante, seguido por la formación de células hijas homocigotas para genes delecionados o inactivados (mutante); o pérdida de cromosoma con el gen normal y duplicación del cromosoma con el gen delecionado o inactivado (mutante). LOH puede ser neutro para copia o puede resultar de una deleción o amplificación.

El término "matriz" como se usa en el presente documento se refiere a una colección intencionadamente creada de moléculas que se puede preparar de forma sintética o biosintética. Las moléculas en la matriz pueden ser idénticas o diferentes entre sí. La matriz puede asumir una variedad de formatos, por ejemplo, bibliotecas de moléculas solubles; bibliotecas de compuestos anclados a bolas de resinas, chips de sílice, micropartículas, nanopartículas u otros soportes sólidos.

El término "complementario" como se usa en el presente documento se refiere a la hibridación o emparejamiento de bases entre nucleótidos o ácidos nucleicos, tal como, por ejemplo, entre dos hebras de una molécula de ADN bicatenaria o entre un cebador oligonucleotídico y un sitio de unión a cebador en un ácido nucleico monocatenario que se va a secuenciar o amplificar. Véase, M. Kanehisa Nucleic Acids Res. 12:203 (1984).

El término "variación del número de copia" o "CNV" se refiere a diferencias en el número de copia de información genética. En muchos aspectos se refiere a diferencias en el número de copia por genoma de una región genómica. Por ejemplo, en un organismo diploide el número de copia esperado para regiones genómicas autosómicas es 2 copias por genoma. Tales regiones genómicas deben estar presentes en 2 copias por célula. Para una revisión reciente véase Zhang *et al. Annu. Rev. Genomics Hum. Genet.* 2009. 10:451-81. CNV es una fuente de diversidad genética en seres humanos y se puede asociar con trastornos complejos y enfermedad, por ejemplo, alterando la dosis génica, disrupción de genes o fusión de genes. También pueden representar variantes polimórficas benignas. Las CNV pueden ser grandes, por ejemplo, mayores de 1 Mb, pero muchas son más pequeñas, por ejemplo, entre 100 pb y 1 Mb. Se han descrito más de 38.000 CNV mayores de 100 pb (y menores de 3 Mb) en seres humanos. Junto con los SNP estas CNV representan una cantidad significativa de variación fenotípica entre individuos. Además de tener impactos perjudiciales, por ejemplo, causan enfermedad, también pueden producir variación ventajosa.

La PCR digital es una técnica donde una dilución limitante de la muestra se hace a través de un gran número de reacciones de PCR separadas de modo que la mayoría de las reacciones no tienen moléculas de molde y dan un resultado de amplificación negativo. Esas reacciones que son positivas en el punto final de la reacción se cuentan como moléculas de molde individuales presentes en la muestra original en una relación de 1 a 1. Véase, Kalina et al. NAR 25:1999-2004 (1997) y Vogelstein y Kinzler, PNAS 96:9236-9241 (1999). Este método es un método de recuento absoluto donde las soluciones se reparten en envases hasta que hay una probabilidad media de una molécula por dos envases o cuando $P_0 = (1\text{-}e^{-n/c}) = ½$; donde *n* es el número de moléculas y *c* es el número de envases, o *n/c* es 0,693. Se asume el reparto cuantitativo, y el intervalo dinámico está regido por el número de envases disponibles para la separación estocástica. Las moléculas se detectan después por PCR y el número de envases positivos se cuenta. Cada amplificación con éxito se cuenta como una molécula, independiente de la cantidad real de producto. Las técnicas basadas en PCR tienen la ventaja adicional de solo contar moléculas que se pueden amplificar, por ejemplo, que son relevantes para la etapa de PCR masivamente paralela en el flujo de trabajo de secuenciación. Puesto que la PCR digital tiene sensibilidad de molécula única, solo se requieren unos pocos cientos de moléculas de genoteca para la cuantificación precisa. La eliminación del cuello de botella de la cuantificación reduce el requisito de aporte de muestra de microgramos a nanogramos o menos, abriendo el camino para muestras diminutas y/o preciadas sobre plataformas de secuenciación de nueva generación sin distorsionar los efectos de la preamplificación. La PCR digital se ha usado para cuantificar genotecas de secuenciación para eliminar incertidumbre asociada con la construcción y aplicación de curvas estándar a cuantificación basada en PCR y permite la secuenciación directa sin carreras de titulación. Véase, White *et al. BMC Genomics* 10: 116 (2009). Para variar el intervalo dinámico, se puede usar microfabricación, para aumentar sustancialmente el número de envases. Véase, Fan et al. *Am J Obstet Gynecol* 200, 543 el (Mayo, 2009).

De forma similar, en marcaje estocástico, se cumplen las mismas condiciones estadísticas cuando $P_0 = (1\text{-}e^{-n/m}) = ½$; donde *m* es el número de ID de cebador, y la mitad de los ID de cebador se usarán al menos una vez cuando *n/m* = 0,693. El intervalo dinámico está regido por el número de ID de cebador usados, y el número de ID de cebador se puede aumentar fácilmente para extender el intervalo dinámico. El número de envases en la PCR digital desempeña el mismo papel que el número de ID de cebador en marcaje estocástico y sustituyendo envases por ID de cebador se pueden aplicar idénticas ecuaciones estadísticas. Usando los principios de separación física, la PCR digital expande estocásticamente moléculas idénticas en *espacio físico*, mientras que el principio que rige el marcaje estocástico se basa en identidad y expande moléculas idénticas en *espacio de identidad*. Véase la solicitud PCT PCT/US11/65291.

El término "hibridación" como se usa en el presente documento se refiere al proceso en el que dos polinucleótidos monocatenarios se unen no covalentemente para formar un polinucleótido bicatenario; la hibridación tricatenaria también es teóricamente posible. El polinucleótido (habitualmente) bicatenario resultante es un "híbrido". La proporción de la población de polinucleótidos que forma híbridos estables se denomina en el presente documento el "grado de hibridación". Las hibridaciones se pueden realizar en condiciones rigurosas, por ejemplo, a una concentración de sal de no más de 1 M y a una temperatura de al menos 25ºC. Por ejemplo, las condiciones SSPE (NaCl 750 mM, fosfato de Na 50 mM, EDTA 5 mM, pH 7,4) 5X y temperatura de 25-30ºC son adecuadas para hibridaciones de sonda

específica de alelo. Para condiciones rigurosas, véase, por ejemplo, Sambrook, Fritsche y Maniatis. "Molecular Cloning A laboratory Manual" 2ª Ed. Cold Spring Harbor Press (1989). En algunos aspectos, las concentraciones de sal para hibridación son preferiblemente entre aproximadamente 200 mM y aproximadamente 1 M o entre aproximadamente 200 mM y aproximadamente 500 mM. Las temperaturas de hibridación pueden ser tan bajas como 5ºC, pero típicamente son mayores de 22ºC, más típicamente mayores de aproximadamente 30ºC, y preferiblemente en exceso de aproximadamente 37ºC. Los fragmentos más largos pueden requerir mayores temperaturas de hibridación para la hibridación específica. Como otros factores pueden afectar la rigurosidad de la hibridación, incluyendo la composición de bases y la longitud de las hebras complementarias, la presencia de solventes orgánicos y el grado de mal apareamiento de bases, la combinación de parámetros es más importante que la medida absoluta de cualquiera solo.

El término “ARNm” o algunas veces referido por “transcritos de ARNm” como se usa en el presente documento, incluye, pero no está limitado a transcrito(s) de pre-ARNm, intermedios de procesamiento de transcrito, ARNm maduro(s) listo(s) para traducción y transcritos del gen o genes, o ácidos nucleicos derivados del/de los transcrito(s) de ARNm. El procesamiento de transcritos puede incluir ayuste, edición y degradación. Como se usa en el presente documento, un ácido nucleico derivado de un transcrito de ARNm se refiere a un ácido nucleico para cuya síntesis el transcrito de ARNm o una subsecuencia del mismo ha servido finalmente como un molde. Por tanto, un ADNc por transcripción inversa de un ARNm, un ARN transcrito de ese ADNc, un ADN amplificado del ADNc, un ARN transcrito del ADN amplificado, etc., todos derivan del transcrito de ARNm y la detección de tales productos derivados es indicativa de la presencia y/o abundancia del transcrito original en una muestra. Por tanto, las muestras derivadas de ARNm incluyen, pero no están limitadas a, transcritos de ARNm del gen o genes, ADNc por transcripción inversa del ARNm, ARNc transcrito del ADNc, ADN amplificado de los genes, ARN transcrito del ADN amplificado, y similares.

También se expresan otras clases de ARN incluyendo, por ejemplo, ARN ribosómico, ARNnp, miARN, y ARNip. Evidencia reciente sugiere que el transcriptoma humano contiene muchos transcritos de ARN funcional que no se traducen a proteínas. Estos ARN no codificantes se han reconocido como importantes en un entendimiento más completo de la biología. Los miARN maduros son dúplex de ARN relativamente pequeños (21-23 nucleótidos) que actúan como represores de traducción de expresión de proteínas. La hebra guía de un miARN interacciona con proteínas para formar complejos de silenciamiento inducido por ARN (RISC) en la célula. Estos complejos de ribonucleoproteína específicos de secuencia se unen a ARNm diana típicamente en la 3’UTR y pueden posteriormente silenciar la expresión génica ya sea mediante degradación de ARNm dirigida o simplemente secuestrando el ARNm diana en una forma ineficaz (Lee et al., Cell (1993), 75: 843-854; Bartel, Cell (2009), 136: 215-233). Se ha demostrado que la regulación basada en miARN desempeña un papel significativo en procesos celulares rutinarios incluyendo metabolismo (Esau et al, Cell Met. 2006, v.3, p 87-98), desarrollo (Carthew et al., Cell 2009, v.137, p. 273-282), e incluso apoptosis (Cheng et al, Nucl. Acids Res. 2005, v.33, p1290-1297). Investigación adicional ha revelado que los miARN desempeñan papeles críticos en diversos procesos de enfermedad tal como hepatitis C (Jopling et al., Science 2005, v.309, p. 1577-1581), diabetes (Poy et al., Nature 2004, v.432, p. 226-230), y de forma más notable múltiples tipos de cáncer (Hammond, Can. Chemo. Pharma. 2006 v.58, s63-s68; Calin et al., Cancer Res. 2006, v.66, p. 7390-7394) incluyendo leucemia (Calin et al., PNAS 2002, v.101, p. 2999-3004) y glioma (Corsten et al., Cancer Res. 2007, v.67, p. 8994-9000). Más de mil miARN se han identificado ahora en animales, pero solo unos pocos miARN individuales se han ligado a funciones específicas. Los métodos de la divulgación divulgados en el presente documento se pueden usar para etiquetar ARN no codificantes reguladores relativamente cortos, tal como micro ARN (miARN), ARN que interaccionan con Piwi (piARN), ARNnop, ARNnp, ARNmo PAR, ARNsd, ARNs-tel, crasiARN y ARN interferentes pequeños (ARNip). Los métodos de la divulgación también se pueden usar para etiquetar ARN no codificantes largos (ARNnc largos), ARNt no codificantes tradicionales y ARN ribosómico (ARNr).

El término “ácido nucleico” como se usa en el presente documento se refiere a una forma polimérica de nucleótidos de cualquier longitud, ya sean ribonucleótidos, desoxirribonucleótidos o ácidos peptidonucleicos (APN), que comprende bases de purina y pirimidina, u otras bases nucleotídicas naturales, química o bioquímicamente modificadas, no naturales o derivadas. El esqueleto del polinucleótido puede comprender azúcares y grupos fosfato, como se puede encontrar típicamente en ARN o ADN, o azúcar o grupo fosfato modificados o sustituidos. Un polinucleótido puede comprender nucleótidos modificados, tal como nucleótidos metilados y análogos de nucleótidos. La secuencia de nucleótidos se puede interrumpir por componentes no nucleotídicos. Por tanto, los términos nucleósido, nucleótido, desoxinucleósido y desoxinucleótido, en general incluyen análogos tales como los descritos en el presente documento. Estos análogos son esas moléculas que tienen algunas características estructurales en común con un nucleósido o nucleótido natural de modo que cuando se incorporan a una secuencia de ácido nucleico u oligonucleósido, permiten la hibridación con una secuencia de ácido nucleico natural en solución. Típicamente, estos análogos derivan de nucleósidos y nucleótidos naturales al sustituir y/o modificar la base, la ribosa o la fracción fosfodiéster. Los cambios se pueden hacer a medida para estabilizar o desestabilizar la formación de híbridos o aumentar la especificidad de hibridación con una secuencia de ácido nucleico complementaria según se desee.

El término “oligonucleótido” o algunas veces denominado “polinucleótido” como se usa en el presente documento se refiere a un ácido nucleico que varía desde al menos 2, preferiblemente al menos 8, y más preferiblemente al menos 20 nucleótidos de longitud o un compuesto que específicamente hibrida con un polinucleótido. Los polinucleótidos de la presente divulgación incluyen secuencias de ácido desoxirribonucleico (ADN) o ácido ribonucleico (ARN) que se pueden aislar de fuentes naturales, producir recombinantemente o sintetizar artificialmente y miméticos de los mismos. Un ejemplo adicional de un polinucleótido de la presente divulgación puede incluir análogos no naturales que pueden

aumentar la especificidad de hibridación, por ejemplo, enlaces de ácido peptidonucleico (APN) y enlaces de ácido nucleico bloqueado (ANB). Los enlaces ANB son análogos de nucleótidos conformacionalmente restringidos que se unen a la diana complementaria con una mayor temperatura de fusión y mayor discriminación de malos emparejamientos. Otras modificaciones que se pueden incluir en sondas incluyen: 2'OMe, 2'Oalilo, 2'O-propargilo, 2'O-alquilo, 2'-fluoro, 2'-arabino, 2'-xilo, 2'-fluoroarabino, fosforotioato, fosforoditioato, fosforamidatos, 2'amino, pirimidina 5-alquil sustituida, pirimidina 5-halo sustituida, purina sustituida con alquilo, purina sustituida con halo, nucleótidos bicíclicos, 2'MOE, moléculas de tipo ANB y derivados de los mismos. La divulgación también abarca situaciones en las que hay un emparejamiento de bases no tradicional tal como emparejamiento de bases de Hoogsteen que se ha identificado en ciertas moléculas de ARNt y postulado que existe en una triple hélice. "Polinucleótido" y "oligonucleótido" se usan de forma intercambiable en esta solicitud.

El término "cebador" como se usa en el presente documento se refiere a un oligonucleótido bicatenario, monocatenario o parcialmente monocatenario. En algunas formas de realización, los cebadores son capaces de actuar como un punto de iniciación para síntesis de ácido nucleico dirigida por molde en condiciones adecuadas, por ejemplo, tampón y temperatura, en presencia de cuatro nucleósidos trifosfato diferentes y un agente para polimerización, tal como, por ejemplo, ADN o ARN polimerasa o transcriptasa inversa. La longitud del cebador, en cualquier caso, depende de, por ejemplo, el uso pretendido del cebador, y en general varía desde 15 a 100 nucleótidos. Las moléculas de cebadores cortos en general requieren temperaturas más frías para formar complejos híbridos suficientemente estables con el molde. Un cebador no necesita reflejar la secuencia exacta del molde, pero deber lo suficientemente complementario para hibridar con tal molde. El sitio cebador es el área del molde con el que hibrida el cebador. El par de cebadores es un conjunto de cebadores que incluye un cebador 5' anterior que hibrida con el extremo 5' de la secuencia que se va a amplificar. Como se usa en el presente documento, el cebador puede comprender una secuencia específica diana y una secuencia ID de cebador. El cebador puede comprender además una secuencia código de barras. La secuencia código de barras se puede usar para identificar la presencia de una secuencia ID de cebador. El cebador también puede comprender una secuencia cebadora de PCR. La secuencia cebadora de PCR se puede usar para iniciar la amplificación de una molécula de ácido nucleico etiquetado.

El término "sonda" como se usa en el presente documento se refiere a una molécula inmovilizada a una superficie que puede ser reconocida por una diana particular. Véase la patente en EE UU No. 6.582.908 para un ejemplo de matrices que tienen todas las combinaciones posibles de sondas con 10, 12 y más bases. Los ejemplos de sondas que se pueden investigar por esta divulgación incluyen, pero no están restringidas a, agonistas y antagonistas para receptores de membrana celular, toxinas y venenos, epítopos víricos, hormonas (por ejemplo, péptidos opioides, esteroides, etc.), receptores de hormonas, péptidos, enzimas, sustratos de enzimas, cofactores, fármacos, lectinas, azúcares, oligonucleótidos, ácidos nucleicos, oligosacáridos, proteínas y anticuerpos monoclonales.

El término "soporte sólido", "soporte" y "sustrato" como se usa en el presente documento se usan de forma intercambiable y se refiere a un material o grupo de materiales que tienen una superficie o superficies rígidas o semirrígidas. En muchas formas de realización, al menos una superficie del soporte sólido será sustancialmente plana, aunque en algunas formas de realización puede ser deseable separar físicamente las regiones de síntesis para diferentes compuestos con, por ejemplo, pocillos, regiones subidas, alfileres, zanjas grabadas, o similares. Según otras formas de realización, el/los soporte(s) sólido(s) tomará(n) la forma de bolas, resinas, geles, microesferas, u otras configuraciones geométricas. Véase, la patente en EE UU 5.744.305 y las publicaciones de patente en EE UU No. 20090149340 y 20080038559 para sustratos ejemplares.

El término "ID de cebador" como se usa en el presente documento se refiere a la información que se añade. Se pueden usar genotecas de cebadores que tienen una diversidad de ID de cebador únicos, por ejemplo, aproximadamente 1.000, aproximadamente 5.000, aproximadamente 10.000, aproximadamente 100.000 o más de 100.000 para identificar exclusivamente apariciones de especies diana marcando de esta manera cada especie con un identificador que se puede usar para distinguir entre dos dianas de otra manera idénticas o casi idénticas. Por ejemplo, cada ID de cebador puede ser una cadena corta de nucleótidos que se puede unir a diferentes copias de un ARNm, por ejemplo, un primer ID de cebador puede ser 5'GCATCTTC3' y un segundo puede ser 5'CAAGTAA3'. Cada uno tiene una identidad única que se puede determinar determinando la identidad y orden de las bases en el ID de cebador.

Aunque los ácidos nucleicos se usan en todo como una forma de realización preferida de ID de cebador, un experto en la materia apreciará que un número de tipos de moléculas o productos que se pueden generar con la diversidad necesaria se pueden usar como ID de cebador. Los ID de cebador deben ser compuestos, estructuras o elementos que son sensibles para al menos un método de detección que permite la discriminación entre diferentes ID de cebador y debe ser asociable en algunos medios con los elementos que se van a contar. Por ejemplo, un conjunto de ID de cebador puede estar compuesto de una colección de diferentes nanocristales semiconductores, compuestos metálicos, péptidos, anticuerpos, moléculas pequeñas, isótopos, partículas o estructuras que tienen diferentes formas, colores, o patrones de difracción asociados con los mismos o embebidos en los mismos, cadenas de números, fragmentos aleatorios de proteínas o ácidos nucleicos, o diferentes isótopos (véase, Abdelrahman, A.I. et al. *Journal of Analytical Atomic Spectrometry* 25 (3):260-268, 2010 para uso de bolas de poliestireno que contienen metal como estándares para citometría de masa). Los grupos de ID de cebador se pueden repartir en distintos conjuntos que se pueden unir a mezclas de muestras separadas y después combinar para análisis posterior. Por ejemplo, un conjunto de 1.000.000 de diferentes ID de cebador se podrían dividir físicamente en 10 conjuntos de 100.000 ID de cebador

diferentes y cada uno se podría usar para ID de cebador de una mezcla diferente. La identidad de los ID de cebador de cada conjunto se puede usar como una indicación de la fuente original. Se puede facilitar el recuento de las múltiples muestras en paralelo.

En algunas formas de realización el ID de cebador también se puede usar junto con un código de barras, que puede tener 2-10 nucleótidos, por ejemplo, 2, 3, 4, 5, 6, 7, 8, 9, o 10 nucleótidos. El código de barras se puede unir directamente al ID de cebador o puede haber una secuencia intermedia entre el código de barras y el ID de cebador. El código de barras puede representar una fecha, tiempo o localización de análisis; un ensayo clínico; una fecha, tiempo o localización de recogida; un número de paciente; un número de muestra; una especie; una subespecie; un subtipo; una pauta terapéutica; o un tipo de tejido. En una forma de realización no limitante, tanto el ID de cebador como el código de barras son monocatenarios. Un código de barras de 3 nucleótidos que representa diferentes fechas de estudio se ejemplifica en el presente documento.

El término “marcador detectable” como se usa en el presente documento se refiere a cualquier fracción química unida a un nucleótido, polímero de nucleótidos, o factor de unión a ácidos nucleicos, en donde la unión puede ser covalente o no covalente. Preferiblemente, el marcador es detectable y hace el nucleótido o polímero de nucleótidos detectable para el practicante de la invención. Los marcadores detectables que se pueden usar en combinación con los métodos divulgados en el presente documento incluyen, por ejemplo, un marcador fluorescente, un marcador quimioluminiscente, un extinguidor, un marcador radioactivo, biotina y oro, o combinaciones de los mismos. Los marcadores detectables incluyen moléculas luminiscentes, fluorocromos, agentes de extinción fluorescente, moléculas coloreadas, radioisótopos o centelleantes. Los marcadores detectables también incluyen cualquier molécula enlazadora útil (tal como biotina, avidina, estreptavidina, HRP, proteína A, proteína G, anticuerpos o fragmentos de los mismos, Grb2, polihistidina, Ni^{2+}, etiquetas FLAG, etiquetas myc), metales pesados, enzimas (los ejemplos incluyen, fosfatasa alcalina, peroxidasa y luciferasa), donantes/aceptores de electrones, ésteres de acridinio, colorantes y sustratos calorimétricos. También se prevé que un cambio en masa se pueda considerar un marcador detectable, como es el caso de detección por resonancia de plasmón de superficie. El experto en la materia reconocería fácilmente marcadores detectables útiles que no se mencionan anteriormente, que se pueden emplear en la operación de la presente divulgación. En algunos casos, se usan marcadores detectables con cebadores. En algunos casos, se usan marcadores detectables con los ID de cebador. En algunos casos, se usan marcadores detectables con la molécula molde de ácido nucleico. En algunos casos, se usan marcadores detectables para detectar amplicones etiquetados. En algunos casos, se usan marcadores detectables para detectar la molécula molde de ácido nucleico.

El término “secuencia consenso” como se usa en el presente documento se refiere a una secuencia formada a partir de dos o más secuencias que contienen un ID de cebador idéntico. En algunos casos, una secuencia consenso es la variante más común de una molécula de ácido nucleico.

En el presente documento se divulga un método para determinar la diversidad génica de una muestra que comprende: (a) proporcionar una muestra que comprende una molécula molde de ácido nucleico; (b) unir un cebador que comprende un ID de cebador a cada molécula molde de ácido nucleico que se va a analizar para generar un molde de ácido nucleico etiquetado, en donde cada molde de ácido nucleico etiquetado se une a un ID de cebador único; (c) amplificar el molde de ácido nucleico etiquetado para producir amplicones etiquetados; y (d) detectar los amplicones etiquetados, determinando de esta manera la diversidad genética de una muestra. En algunas formas de realización, el ID de cebador comprende una secuencia degenerada. En algunas formas de realización, el ID de cebador comprende una secuencia semidegenerada. En algunas formas de realización, el ID de cebador comprende una secuencia mixta. En algunas formas de realización, el ID de cebador comprende una secuencia ambigua. En algunas formas de realización, el ID de cebador comprende una secuencia titubeante. En algunas formas de realización, el ID de cebador comprende una secuencia aleatoria. En algunas formas de realización, el ID de cebador comprende una secuencia predeterminada. En algunas formas de realización de ID de cebador está unido al molde por ligación. En algunas formas de realización de ID de cebador está unido al molde por hibridación. En algunas formas de realización de ID de cebador está unido al molde a través de PCR. En algunas formas de realización, se analiza al menos una molécula molde. En algunas formas de realización, se analizan al menos dos moléculas molde diferentes. En algunas formas de realización, detectar los amplicones etiquetados comprende secuenciar los amplicones etiquetados. La secuenciación de los amplicones etiquetados se puede producir por una variedad de métodos, incluyendo, pero no limitados al método de secuenciación de Maxam-Gilbert, el método de secuenciación del dideoxi de Sanger, el método de secuenciación de terminador colorante, pirosecuenciación, secuenciación de ADN con cebador múltiple, secuenciación aleatoria, y desplazamiento sobre el cebador. En algunas formas de realización, la secuenciación comprende pirosecuenciación. En algunas formas de realización, detectar los amplicones etiquetados comprende además contar un número de diferentes ID de cebador asociados con los amplicones etiquetados, en donde el número de diferentes ID de cebador asociados con los amplicones etiquetados refleja el número de moldes muestreados. En algunas formas de realización el método comprende además formar una secuencia consenso para amplicones etiquetados que comprenden el mismo ID de cebador. En algunas formas de realización, el molde de ácido nucleico comprende un molde de ADN. En algunas formas de realización, el molde de ácido nucleico comprende un molde de ARN. En algunas formas de realización, amplificar comprende un método basado en PCR. En algunas formas de realización, el método basado en PCR comprende PCR. En algunas formas de realización, el método basado en PCR comprende PCR cuantitativa. En algunas formas de realización, el método basado en PCR comprende PCR en

emulsión. En algunas formas de realización, el método basado en PCR comprende PCR en gota. En algunas formas de realización, el método basado en PCR comprende PCR de inicio en caliente. En algunas formas de realización, el método basado en PCR comprende PCR in situ. En algunas formas de realización, el método basado en PCR comprende PCR inversa. En algunas formas de realización, el método basado en PCR comprende PCR multiplex. En algunas formas de realización, el método basado en PCR comprende PCR de número variables de repeticiones en tándem (VNTR). En algunas formas de realización, el método basado en PCR comprende PCR asimétrica. En algunas formas de realización, el método basado en PCR comprende PCR larga. En algunas formas de realización, el método basado en PCR comprende PCR anidada. En algunas formas de realización, el método basado en PCR comprende PCR hemianidada. En algunas formas de realización, el método basado en PCR comprende PCR touchdown. En algunas formas de realización, el método basado en PCR comprende PCR de ensamblaje. En algunas formas de realización, el método basado en PCR comprende PCR en colonia. En algunas formas de realización, amplificar comprende un método no basado en PCR. En algunas formas de realización, el método no basado en PCR comprende amplificación por desplazamiento múltiple (MDA). En algunas formas de realización, el método no basado en PCR comprende amplificación mediada por transcripción (TMA). En algunas formas de realización, el método no basado en PCR comprende amplificación basada en secuencia de ácido nucleico (NASBA). En algunas formas de realización, el método no basado en PCR comprende amplificación por desplazamiento de la hebra (SDA). En algunas formas de realización, el método no basado en PCR comprende SDA en tiempo real. En algunas formas de realización, el método no basado en PCR comprende amplificación por círculo rodante. En algunas formas de realización, el método no basado en PCR comprende amplificación de círculo a círculo.

Las tecnologías de secuenciación de nueva generación adecuadas están ampliamente disponibles para uso en relación con los métodos descritos en el presente documento. Los ejemplos incluyen la plataforma 454 de Life Sciences (Roche, Branford, CT) (Margulies *et al.* 2005 *Nature,* 437, 376-380); el analizador de genoma de Illumina, Ensayo de metilación GoldenGate, o ensayos de metilación Infinium, es decir, la matriz de metilación Infinium HumanMethylation 27K BeadArray o VeraCode GoldenGate (Illumina, San Diego, CA; Bibkova *et al.,* 2006, *Genome Res.* 16, 383-393; patentes en EE UU No. 6.306.597 y 7.598.035 (Macevicz); 7.232.656 (Balasubramanian *et al.*)); o Secuenciación de ADN por ligación, sistema SOLiD (Applied Biosystems/Life Technologies; patentes en EE UU No. 6.797.470, 7.083.917, 7.166.434, 7.320.865, 7.332.285, 7.364.858, y 7.429.453 (Barany *et al.);* o la tecnología de secuenciación de ADN de molécula única Helicos True Single Molecule DNA (Harris *et al.,* 2008 *Science,* 320, 106-109; patentes en EE UU No. 7.037.687 y 7.645.596 (Williams *et al.*); 7.169.560 (Lapidus *et al.*); 7.769.400 (Harris)), la tecnología de molécula única en tiempo real (SMRT™) de Pacific Biosciences, y secuenciación (Soni y Meller, 2007, *Clin. Chem.* 53, 1996-2001). Estos sistemas permiten la secuenciación de muchas moléculas de ácido nucleico aisladas de una muestra en altos órdenes de multiplexación de una manera paralela (Dear, 2003, *Brief Funct. Genomic Proteomic,* 1(4), 397-416 y McCaughan y Dear, 2010, *J. Pathol.,* 220, 297-306). Cada una de estas plataformas permite la secuenciación de moléculas únicas clonalmente expandidas o no amplificadas de fragmentos de ácido nucleico. Ciertas plataformas implican, por ejemplo, (i) secuenciar por ligación de sondas modificadas con colorante (incluyendo ligación cíclica y corte), (ii) pirosecuenciación, y (iii) secuenciación de molécula única.

La pirosecuenciación es un método de secuenciación de ácidos nucleicos basado en secuenciar por síntesis, que se basa en la detección de un pirofosfato liberado en la incorporación de nucleótido. En general, secuenciar por síntesis implica sintetizar, un nucleótido cada vez, una hebra de ADN complementaria a la hebra cuya secuencia se busca. Los ácidos nucleicos de estudio se pueden inmovilizar a un soporte sólido, hibridar con un cebador de secuenciación, incubar con ADN polimerasa, ATP sulfurilasa, luciferasa, apirasa, adenosina 5' fosfosulfato y luciferina. Las soluciones de nucleótidos se añaden y eliminan secuencialmente. La incorporación correcta de un nucleótido libera un pirofosfato, que interacciona con ATP sulfurilasa y produce ATP en presencia de adenosina 5' fosfosulfato, alimentando la reacción de luciferina, que produce una señal quimioluminiscente que permite la determinación de la secuencia. Máquinas para pirosecuenciación y reactivos específicos de metilación están disponibles de Qiagen, Inc. (Valencia, CA). Véase también Tost y Gut, 2007, *Nat. Prot.* 2 2265-2275. Un ejemplo de un sistema que puede usar un experto en la materia basado en pirosecuenciación en general implica las siguientes etapas: ligar un ácido nucleico adaptador a un ácido nucleico de estudio e hibridar el ácido nucleico de estudio a una bola; amplificar una secuencia de nucleótidos en el ácido nucleico de estudio en una emulsión; separar las bolas usando un soporte sólido multipocillo de picolitros; y secuenciar las secuencias de nucleótidos amplificadas por metodología de pirosecuenciación (por ejemplo, Nakano *et al.,* 2003, *J. Biotech.* 102, 117-124). Tal sistema se puede usar para amplificar exponencialmente productos de amplificación generados por un proceso descrito en el presente documento, por ejemplo, ligando un ácido nucleico heterólogo al primer producto de amplificación generado por un proceso descrito en el presente documento.

Ciertas formas de realización de secuenciación de molécula única se basan en el principio de secuenciación por síntesis, y utilizan transferencia de energía de resonancia de fluorescencia de par único (FRET de par único) como un mecanismo mediante el que se emiten fotones como resultado de incorporación de nucleótido con éxito. Los fotones emitidos con frecuencia se detectan usando dispositivos de carga acoplada enfriados intensificados o de alta sensibilidad junto con microscopia de reflexión interna total (TIRM). Los fotones se emiten solo cuando la solución de reacción introducida contiene el nucleótido correcto para la incorporación en la cadena de ácido nucleico creciente que se sintetiza como resultado del proceso de secuenciación. En la secuenciación o detección de molécula única basada en FRET, la energía se transfiere entre dos colorantes fluorescentes, algunas veces colorantes de polimetina cianina Cy3 y Cy5, a través de interacciones de dipolo de largo alcance. El donante se excita en su longitud de onda de excitación específica y la energía del estado excitado se transfiere, de forma no radioactiva al colorante aceptor,

que a su vez se excita. El colorante aceptor eventualmente vuelve al estado basal por emisión radioactiva de un fotón. Los dos colorantes usados en el proceso de transferencia de energía representan el “par único”, en FRET de par único. Cy3 se usa con frecuencia como el fluoróforo donante y con frecuencia se incorpora como el primer nucleótido marcado. Cy5 con frecuencia se usa como el fluoróforo aceptor y se usa como el marcador de nucleótido para sucesivas adiciones de nucleótidos después de la incorporación de un primer nucleótido marcado con Cy3. Los fluoróforos en general están a 10 nanómetros uno de otro para que la transferencia de energía se produzca con éxito. Bailey y col recientemente describieron un método muy sensible (ADN metilado 15pg) que usa puntos cuánticos para detectar el estado de metilación usando transferencia de energía de resonancia fluorescente (MS-qFRET) (Bailey *et al.* 2009, *Genome Res.* 19(8), 1455-1461).

Un ejemplo de un sistema que se puede usar basado en secuenciación de molécula única en general implica hibridar un cebador a un ácido nucleico de estudio para generar un complejo; asociar el complejo con una fase sólida; extender iterativamente el cebador por un nucleótido etiquetado con una molécula fluorescente; y capturar una imagen de señales de transferencia de energía de fluorescencia después de cada iteración (por ejemplo, Braslaysky *et al.,* PNAS 100(7): 3960-3964 (2003); patente en EE UU No. 7.297.518 (Quake *et al.*)). Tal sistema se puede usar para secuenciar directamente productos de amplificación generados por procesos descritos en el presente documento. En algunas formas de realización, el producto de amplificación lineal liberado se puede hibridar con un cebador que contiene secuencias complementarias a secuencias de captura inmovilizadas presentes en un soporte sólido, una bola o un portaobjetos de vidrio, por ejemplo. La hibridación de los complejos cebador-producto de amplificación lineal liberado con las secuencias de captura inmovilizadas, inmoviliza los productos de amplificación lineales liberados a soportes sólidos para secuenciación basada en FRET de par único por síntesis. El cebador con frecuencia es fluorescente, de modo que se puede generar una imagen de referencia inicial de la superficie del portaobjetos con ácidos nucleicos inmovilizados. La imagen de referencia inicial es útil para determinar localizaciones en las que se produce incorporación de nucleótidos verdaderos. Las señales fluorescentes detectadas en localizaciones de la matriz no inicialmente identificadas en la imagen de referencia de “solo cebador” se descartan como fluorescencia no específica. Después de la inmovilización los complejos cebador-producto de amplificación lineal liberado, los ácidos nucleicos unidos con frecuencia se secuencian en paralelo por las etapas iterativas de a) extensión de polimerasa en presencia de un nucleótido fluorescentemente marcado, b) detección de fluorescencia usando microscopía apropiada, TIRM, por ejemplo, c) eliminación de nucleótido fluorescente, y d) vuelta a la etapa a con un nucleótido fluorescentemente marcado diferente.

En algunas formas de realización, al menos 2 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 3 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 4 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 5 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 6 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 7 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 8 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 9 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 10 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 15 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 20 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 25, al menos 30, al menos 35, al menos 40, al menos 45, al menos 50, al menos 55, al menos 60, al menos 65, al menos 70, al menos 75, al menos 80, al menos 85, al menos 90, al menos 95, al menos 100, al menos 125, al menos 150, al menos 175, al menos 200, al menos 250, al menos 300, al menos 350, o al menos 400 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 500 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 1.000 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 5.000 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 10.000 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 20.000 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 30.000 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 40.000 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 50.000 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 60.000 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 70.000 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 80.000 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 90.000 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 100.000 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, los ID de cebador se unen a las moléculas molde de ácido nucleico simultáneamente. En algunas formas de realización, los ID de cebador se unen a las moléculas molde de ácido nucleico secuencialmente. En algunas formas de realización, las moléculas molde de ácido nucleico se amplifican y/o detectan simultáneamente. En algunas formas de realización, las moléculas molde de ácido nucleico se amplifican y/o detectan secuencialmente. En algunas formas de realización, el ID de cebador comprende una secuencia de ácido nucleico. En algunas formas de realización, el ID de cebador comprende una secuencia de ácido desoxirribonucleico. En algunas formas de realización, el ID de cebador comprende una secuencia de ácido ribonucleico. En algunas formas de realización, el ID de cebador comprende 5-100 nucleótidos. En algunas formas de realización, el ID de cebador comprende 5-50 nucleótidos. En algunas formas de realización, el ID de cebador

comprende al menos 6 nucleótidos. En algunas formas de realización, el ID de cebador comprende al menos 7 nucleótidos. En algunas formas de realización, el ID de cebador comprende al menos 8 nucleótidos. En algunas formas de realización, el ID de cebador comprende al menos 9 nucleótidos. En algunas formas de realización, el ID de cebador comprende al menos 10 nucleótidos. En algunas formas de realización, el ID de cebador comprende al menos 12 nucleótidos. En algunas formas de realización, el ID de cebador comprende al menos 15 nucleótidos. En algunas formas de realización, el ID de cebador comprende al menos 20 nucleótidos. En algunas formas de realización, el ID de cebador comprende al menos 25 nucleótidos. En algunas formas de realización, el ID de cebador comprende al menos 35 nucleótidos.

Se proporciona además un método para detectar variantes genéticas que comprende: (a) proporcionar una muestra que comprende una molécula molde de ácido nucleico; (b) unir un cebador que comprenden un ID de cebador a cada molécula molde de ácido nucleico que se va a analizar para generar un molde de ácido nucleico etiquetado, en donde cada molde de ácido nucleico etiquetado está unido a un ID de cebador único; (c) amplificar el molde de ácido nucleico etiquetado para producir amplicones etiquetados; y (d) detectar los amplicones etiquetados, determinando mediante ello las variantes genéticas. En algunas formas de realización, detectar las variantes genéticas comprende determinar la prevalencia de mutaciones. En algunas formas de realización, detectar las variantes genéticas comprende formar una secuencia consenso para moldes de ácido nucleico etiquetados que comprenden el mismo ID de cebador. En algunas formas de realización, detectar las variantes genéticas comprende secuenciar los amplicones etiquetados. Secuenciar los amplicones etiquetados se puede producir por una variedad de métodos incluyendo, pero no limitados a, método de secuenciación de Maxam-Gilbert, el método de secuenciación del dideoxi de Sanger, método de secuenciación de terminador con colorante, pirosecuenciación, secuenciación de ADN con cebadores múltiples, secuenciación aleatoria, desplazamiento por el cebador. En algunas formas de realización secuenciar comprende pirosecuenciación. En algunas formas de realización, detectar las variantes genéticas comprende contar un número de amplicones etiquetados diferentes. En algunas formas de realización, la variante genética comprende un polimorfismo. En algunas formas de realización, el polimorfismo comprende un polimorfismo de nucleótido único. En algunos casos, el polimorfismo se produce a una frecuencia de menos del 0,5%. En algunos casos, el polimorfismo se produce a una frecuencia de menos del 1%. En algunos casos, el polimorfismo se produce a una frecuencia de menos del 2%. En algunos casos, el polimorfismo se produce a una frecuencia de menos del 5%. En algunos casos, el polimorfismo se produce a una frecuencia de más del 1%. En algunos casos, el polimorfismo se produce a una frecuencia de más del 5%. En algunos casos, el polimorfismo se produce a una frecuencia de más del 10%. En algunos casos, el polimorfismo se produce a una frecuencia de más del 20%. En algunos casos, el polimorfismo se produce a una frecuencia de más del 30%. En algunas formas de realización, la variante genética comprende una mutación. En algunas formas de realización, la variante genética comprende una deleción. En algunas formas de realización, la variante genética comprende una inserción. En algunas formas de realización, el ID de cebador comprende una secuencia degenerada. En algunas formas de realización, el ID de cebador comprende una secuencia semidegenerada. En algunas formas de realización, el ID de cebador comprende una secuencia mixta. En algunas formas de realización, el ID de cebador comprende una secuencia ambigua. En algunas formas de realización, el ID de cebador comprende una secuencia titubeante. En algunas formas de realización, el ID de cebador comprende una secuencia aleatoria. En algunas formas de realización, el ID de cebador comprende una secuencia predeterminada. En algunas formas de realización de ID de cebador está unido al molde por ligación. En algunas formas de realización de ID de cebador está unido al molde por hibridación. En algunas formas de realización de ID de cebador está unido al molde a través de PCR. En algunas formas de realización, se analiza al menos una molécula molde. En algunas formas de realización, se analizan al menos dos moléculas molde diferentes. En algunas formas de realización, detectar los amplicones etiquetados comprende además contar un número de diferentes ID de cebador asociados con los amplicones etiquetados, en donde el número de diferentes ID de cebador asociados con los amplicones etiquetados refleja el número de moldes muestreados. En algunas formas de realización el método comprende además formar una secuencia consenso para amplicones etiquetados que comprenden el mismo ID de cebador. En algunas formas de realización, amplificar comprende un método basado en PCR. En algunas formas de realización, el método basado en PCR comprende PCR. En algunas formas de realización, el método basado en PCR comprende PCR cuantitativa. En algunas formas de realización, el método basado en PCR comprende PCR en emulsión. En algunas formas de realización, el método basado en PCR comprende PCR en gota. En algunas formas de realización, el método basado en PCR comprende PCR de inicio en caliente. En algunas formas de realización, el método basado en PCR comprende PCR in situ. En algunas formas de realización, el método basado en PCR comprende PCR inversa. En algunas formas de realización, el método basado en PCR comprende PCR multiplex. En algunas formas de realización, el método basado en PCR comprende PCR de número variables de repeticiones en tándem (VNTR). En algunas formas de realización, el método basado en PCR comprende PCR asimétrica. En algunas formas de realización, el método basado en PCR comprende PCR larga. En algunas formas de realización, el método basado en PCR comprende PCR anidada. En algunas formas de realización, el método basado en PCR comprende PCR hemianidada. En algunas formas de realización, el método basado en PCR comprende PCR touchdown. En algunas formas de realización, el método basado en PCR comprende PCR de ensamblaje. En algunas formas de realización, el método basado en PCR comprende PCR en colonia. En algunas formas de realización, amplificar comprende un método no basado en PCR. En algunas formas de realización, el método no basado en PCR comprende amplificación por desplazamiento múltiple (MDA). En algunas formas de realización, el método no basado en PCR comprende amplificación mediada por transcripción (TMA). En algunas formas de realización, el método no basado en PCR comprende amplificación basada en secuencia de ácido nucleico (NASBA). En algunas formas de realización, el método no basado en PCR comprende amplificación por desplazamiento de la hebra (SDA). En algunas formas de

realización, el método no basado en PCR comprende SDA en tiempo real. En algunas formas de realización, el método no basado en PCR comprende amplificación por círculo rodante. En algunas formas de realización, el método no basado en PCR comprende amplificación de círculo a círculo. En algunas formas de realización, el molde de ácido nucleico comprende un molde de ADN. En algunas formas de realización, el molde de ácido nucleico comprende un molde de ARN. En algunas formas de realización, al menos 2 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 3 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 4 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 5 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 6 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 7 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 8 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 9 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 10 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 15 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 20 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 25, al menos 30, al menos 35, al menos 40, al menos 45, al menos 50, al menos 55, al menos 60, al menos 65, al menos 70, al menos 75, al menos 80, al menos 85, al menos 90, al menos 95, al menos 100, al menos 125, al menos 150, al menos 175, al menos 200, al menos 250, al menos 300, al menos 350, o al menos 400 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 500 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 1.000 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 5.000 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 10.000 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 20.000 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 30.000 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 40.000 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 50.000 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 60.000 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 70.000 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 80.000 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 90.000 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 100.000 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, los ID de cebador se unen a las moléculas molde de ácido nucleico simultáneamente. En algunas formas de realización, los ID de cebador se unen a las moléculas molde de ácido nucleico secuencialmente. En algunas formas de realización, las moléculas molde de ácido nucleico se amplifican y/o detectan simultáneamente. En algunas formas de realización, las moléculas molde de ácido nucleico se amplifican y/o detectan secuencialmente. En algunas formas de realización, el ID de cebador comprende 5-100 nucleótidos. En algunas formas de realización, el ID de cebador comprende 5-50 nucleótidos. En algunas formas de realización, el ID de cebador comprende al menos 6 nucleótidos. En algunas formas de realización, el ID de cebador comprende al menos 7 nucleótidos. En algunas formas de realización, el ID de cebador comprende al menos 8 nucleótidos. En algunas formas de realización, el ID de cebador comprende al menos 9 nucleótidos. En algunas formas de realización, el ID de cebador comprende al menos 10 nucleótidos. En algunas formas de realización, el ID de cebador comprende al menos 12 nucleótidos. En algunas formas de realización, el ID de cebador comprende al menos 15 nucleótidos. En algunas formas de realización, el ID de cebador comprende al menos 20 nucleótidos. En algunas formas de realización, el ID de cebador comprende al menos 25 nucleótidos. En algunas formas de realización, el ID de cebador comprende al menos 35 nucleótidos.

También se proporciona en el presente documento un método para determinar o cribar variantes resistentes a fármaco que comprende: (a) proporcionar una muestra que comprende una molécula molde de ácido nucleico; (b) unir un cebador que comprenden un ID de cebador a cada molécula molde de ácido nucleico que se va a analizar para generar un molde de ácido nucleico etiquetado, en donde cada molde de ácido nucleico etiquetado está unido a un ID de cebador único; (c) amplificar el molde de ácido nucleico etiquetado para producir amplicones etiquetados; y (d) detectar los amplicones etiquetados, determinando o cribando mediante ello las variantes resistentes a fármaco. En algunas formas de realización, detectar amplicones etiquetados comprende secuenciar los amplicones etiquetados. Secuenciar los amplicones etiquetados se puede producir por una variedad de métodos incluyendo, pero no limitados a, método de secuenciación de Maxam-Gilbert, el método de secuenciación del dideoxi de Sanger, método de secuenciación de terminador con colorante, pirosecuenciación, secuenciación de ADN con cebadores múltiples, secuenciación aleatoria, desplazamiento por el cebador. En algunas formas de realización secuenciar comprende pirosecuenciación. En algunas formas de realización, detectar amplicones etiquetados comprende además formar una secuencia consenso para los amplicones etiquetados que tienen el mismo ID de cebador. En algunas formas de realización, la molécula molde de ácido nucleico comprende una secuencia vírica. En algunas formas de realización, la molécula molde de ácido nucleico comprende una secuencia bacteriana. En algunas formas de realización, la muestra es de un individuo que padece una infección vírica. En algunas formas de realización, la muestra es de un individuo que padece una infección bacteriana. En algunas formas de realización, la muestra es de un individuo que padece cáncer. En algunas formas de realización, la muestra es de un individuo que padece un trastorno autoinmunitario. En algunas formas de realización, el ID de cebador comprende una secuencia degenerada. En algunas formas de realización, el ID de cebador comprende una secuencia semidegenerada. En algunas formas de realización, el ID de cebador comprende una secuencia mixta. En algunas formas de realización, el ID de cebador

comprende una secuencia ambigua. En algunas formas de realización, el ID de cebador comprende una secuencia titubeante. En algunas formas de realización, el ID de cebador comprende una secuencia aleatoria. En algunas formas de realización, el ID de cebador comprende una secuencia predeterminada. En algunas formas de realización de ID de cebador está unido al molde por ligación. En algunas formas de realización de ID de cebador está unido al molde por hibridación. En algunas formas de realización de ID de cebador está unido al molde a través de PCR. En algunas formas de realización, se analiza al menos una molécula molde. En algunas formas de realización, se analizan al menos dos moléculas molde diferentes. En algunas formas de realización, detectar los amplicones etiquetados comprende además contar un número de diferentes ID de cebador asociados con los amplicones etiquetados, en donde el número de diferentes ID de cebador asociados con los amplicones etiquetados refleja el número de moldes muestreados. En algunas formas de realización, amplificar comprende un método basado en PCR. En algunas formas de realización, el método basado en PCR comprende PCR. En algunas formas de realización, el método basado en PCR comprende PCR cuantitativa. En algunas formas de realización, el método basado en PCR comprende PCR en emulsión. En algunas formas de realización, el método basado en PCR comprende PCR en gota. En algunas formas de realización, el método basado en PCR comprende PCR de inicio en caliente. En algunas formas de realización, el método basado en PCR comprende PCR in situ. En algunas formas de realización, el método basado en PCR comprende PCR inversa. En algunas formas de realización, el método basado en PCR comprende PCR multiplex. En algunas formas de realización, el método basado en PCR comprende PCR de número variables de repeticiones en tándem (VNTR). En algunas formas de realización, el método basado en PCR comprende PCR asimétrica. En algunas formas de realización, el método basado en PCR comprende PCR larga. En algunas formas de realización, el método basado en PCR comprende PCR anidada. En algunas formas de realización, el método basado en PCR comprende PCR hemianidada. En algunas formas de realización, el método basado en PCR comprende PCR touchdown. En algunas formas de realización, el método basado en PCR comprende PCR de ensamblaje. En algunas formas de realización, el método basado en PCR comprende PCR en colonia. En algunas formas de realización, amplificar comprende un método no basado en PCR. En algunas formas de realización, el método no basado en PCR comprende amplificación por desplazamiento múltiple (MDA). En algunas formas de realización, el método no basado en PCR comprende amplificación mediada por transcripción (TMA). En algunas formas de realización, el método no basado en PCR comprende amplificación basada en secuencia de ácido nucleico (NASBA). En algunas formas de realización, el método no basado en PCR comprende amplificación por desplazamiento de la hebra (SDA). En algunas formas de realización, el método no basado en PCR comprende SDA en tiempo real. En algunas formas de realización, el método no basado en PCR comprende amplificación por círculo rodante. En algunas formas de realización, el método no basado en PCR comprende amplificación de círculo a círculo. En algunas formas de realización, al menos 2 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 3 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 4 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 5 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 6 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 7 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 8 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 9 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 10 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 15 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 20 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 25, al menos 30, al menos 35, al menos 40, al menos 45, al menos 50, al menos 55, al menos 60, al menos 65, al menos 70, al menos 75, al menos 80, al menos 85, al menos 90, al menos 95, al menos 100, al menos 125, al menos 150, al menos 175, al menos 200, al menos 250, al menos 300, al menos 350, o al menos 400 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 500 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 1.000 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 5.000 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 10.000 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 20.000 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 30.000 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 40.000 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 50.000 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 60.000 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 70.000 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 80.000 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 90.000 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 100.000 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, los ID de cebador se unen a las moléculas molde de ácido nucleico simultáneamente. En algunas formas de realización, los ID de cebador se unen a las moléculas molde de ácido nucleico secuencialmente. En algunas formas de realización, las moléculas molde de ácido nucleico se amplifican y/o detectan simultáneamente. En algunas formas de realización, las moléculas molde de ácido nucleico se amplifican y/o detectan secuencialmente. En algunas formas de realización, el ID de cebador comprende 5-100 nucleótidos. En algunas formas de realización, el ID de cebador comprende 5-50 nucleótidos. En algunas formas de realización, el ID de cebador comprende al menos 6 nucleótidos. En algunas formas de realización, el ID de cebador comprende al menos 7 nucleótidos. En algunas formas de realización, el ID de cebador comprende al menos 8 nucleótidos. En algunas formas de realización, el ID de cebador comprende al menos 9 nucleótidos. En algunas formas de realización, el ID de cebador comprende al menos 10

nucleótidos. En algunas formas de realización, el ID de cebador comprende al menos 12 nucleótidos. En algunas formas de realización, el ID de cebador comprende al menos 15 nucleótidos. En algunas formas de realización, el ID de cebador comprende al menos 20 nucleótidos. En algunas formas de realización, el ID de cebador comprende al menos 25 nucleótidos. En algunas formas de realización, el ID de cebador comprende al menos 35 nucleótidos.

Se divulga además en el presente documento un método para determinar remuestreo de PCR en una reacción de amplificación que comprende: (a) proporcionar una muestra que comprende una molécula molde de ácido nucleico; (b) unir un cebador que comprenden un ID de cebador a cada molécula molde de ácido nucleico que se va a analizar para generar un molde de ácido nucleico etiquetado, en donde cada molde de ácido nucleico etiquetado está unido a un ID de cebador único; (c) amplificar el molde de ácido nucleico etiquetado para producir amplicones etiquetados; y (d) detectar los amplicones etiquetados, determinando mediante ello el remuestreo de PCR en una reacción de amplificación. En algunas formas de realización, el ID de cebador comprende una secuencia degenerada. En algunas formas de realización, el ID de cebador comprende una secuencia semidegenerada. En algunas formas de realización, el ID de cebador comprende una secuencia mixta. En algunas formas de realización, el ID de cebador comprende una secuencia ambigua. En algunas formas de realización, el ID de cebador comprende una secuencia titubeante. En algunas formas de realización, el ID de cebador comprende una secuencia aleatoria. En algunas formas de realización, el ID de cebador comprende una secuencia predeterminada. En algunas formas de realización de ID de cebador está unido al molde por ligación. En algunas formas de realización de ID de cebador está unido al molde por hibridación. En algunas formas de realización de ID de cebador está unido al molde a través de PCR. En algunas formas de realización, se analiza al menos una molécula molde. En algunas formas de realización, se analizan al menos dos moléculas molde diferentes. En algunas formas de realización, detectar amplicones etiquetados comprende secuenciar los amplicones etiquetados. Secuenciar los amplicones etiquetados se puede producir por una variedad de métodos incluyendo, pero no limitados a, método de secuenciación de Maxam-Gilbert, el método de secuenciación del dideoxi de Sanger, método de secuenciación de terminador con colorante, pirosecuenciación, secuenciación de ADN con cebadores múltiples, secuenciación aleatoria, desplazamiento por el cebador. En algunas formas de realización secuenciar comprende pirosecuenciación. En algunas formas de realización, detectar los amplicones etiquetados comprende además contar un número de diferentes ID de cebador asociados con los amplicones etiquetados, en donde el número de diferentes ID de cebador asociados con los amplicones etiquetados refleja el número de moldes muestreados. En algunas formas de realización, el método comprende además formar una secuencia consenso para amplicones etiquetados que comprenden el mismo ID de cebador. En algunas formas de realización, el molde de ácido nucleico comprende un molde de ADN. En algunas formas de realización, el molde de ácido nucleico comprende un molde de ARN. En algunas formas de realización, amplificar comprende un método basado en PCR. En algunas formas de realización, el método basado en PCR comprende PCR. En algunas formas de realización, el método basado en PCR comprende PCR cuantitativa. En algunas formas de realización, el método basado en PCR comprende PCR en emulsión. En algunas formas de realización, el método basado en PCR comprende PCR en gota. En algunas formas de realización, el método basado en PCR comprende PCR de inicio en caliente. En algunas formas de realización, el método basado en PCR comprende PCR in situ. En algunas formas de realización, el método basado en PCR comprende PCR inversa. En algunas formas de realización, el método basado en PCR comprende PCR multiplex. En algunas formas de realización, el método basado en PCR comprende PCR de número variables de repeticiones en tándem (VNTR). En algunas formas de realización, el método basado en PCR comprende PCR asimétrica. En algunas formas de realización, el método basado en PCR comprende PCR larga. En algunas formas de realización, el método basado en PCR comprende PCR anidada. En algunas formas de realización, el método basado en PCR comprende PCR hemianidada. En algunas formas de realización, el método basado en PCR comprende PCR touchdown. En algunas formas de realización, el método basado en PCR comprende PCR de ensamblaje. En algunas formas de realización, el método basado en PCR comprende PCR en colonia. En algunas formas de realización, amplificar comprende un método no basado en PCR. En algunas formas de realización, el método no basado en PCR comprende amplificación por desplazamiento múltiple (MDA). En algunas formas de realización, el método no basado en PCR comprende amplificación mediada por transcripción (TMA). En algunas formas de realización, el método no basado en PCR comprende amplificación basada en secuencia de ácido nucleico (NASBA). En algunas formas de realización, el método no basado en PCR comprende amplificación por desplazamiento de la hebra (SDA). En algunas formas de realización, el método no basado en PCR comprende SDA en tiempo real. En algunas formas de realización, el método no basado en PCR comprende amplificación por círculo rodante. En algunas formas de realización, el método no basado en PCR comprende amplificación de círculo a círculo. En algunas formas de realización, al menos 2 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 3 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 4 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 5 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 6 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 7 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 8 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 9 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 10 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 15 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 20 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 25, al menos 30, al menos 35, al menos 40, al menos 45, al menos 50, al menos 55, al menos 60, al menos 65, al menos 70, al menos 75, al menos 80, al menos 85, al menos 90, al menos 95, al menos 100, al menos 125, al menos 150, al menos 175, al menos 200, al menos 250, al menos 300, al menos 350, o al menos

400 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 500 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 1.000 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 5.000 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 10.000 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 20.000 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 30.000 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 40.000 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 50.000 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 60.000 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 70.000 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 80.000 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 90.000 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 100.000 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, los ID de cebador se unen a las moléculas molde de ácido nucleico simultáneamente. En algunas formas de realización, los ID de cebador se unen a las moléculas molde de ácido nucleico secuencialmente. En algunas formas de realización, las moléculas molde de ácido nucleico se amplifican y/o detectan simultáneamente. En algunas formas de realización, las moléculas molde de ácido nucleico se amplifican y/o detectan secuencialmente. En algunas formas de realización, el ID de cebador comprende 5-100 nucleótidos. En algunas formas de realización, el ID de cebador comprende 5-50 nucleótidos. En algunas formas de realización, el ID de cebador comprende al menos 6 nucleótidos. En algunas formas de realización, el ID de cebador comprende al menos 7 nucleótidos. En algunas formas de realización, el ID de cebador comprende al menos 8 nucleótidos. En algunas formas de realización, el ID de cebador comprende al menos 9 nucleótidos. En algunas formas de realización, el ID de cebador comprende al menos 10 nucleótidos. En algunas formas de realización, el ID de cebador comprende al menos 12 nucleótidos. En algunas formas de realización, el ID de cebador comprende al menos 15 nucleótidos. En algunas formas de realización, el ID de cebador comprende al menos 20 nucleótidos. En algunas formas de realización, el ID de cebador comprende al menos 25 nucleótidos. En algunas formas de realización, el ID de cebador comprende al menos 35 nucleótidos.

Se divulga además en el presente documento un método para determinar errores de PCR y/o errores de secuenciación que comprende: (a) proporcionar una muestra que comprende una molécula molde de ácido nucleico; (b) unir un cebador que comprenden un ID de cebador a cada molécula molde de ácido nucleico que se va a analizar para generar un molde de ácido nucleico etiquetado, en donde cada molde de ácido nucleico etiquetado está unido a un ID de cebador único; (c) amplificar el molde de ácido nucleico etiquetado para producir amplicones etiquetados; y (d) detectar los amplicones etiquetados, determinando mediante ello errores de PCR y/o errores de secuenciación. En algunas formas de realización, determinar el error de PCR y/o error de secuenciación comprende determinar la fidelidad de una polimerasa. En algunas formas de realización, determinar el error de PCR y/o error de secuenciación comprende determinar la precisión de oligonucleótidos sintetizados in vitro. En algunas formas de realización, determinar el error de PCR y/o error de secuenciación comprende determinar la precisión de la reacción de secuenciación. En algunas formas de realización, el ID de cebador comprende una secuencia degenerada. En algunas formas de realización, el ID de cebador comprende una secuencia semidegenerada. En algunas formas de realización, el ID de cebador comprende una secuencia mixta. En algunas formas de realización, el ID de cebador comprende una secuencia ambigua. En algunas formas de realización, el ID de cebador comprende una secuencia titubeante. En algunas formas de realización, el ID de cebador comprende una secuencia aleatoria. En algunas formas de realización, el ID de cebador comprende una secuencia predeterminada. En algunas formas de realización de ID de cebador está unido al molde por ligación. En algunas formas de realización de ID de cebador está unido al molde por hibridación. En algunas formas de realización de ID de cebador está unido al molde a través de PCR. En algunas formas de realización, se analiza al menos una molécula molde. En algunas formas de realización, se analizan al menos dos moléculas molde diferentes. En algunas formas de realización, detectar los amplicones etiquetados comprende además contar un número de diferentes ID de cebador asociados con los amplicones etiquetados, en donde el número de diferentes ID de cebador asociados con los amplicones etiquetados refleja el número de moldes muestreados. En algunas formas de realización, el método comprende además formar una secuencia consenso para amplicones etiquetados que comprenden el mismo ID de cebador. En algunas formas de realización, el molde de ácido nucleico comprende un molde de ADN. En algunas formas de realización, el molde de ácido nucleico comprende un molde de ARN. En algunas formas de realización, amplificar comprende un método basado en PCR. En algunas formas de realización, el método basado en PCR comprende PCR. En algunas formas de realización, el método basado en PCR comprende PCR cuantitativa. En algunas formas de realización, el método basado en PCR comprende PCR en emulsión. En algunas formas de realización, el método basado en PCR comprende PCR en gota. En algunas formas de realización, el método basado en PCR comprende PCR de inicio en caliente. En algunas formas de realización, el método basado en PCR comprende PCR in situ. En algunas formas de realización, el método basado en PCR comprende PCR inversa. En algunas formas de realización, el método basado en PCR comprende PCR multiplex. En algunas formas de realización, el método basado en PCR comprende PCR de número variables de repeticiones en tándem (VNTR). En algunas formas de realización, el método basado en PCR comprende PCR asimétrica. En algunas formas de realización, el método basado en PCR comprende PCR larga. En algunas formas de realización, el método basado en PCR comprende PCR anidada. En algunas formas de realización, el método basado en PCR comprende PCR hemianidada. En algunas formas de realización, el método basado en PCR comprende PCR touchdown. En algunas formas de realización, el método basado en PCR comprende PCR de ensamblaje. En algunas formas de realización,

el método basado en PCR comprende PCR en colonia. En algunas formas de realización, amplificar comprende un método no basado en PCR. En algunas formas de realización, el método no basado en PCR comprende amplificación por desplazamiento múltiple (MDA). En algunas formas de realización, el método no basado en PCR comprende amplificación mediada por transcripción (TMA). En algunas formas de realización, el método no basado en PCR comprende amplificación basada en secuencia de ácido nucleico (NASBA). En algunas formas de realización, el método no basado en PCR comprende amplificación por desplazamiento de la hebra (SDA). En algunas formas de realización, el método no basado en PCR comprende SDA en tiempo real. En algunas formas de realización, el método no basado en PCR comprende amplificación por círculo rodante. En algunas formas de realización, el método no basado en PCR comprende amplificación de círculo a círculo. En algunas formas de realización, al menos 2 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 3 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 4 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 5 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 6 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 7 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 8 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 9 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 10 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 15 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 20 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 25, al menos 30, al menos 35, al menos 40, al menos 45, al menos 50, al menos 55, al menos 60, al menos 65, al menos 70, al menos 75, al menos 80, al menos 85, al menos 90, al menos 95, al menos 100, al menos 125, al menos 150, al menos 175, al menos 200, al menos 250, al menos 300, al menos 350, o al menos 400 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 500 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 1.000 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 5.000 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 10.000 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 20.000 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 30.000 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 40.000 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 50.000 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 60.000 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 70.000 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 80.000 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 90.000 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 100.000 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, los ID de cebador se unen a las moléculas molde de ácido nucleico simultáneamente. En algunas formas de realización, los ID de cebador se unen a las moléculas molde de ácido nucleico secuencialmente. En algunas formas de realización, las moléculas molde de ácido nucleico se amplifican y/o detectan simultáneamente. En algunas formas de realización, las moléculas molde de ácido nucleico se amplifican y/o detectan secuencialmente. En algunas formas de realización, el ID de cebador comprende 5-100 nucleótidos. En algunas formas de realización, el ID de cebador comprende 5-50 nucleótidos. En algunas formas de realización, el ID de cebador comprende al menos 6 nucleótidos. En algunas formas de realización, el ID de cebador comprende al menos 7 nucleótidos. En algunas formas de realización, el ID de cebador comprende al menos 8 nucleótidos. En algunas formas de realización, el ID de cebador comprende al menos 9 nucleótidos. En algunas formas de realización, el ID de cebador comprende al menos 10 nucleótidos. En algunas formas de realización, el ID de cebador comprende al menos 12 nucleótidos. En algunas formas de realización, el ID de cebador comprende al menos 15 nucleótidos. En algunas formas de realización, el ID de cebador comprende al menos 20 nucleótidos. En algunas formas de realización, el ID de cebador comprende al menos 25 nucleótidos. En algunas formas de realización, el ID de cebador comprende al menos 35 nucleótidos.

Se divulga además en el presente documento un método para corregir errores de PCR y/o errores de secuenciación que comprende: (a) proporcionar una muestra que comprende una molécula molde de ácido nucleico; (b) unir un cebador que comprenden un ID de cebador a cada molécula molde de ácido nucleico que se va a analizar para generar un molde de ácido nucleico etiquetado, en donde cada molde de ácido nucleico etiquetado está unido a un ID de cebador único; (c) amplificar el molde de ácido nucleico etiquetado para producir amplicones etiquetados; y (d) detectar los amplicones etiquetados, corrigiendo mediante ello error de PCR y/o error de secuenciación. En algunas formas de realización, el ID de cebador comprende una secuencia degenerada. En algunas formas de realización, el ID de cebador comprende una secuencia semidegenerada. En algunas formas de realización, el ID de cebador comprende una secuencia mixta. En algunas formas de realización, el ID de cebador comprende una secuencia ambigua. En algunas formas de realización, el ID de cebador comprende una secuencia titubeante. En algunas formas de realización, el ID de cebador comprende una secuencia aleatoria. En algunas formas de realización, el ID de cebador comprende una secuencia predeterminada. En algunas formas de realización de ID de cebador está unido al molde por ligación. En algunas formas de realización de ID de cebador está unido al molde por hibridación. En algunas formas de realización de ID de cebador está unido al molde a través de PCR. En algunas formas de realización, se analiza al menos una molécula molde. En algunas formas de realización, se analizan al menos dos moléculas molde diferentes. En algunas formas de realización, detectar los amplicones etiquetados comprende además contar un número de diferentes ID de cebador asociados con los amplicones etiquetados, en donde el número de diferentes ID de cebador

asociados con los amplicones etiquetados refleja el número de moldes muestreados. En algunas formas de realización, el método comprende además formar una secuencia consenso para amplicones etiquetados que comprenden el mismo ID de cebador. En algunas formas de realización, el molde de ácido nucleico comprende un molde de ADN. En algunas formas de realización, el molde de ácido nucleico comprende un molde de ARN. En algunas formas de realización, amplificar comprende un método basado en PCR. En algunas formas de realización, el método basado en PCR comprende PCR. En algunas formas de realización, el método basado en PCR comprende PCR cuantitativa. En algunas formas de realización, el método basado en PCR comprende PCR en emulsión. En algunas formas de realización, el método basado en PCR comprende PCR en gota. En algunas formas de realización, el método basado en PCR comprende PCR de inicio en caliente. En algunas formas de realización, el método basado en PCR comprende PCR in situ. En algunas formas de realización, el método basado en PCR comprende PCR inversa. En algunas formas de realización, el método basado en PCR comprende PCR multiplex. En algunas formas de realización, el método basado en PCR comprende PCR de número variables de repeticiones en tándem (VNTR). En algunas formas de realización, el método basado en PCR comprende PCR asimétrica. En algunas formas de realización, el método basado en PCR comprende PCR larga. En algunas formas de realización, el método basado en PCR comprende PCR anidada. En algunas formas de realización, el método basado en PCR comprende PCR hemianidada. En algunas formas de realización, el método basado en PCR comprende PCR touchdown. En algunas formas de realización, el método basado en PCR comprende PCR de ensamblaje. En algunas formas de realización, el método basado en PCR comprende PCR en colonia. En algunas formas de realización, amplificar comprende un método no basado en PCR. En algunas formas de realización, el método no basado en PCR comprende amplificación por desplazamiento múltiple (MDA). En algunas formas de realización, el método no basado en PCR comprende amplificación mediada por transcripción (TMA). En algunas formas de realización, el método no basado en PCR comprende amplificación basada en secuencia de ácido nucleico (NASBA). En algunas formas de realización, el método no basado en PCR comprende amplificación por desplazamiento de la hebra (SDA). En algunas formas de realización, el método no basado en PCR comprende SDA en tiempo real. En algunas formas de realización, el método no basado en PCR comprende amplificación por círculo rodante. En algunas formas de realización, el método no basado en PCR comprende amplificación de círculo a círculo. En algunas formas de realización, al menos 2 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 3 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 4 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 5 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 6 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 7 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 8 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 9 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 10 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 15 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 20 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 25, al menos 30, al menos 35, al menos 40, al menos 45, al menos 50, al menos 55, al menos 60, al menos 65, al menos 70, al menos 75, al menos 80, al menos 85, al menos 90, al menos 95, al menos 100, al menos 125, al menos 150, al menos 175, al menos 200, al menos 250, al menos 300, al menos 350, o al menos 400 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 500 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 1.000 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 5.000 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 10.000 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 20.000 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 30.000 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 40.000 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 50.000 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 60.000 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 70.000 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 80.000 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 90.000 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, al menos 100.000 moléculas molde de ácido nucleico diferentes se analizan. En algunas formas de realización, los ID de cebador se unen a las moléculas molde de ácido nucleico simultáneamente. En algunas formas de realización, los ID de cebador se unen a las moléculas molde de ácido nucleico secuencialmente. En algunas formas de realización, las moléculas molde de ácido nucleico se amplifican y/o detectan simultáneamente. En algunas formas de realización, las moléculas molde de ácido nucleico se amplifican y/o detectan secuencialmente. En algunas formas de realización, el ID de cebador comprende 5-100 nucleótidos. En algunas formas de realización, el ID de cebador comprende 5-50 nucleótidos. En algunas formas de realización, el ID de cebador comprende al menos 6 nucleótidos. En algunas formas de realización, el ID de cebador comprende al menos 7 nucleótidos. En algunas formas de realización, el ID de cebador comprende al menos 8 nucleótidos. En algunas formas de realización, el ID de cebador comprende al menos 9 nucleótidos. En algunas formas de realización, el ID de cebador comprende al menos 10 nucleótidos. En algunas formas de realización, el ID de cebador comprende al menos 12 nucleótidos. En algunas formas de realización, el ID de cebador comprende al menos 15 nucleótidos. En algunas formas de realización, el ID de cebador comprende al menos 20 nucleótidos. En algunas formas de realización, el ID de cebador comprende al menos 25 nucleótidos. En algunas formas de realización, el ID de cebador comprende al menos 35 nucleótidos.

En el presente documento se divulga un método para analizar secuencias de ácido nucleico, que comprende (a) unir un ID de cebador a un primer extremo de cada uno de una pluralidad de fragmentos de ácido nucleico para formar moldes de ácido nucleico etiquetados; (b) determinar redundantemente la secuencia de nucleótidos de un molde de ácido nucleico etiquetado, en donde las secuencias de nucleótidos determinadas que comparten ID de cebador forman una familia de miembros; y (c) identificar una secuencia de nucleótidos que represente de forma precisa un fragmento de ácido nucleico analito cuando al menos el 1% de miembros de la familia contiene la secuencia.

La secuencia de nucleótidos se puede identificar cuando al menos el 5% de los miembros de la familia contienen la secuencia. La secuencia de nucleótidos se puede identificar cuando al menos el 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, 50%, 55%, 60%, 65%, 70%, 75%, 80%, 85%, 90%, 92%, 95%, 97%, 98%, 99% o más miembros de la familia contienen la secuencia. La secuencia de nucleótidos se puede identificar cuando al menos aproximadamente del 75% a aproximadamente el 99% de miembros de la familia contienen la secuencia. La secuencia de nucleótidos se puede identificar cuando al menos aproximadamente del 85% a aproximadamente el 99% de miembros de la familia contienen la secuencia. La secuencia de nucleótidos se puede identificar cuando al menos aproximadamente del 92% a aproximadamente el 98% de miembros de la familia contienen la secuencia.

Se puede unir un primer sitio cebador universal a un segundo extremo de cada uno de una pluralidad de fragmentos de ácido nucleico analito.

Se pueden realizar al menos dos ciclos de reacción en cadena de la polimerasa de modo que se puede formar una familia de moldes de ácido nucleico etiquetado que tienen un ID de cebador en el primer extremo y un primer sitio cebador universal en un segundo extremo.

El ID de cebador se puede unir covalentemente a un segundo sitio cebador universal. El ID de cebador se puede unir al extremo 5' de un fragmento de ácido nucleico y el segundo sitio cebador universal puede estar 5' respecto al ID de cebador. El ID de cebador se puede unir al extremo 3' de un fragmento de ácido nucleico y el segundo sitio cebador universal puede estar 3' respecto al ID de cebador.

Los fragmentos de ácido nucleico se pueden formar aplicando una fuerza de cizalla a un ácido nucleico. Alternativamente, los fragmentos de ácido nucleico se pueden formar por una o más endonucleasas de restricción.

El método puede además comprender, antes de la etapa de determinar redundantemente, amplificar los moldes de ácido nucleico etiquetado. El método puede además comprender, antes de la etapa de determinar redundantemente, amplificar los moldes de ácido nucleico etiquetado usando un par de cebadores que pueden ser complementarios al primer y segundo sitios cebadores universales, respectivamente.

El método puede además comprender, antes de la etapa de determinar redundantemente, amplificar los moldes de ácido nucleico etiquetado, y en donde antes de dicha amplificación, se puede usar una exonucleasa específica de hebra única para digerir los cebadores en exceso usados para unir el ID de cebador a los fragmentos de ácido nucleico.

El método puede además comprender, antes de la etapa de determinar redundantemente, amplificar los moldes de ácido nucleico etiquetado, y en donde antes de dicha amplificación, la exonucleasa específica de hebra única se puede inactivar, inhibir, o eliminar. La exonucleasa específica de hebra única se puede inactivar por tratamiento con calor.

Los cebadores usados en dicha amplificación pueden comprender una o más modificaciones químicas que los hacen resistentes a exonucleasas. Los cebadores usados en dicha amplificación pueden comprender uno o más enlaces fosforotioato.

El método puede además comprender, antes de la etapa de amplificación, tratar el ADN con bisulfito para convertir bases de citosina no metilada a uracilo.

El método puede además comprender la etapa de comparar el número de familias que representan un primer fragmento de ADN a un número de familias que representan un segundo fragmento de ADN para determinar una concentración relativa de un primer fragmento de ADN respecto a un segundo fragmento de ADN en la pluralidad de fragmentos de ADN.

Se divulga en el presente documento un método para analizar secuencias de ADN que comprende (a) unir un ID de cebador a un primer extremo de cada uno de una pluralidad de fragmentos de ADN usando al menos dos ciclos de amplificación con un primer y un segundo cebadores para formar fragmentos de ADN etiquetado, en donde los ID de cebador están en exceso de los fragmentos de ADN durante la amplificación, en donde el primer cebador comprende (i) un primer segmento complementario a un amplicón deseado; (ii) un segundo segmento que contiene el ID de cebador; y (iii) un tercer segmento que contiene un sitio cebador universal para la posterior amplificación; y en donde el segundo cebador comprende un sitio cebador universal para la posterior amplificación; en donde cada ciclo de amplificación une un sitio cebador universal a la hebra; (b) amplificar los fragmentos de ADN etiquetado para formar una familia de fragmentos de ADN etiquetado de cada fragmento de ADN etiquetado; y (c) determinar las secuencias de nucleótidos de una pluralidad de miembros de la familia.

El método puede además comprender las etapas de (d) comparar secuencias de una familia de fragmentos de ADN etiquetado; y (e) identificar una secuencia de nucleótidos como que representa de forma precisa un fragmento de ADN cuando al menos el 1% de los miembros de la familia contiene la secuencia.

La secuencia de nucleótidos se puede identificar cuando al menos el 5% de los miembros de la familia contienen la secuencia. La secuencia de nucleótidos se puede identificar cuando al menos el 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, 50%, 55%, 60%, 65%, 70%, 75%, 80%, 85%, 90%, 92%, 95%, 97%, 98%, 99% o más miembros de la familia contienen la secuencia. La secuencia de nucleótidos se puede identificar cuando al menos aproximadamente del 75% a aproximadamente el 99% de miembros de la familia contienen la secuencia. La secuencia de nucleótidos se puede identificar cuando al menos aproximadamente del 85% a aproximadamente el 99% de miembros de la familia contienen la secuencia. La secuencia de nucleótidos se puede identificar cuando al menos aproximadamente del 92% a aproximadamente el 98% de miembros de la familia contienen la secuencia.

Los segundos cebadores pueden comprender cada uno de ID de cebador.

Los ID de cebador pueden tener de 2 a 4000 bases o pares de bases inclusive. Los ID de cebador pueden tener de 20 a 100 bases o pares de bases inclusive. Los ID de cebador pueden tener de 20 a 80 bases o pares de bases inclusive. Los ID de cebador pueden tener de 20 a 60 bases o pares de bases inclusive. Los ID de cebador pueden tener al menos aproximadamente 10, 15, 20, 25, 30, 35, 40, 45, 50, 60, 70, 80, 90, 100 o más bases o pares de bases. Los ID de cebador pueden tener al menos aproximadamente 125, 150, 175, 200, 250, 300, 350, 400, 450, 500 o más bases o pares de bases. Los ID de cebador pueden tener menos de aproximadamente 400, 300, 200, 100, 90, 80, 70, 60 o menos bases o pares de bases.

El método puede además comprender, antes de la etapa de amplificar los fragmentos de ADN etiquetado, digerir los cebadores en exceso usados para unir el ID de cebador a los fragmentos de ADN con una exonucleasa específica de hebra única. El método puede además comprender, antes de la etapa de amplificar, inactivar, inhibir o eliminar la exonucleasa específica de hebra única. La exonucleasa específica de hebra única se puede inactivar por tratamiento con calor. Los cebadores usados en la etapa de amplificar pueden comprender uno o más enlaces fosforotioato.

El método puede además comprender, antes de la etapa de amplificación, tratar el ADN con bisulfito para convertir bases de citosina no metilada a uracilo. El método puede además comprender la etapa de comparar el número de familias que representan un primer fragmento de ADN a un número de familias que representan un segundo fragmento de ADN para determinar una concentración relativa de un primer fragmento de ADN respecto a un segundo fragmento de ADN en la pluralidad de fragmentos de ADN.

Se divulga en el presente documento un método para analizar ADN usando identificadores únicos endógenos, que comprende (a) unir oligonucleótidos adaptadores a extremos de fragmentos de ADN de entre 30 a 2000 bases, inclusive, para formar fragmentos adaptados, en donde cada extremo de un fragmento antes de dicha unión es un identificador único endógeno para el fragmento; (b) amplificar los fragmentos adaptados usando cebadores complementarios a los oligonucleótidos adaptadores para formar familias de fragmentos adaptados; (c) determinar la secuencia de nucleótidos de una pluralidad de miembros de una familia; comparando las secuencias de nucleótidos de la pluralidad de miembros de la familia; y (d) identificar una secuencia de nucleótidos como que representa de forma precisa un fragmento de ADN cuando al menos el 1% de los miembros de la familia contiene la secuencia.

El método puede además comprender enriquecer para fragmentos que representan uno o más genes seleccionados por medio de amplificar fragmentos complementarios a los genes seleccionados.

La etapa de unión puede ser anterior a la etapa de enriquecimiento.

Los fragmentos se pueden formar por corte. Los fragmentos se pueden formar por digestión con una o más enzimas de restricción.

La secuencia de nucleótidos se puede identificar cuando al menos el 5% de los miembros de la familia contienen la secuencia. La secuencia de nucleótidos se puede identificar cuando al menos el 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, 50%, 55%, 60%, 65%, 70%, 75%, 80%, 85%, 90%, 92%, 95%, 97%, 98%, 99% o más miembros de la familia contienen la secuencia. La secuencia de nucleótidos se puede identificar cuando al menos aproximadamente del 75% a aproximadamente el 99% de miembros de la familia contienen la secuencia. La secuencia de nucleótidos se puede identificar cuando al menos aproximadamente del 85% a aproximadamente el 99% de miembros de la familia contienen la secuencia. La secuencia de nucleótidos se puede identificar cuando al menos aproximadamente del 92% a aproximadamente el 98% de miembros de la familia contienen la secuencia.

El método puede además comprender, antes de la etapa de amplificación, tratar el ADN con bisulfito para convertir bases de citosina no metilada a uracilo.

El método puede además comprender la etapa de comparar el número de familias que representan un primer fragmento de ADN a un número de familias que representan un segundo fragmento de ADN para determinar una concentración relativa de un primer fragmento de ADN respecto a un segundo fragmento de ADN en la pluralidad de fragmentos de ADN.

Se divulga en el presente documento una población de pares de cebadores, en donde cada par comprende un primer y un segundo cebador para amplificar e identificar un gen o porción de gen, en donde (a) el primer cebador comprende una primera porción de 10-100 nucleótidos complementarios al gen o porción de gen y una segunda porción de 10 a 100 nucleótidos que comprende un sitio para hibridación a un tercer cebador; (b) el segundo cebador comprende una primera porción de 10-100 nucleótidos complementarios al gen o porción de gen y una segunda porción de 10 a 100 nucleótidos que comprende un sitio para hibridación a un cuarto cebador, en donde interpuesta entre la primera porción y la segunda porción del segundo cebador hay una tercera porción que consiste en 2 a 4000 nucleótidos que forman un ID de cebador, en donde los ID de cebador en la población tienen al menos 4 secuencias diferentes, en donde el primer y el segundo cebador son complementarios a hebras opuestas del gen o porción de gen.

La primera porción del primer cebador y/o la primera porción del segundo cebador puede comprender al menos aproximadamente 10, 15, 20, 25, 30 o más nucleótidos complementarios al gen o porción del gen. La primera porción del primer cebador y/o la primera porción del segundo cebador puede comprender menos de aproximadamente 80, 70, 60, 50 o menos nucleótidos complementarios al gen o porción del gen. La primera porción del primer cebador y/o la primera porción del segundo cebador puede comprender entre aproximadamente 10 a aproximadamente 90, entre aproximadamente 10 a aproximadamente 80, entre aproximadamente 10 a aproximadamente 70, entre aproximadamente 10 hasta aproximadamente 60 nucleótidos complementarios al gen o porción del gen.

El primer cebador puede además comprender un ID de cebador.

Los ID de cebadores en la población pueden tener al menos 10, 15, 20, 25, 30, 35, 40, 45, 50, 60, 70, 80, 90, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000 o más secuencias diferentes. Los ID de cebadores en la población pueden tener al menos al menos 2.000; 3.000; 4.000; 5.000; 6.000; 7.000; 8.000; 9.000; 10.000; 20.000; 25.000; 30.000; 35.000; 40.000; 45.000; 50.000; 60.000; 70.000; 80.000; 90.000; 100.000; 200.000; 300.000; 400.000; 500.000; 600.000; 700.000; 800.000; 900.000; 1.000.000 o más secuencias diferentes.

Se divulga además en el presente documento un kit que comprende una población de cebadores de cebador, en donde cada par comprende un primer y un segundo cebador para amplificar e identificar un gen o porción de gen, en donde (a) el primer cebador comprende una primera porción de 10-100 nucleótidos complementarios al gen o porción de gen y una segunda porción de 10 a 100 nucleótidos que comprende un sitio para hibridación a un tercer cebador; (b) el segundo cebador comprende una primera porción de 10-100 nucleótidos complementarios al gen o porción de gen y una segunda porción de 10 a 100 nucleótidos que comprende un sitio para hibridación a un cuarto cebador, en donde interpuesta entre la primera porción y la segunda porción del segundo cebador hay una tercera porción que consiste en 2 a 4000 nucleótidos que forman un ID de cebador, en donde los ID de cebador en la población tienen al menos 4 secuencias diferentes, en donde el primer y el segundo cebador son complementarios a hebras opuestas del gen o porción de gen.

El kit puede comprender además un tercer y cuarto cebadores complementarios a las segundas porciones de cada uno del primer y el segundo cebadores. La primera porción del primer cebador y/o la primera porción del segundo cebador puede comprender al menos aproximadamente 10, 15, 20, 25, 30 o más nucleótidos complementarios al gen o porción del gen. La primera porción del primer cebador y/o la primera porción del segundo cebador puede comprender menos de aproximadamente 80, 70, 60, 50 o menos nucleótidos complementarios al gen o porción del gen. La primera porción del primer cebador y/o la primera porción del segundo cebador puede comprender entre aproximadamente 10 a aproximadamente 90, entre aproximadamente 10 a aproximadamente 80, entre aproximadamente 10 a aproximadamente 70, entre aproximadamente 10 hasta aproximadamente 60 nucleótidos complementarios al gen o porción del gen.

El primer cebador puede además comprender un ID de cebador. Los ID de cebadores en la población pueden tener al menos 10, 15, 20, 25, 30, 35, 40, 45, 50, 60, 70, 80, 90, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000 o más secuencias diferentes. Los ID de cebadores en la población pueden tener al menos al menos 2.000; 3.000; 4.000; 5.000; 6.000; 7.000; 8.000; 9.000; 10.000; 20.000; 25.000; 30.000; 35.000; 40.000; 45.000; 50.000; 60.000; 70.000; 80.000; 90.000; 100.000; 200.000; 300.000; 400.000; 500.000; 600.000; 700.000; 800.000; 900.000; 1.000.000 o más secuencias diferentes.

La divulgación proporciona kits para detectar y/o medir tipos y/o niveles. En una forma de realización no limitante, los kits para llevar a cabo ensayos diagnósticos de la divulgación típicamente incluyen, un medio envase adecuado, (i) una sonda que comprende una secuencia de ácido nucleico (incluyendo ID de cebador y opcionalmente código de barras) que se une específicamente a un polinucleótido de interés; (ii) un marcador para detectar la presencia de la sonda; y (iii) instrucciones para cómo usar y/o interpretar los resultados. El medio de envase de los kits en general incluirá al menos un vial, tubo de ensayo, matraz, botella, jeringa y/u otro envase en el que un ácido nucleico específico para uno de un polinucleótido de interés de la presente divulgación se puede colocar y/o hacer alícuotas

adecuadamente. Donde se proporciona un segundo, tercero y/o componente adicional, el kit también contendrá en general un segundo, tercero y/u otros envases adicionales en el que se puede colocar este componente. Alternativamente, un envase puede contener una mezcla de más de un reactivo de ácido nucleico, cada reactivo se une específicamente a un marcador diferente según la presente divulgación. Los kits de la presente divulgación también incluirán típicamente medios para contener las sondas de ácido nucleico en confinamiento cerrado para venta comercial. Tales envases pueden incluir envases de plástico moldeados por inyección y/o soplado en los que se retienen los viales deseados.

Los kits pueden comprender además controles positivo y negativo, así como instrucciones para el uso de los componentes del kit contenidos en el mismo, según los métodos de la presente divulgación. El kit puede incluir también un prospecto de envase con instrucciones en relación con la indicación aprobada.

El ID de cebador puede comprender secuencias seleccionadas aleatoriamente. El ID de cebador puede comprender secuencias de nucleótidos predefinidas. El ID de cebador puede comprender tanto secuencias seleccionadas aleatoriamente como nucleótidos predefinidos.

El método puede además comprender, antes de la etapa de amplificación, tratar el ADN con bisulfito para convertir las bases de citosina no metilada a uracilo.

El método puede además comprender la etapa de comparar el número de familias que representan un primer fragmento de ADN con el número de familias que representan un segundo fragmento de ADN para determinar una concentración relativa de un primer fragmento de ADN respecto a un segundo fragmento de ADN en la pluralidad de fragmentos de ADN. Véase también, Jabara *et al.* 2011 PNAS 20166-20171.

A lo largo de la especificación la palabra “comprender”, o variaciones tales como “comprende” o “que comprende”, se entenderá que implica la inclusión de un elemento, número entero o etapa expuestos, o grupo de elementos, números enteros o etapas, pero no la exclusión de cualquier otro elemento, número entero o etapa, o grupo de elementos, números enteros o etapas. La presente divulgación puede adecuadamente “comprender”, “consistir en”, o “consistir esencialmente en”, las etapas, elementos y/o reactivos descritos en las reivindicaciones.

Se advierte además que las reivindicaciones se pueden redactar para excluir cualquier elemento opcional. Como tal, esta manifestación se pretende que sirva como base antecedente para uso de tal terminología exclusiva tal como “solamente”, “solo” y similares en relación con la enumeración de elementos de reivindicación, o el uso de una limitación “negativa”.

Los siguientes ejemplos ilustran adicionalmente la divulgación y no se pretende que limiten el ámbito de la divulgación. En particular, se debe entender que esta divulgación no está limitada a formas de realización particulares descritas, ya que tales pueden, por supuesto, variar. También se debe entender que la terminología usada en el presente documento es para el fin de describir formas de realización particulares solo, y no se pretende que sea limitante, ya que el ámbito de la presente invención está limitado solo por las reivindicaciones adjuntas.

Ejemplos

Ejemplo 1. Uso de los ID de cebador para el análisis de una molécula molde vírica

La secuenciación de alto rendimiento permite la adquisición de grandes cantidades de datos de secuencias que pueden abarcar genomas enteros. Con suficientes cantidades de ADN inicial, no se necesita PCR antes de la etapa de preparación de genoteca del protocolo de secuenciación. Las malas asignaciones de secuenciación inherentes en enfoques de secuenciación de alto rendimiento se resuelven usando múltiples lecturas sobre una base determinada.

La secuenciación profunda también puede capturar la diversidad genética de poblaciones víricas, incluyendo poblaciones intrahuésped derivadas de muestras clínicas. Este enfoque ofrece la oportunidad de ver la diversidad y dinámica de población y evolución vírica en detalle sin precedentes. Un lugar donde la presencia de variantes minoritarias es de importancia práctica inmediata es en la detección de variantes resistentes a fármacos. Los métodos de secuenciación masiva estándar típicamente pierden variantes alélicas por debajo del 20% en frecuencia dentro de una población. Ensayos alternativos pueden detectar variantes menos abundantes que confieren resistencia a fármacos, pero requieren selección a priori de sitios y variantes. Por tanto, los enfoques de secuenciación profunda ofrecen la oportunidad de identificar variantes minoritarias asociadas con resistencia de novo con el fin de entender su papel en el fallo de terapia.

Aunque cribar para variantes resistentes a fármacos es una aplicación práctica de la tecnología de secuenciación profunda, esta tecnología también aborda preguntas más amplias de diversidad de secuencia y estructura para una población compleja como VIH-1. Sin embargo, las tasas de errores de secuenciación relativamente altas de estas tecnologías aumentan artificialmente la diversidad genética, que confunde la detección de variación genética natural especialmente cuando se secuencia una población vírica muy heterogénea. Además, el uso de PCR para amplificar la cantidad de material antes de empezar el protocolo de secuenciación añade el potencial de varios artefactos serios:

Primero, la mala incorporación de nucleótidos por la polimerasa durante muchas rondas de amplificación aumenta artificialmente la diversidad de secuencia; segundo, la recombinación artefactual durante la amplificación se produce cuando productos de terminación prematura ceban una ronda posterior de síntesis, lo que puede oscurecer el ligamiento de dos polimorfismos de secuencia; tercero, la amplificación diferencial puede distorsionar las frecuencias alélicas; y cuarto, la amplificación por PCR puede crear una masa significativa de ADN a partir de un pequeño número de moldes iniciales, que oscurece el verdadero muestreo de la población original ya que estos pocos moldes/genomas iniciales se remuestrean en el producto de PCR, lo que crea remuestreo de secuencia más que la observación de genomas independientes. En conjunto, estos sesgos disminuyen artificialmente la diversidad verdadera al tiempo que introducen diversidad artefactual y también distorsionan las frecuencias alélicas, lo que puede llevar a incongruencia entre las poblaciones víricas real y observada. La mayoría de los investigadores usan herramientas estadísticas para intentar controlar los tipos de errores de secuenciación que se asocian con cada plataforma de secuenciación.

Para hacer la secuenciación profunda útil para poblaciones complejas, es necesario superar el remuestreo de PCR, que se confunde por muestreo de la población original, y errores de PCR y secuenciación, que se pueden confundir por diversidad. Como la mala incorporación de nucleótidos es en gran medida aleatoria a través de sitios y el cambio/recombinación de molde es más probable que se produzca en los últimos ciclos de una PCR, estrategias para crear una secuencia masiva o consenso para cada molde muestreado asignarán la base correcta en cada posición. Un enfoque para muestrear poblaciones muy heterólogas, tal como el gen env de VIH-1, es mediante titulación de dilución de punto final del molde antes de PCR anidada, de modo que un único molde está presente en cada amplificación de PCR. Además de enmascarar las malas incorporaciones, la recombinación mediada por PCR produce moldes recombinantes idénticos a la secuencia parental. Aunque muy precisa, esta técnica necesita mucho trabajo y, como el muestreo de la población depende del número de moldes secuenciados, esta metodología no se presta a la identificación de variantes minoritarias o a entender la estructura de una población compleja, ni es fácilmente adaptable a un enfoque de alto rendimiento.

Hemos desarrollado una técnica de alto rendimiento para resolver directamente la diversidad genética de una población vírica. Esta técnica evita el registro de errores de PCR y secuenciación que crean diversidad artificial, y corrige la distorsión alélica artificial y el remuestreo de PCR, revelando los genomas originales en la población. Esto se logra embebiendo un bloque degenerado de nucleótidos en el cebador usado en la primera ronda de síntesis de ADNc. Esto crea una genoteca aleatoria de secuencias en la población de cebadores. Como los cebadores se usan individualmente de esta genoteca, cada molde vírico se copia de modo que el complemento (ADNc) incluye ahora una etiqueta de secuencia única, o ID de cebador. Este ID de cebador se porta a través de todas las manipulaciones posteriores para marcar todas las secuencias que derivan de cada suceso de molde independiente, y el remuestreo de PCR se convierte entonces en cobertura excesiva para cada molde para crear una secuencia consenso de ese molde. Usando este enfoque, se pudo directamente eliminar error, corregir el remuestreo de PCR, y capturar la fluctuación de variantes minoritarias en la población vírica en un huésped. También se resolvieron variantes resistentes a fármacos minoritarias por debajo del 1% en frecuencia antes de la iniciación de terapia antirretroviral, y se pudo correlacionar estas variantes con la aparición de resistencia a fármacos.

Resultados

Se puede usar un cebador de síntesis de ADNc que contiene un ID de cebador para seguir moldes víricos individuales. Se diseñó una población de cebadores de síntesis de ADNc para cebar síntesis de ADN después del (pro) gen de proteasa de VIH-1, con el cebador que contiene dos bloques adicionales de información identificadora (Fig. 1A). El primer bloque era una cadena de ocho nucleótidos degenerados que creó 65.536 combinaciones de secuencia distintas, o ID de cebador. Esta región estaba flanqueada por un código de barras de tres nucleótidos seleccionado a priori, que crea un bloque de identificación de muestra de modo que múltiples muestras se pudieran juntar en una carrera de secuenciación. Una secuencia diseñada en el extremo 5' del cebador de ADNc se usó para posterior amplificación de las secuencias del ADNc por PCR anidada.

Se extrajo el ARN vírico de muestras de plasma sanguíneo longitudinales de un individuo infectado con el subtipo B de VIH-1 que participaba en un ensayo de eficacia de un inhibidor de la proteasa (M94-247) (Fig. 4). Se usaron aproximadamente 10.000 copias de ARN vírico de cada muestra en una reacción de transcripción inversa para la síntesis de ADNc y etiquetado usando el ID de cebador. El producto de ADNc se separó de los cebadores de ADNc no usados, y después las secuencias víricas se amplificaron por PCR anidada y se secuenciaron en el 454 GS FLX Titanium. Nuestros datos se destilaron de lecturas totales de 20.419, 24.658 y 27.075 para los tres puntos de tiempo (T1, T2 y T3, respectivamente). Las lecturas de secuencia en bruto se evaluaron para el cebador de etiquetado de ADNc y una secuencia pro gen de longitud completa (297 nucleótidos de longitud que representan 99 codones), y cuando tres o más secuencias en una muestra contenían un ID de cebador idéntico, se formó una secuencia consenso para representar una secuencia/genoma en la población (Fig. 1B y 1C y Fig. 5).

Con estas manipulaciones se generaron 857, 1.609 y 2.213 secuencias consenso, respectivamente, para los tres puntos de tiempo (Fig. 1C). El número mediana de lecturas por ID de cebador fue 6, que variaba de 1 a 96 (Fig. 6A). La distribución de ID de cebador idénticos no formó una distribución normal como se esperaría si todos los moldes se amplificaran por igual. Vimos un número mayor del esperado de lecturas únicas de los ID de cebador; aunque no sabemos la razón para esto, tal resultado es consistente con diferentes moldes de ADNc que entran en la PCR en

diferentes ciclos. Puesto que cada molde está etiquetado individualmente el número diferente de lecturas es una indicación de distorsión alélica, como se indica esto puede ser casi 100 veces, En un análisis de un número de variantes de baja abundancia se vio un intervalo de 20 veces de representación mediante distorsión alélica, con la mitad de las variantes hasta 2 a 3 veces más abundantes que la media, y la otra mitad hasta 5 a 10 veces menos abundante (Fig. 7).

Estimamos conservadoramente que la tasa de errores in vitro combinada de la etapa de síntesis de ADNc por transcriptasa inversa (RT) y la síntesis de la primera hebra por la Taq polimerasa está en el orden de 1 mutación en 10.000 bases, o aproximadamente una mutación por 33 secuencias pro gen, basado en una tasa de error de RT de 1 en 22.000 nucleótidos (38) y una tasa de error de Taq polimerasa de 1,1 en 10.000 nucleótidos (39), pero reducido a la mitad porque solo la primera ronda de síntesis es relevante y una mala incorporación en esta etapa da una mezcla. Las rondas posteriores de errores de Taq polimerasa se deben perder en su mayor parte mediante la creación de la secuencia consenso. Por tanto, se esperarían que estuvieran presentes 139 malas incorporaciones de secuencia en el conjunto de datos de 4.679 secuencias totales que representan T1+T2+T3, y con un exceso de transiciones. Se esperaría que estas se produjeran como 113 polimorfismos de nucleótido único (SNP) de copia única y 13 SNP que aparecían dos veces. Observamos 98 SNP de copia única en el conjunto de datos con un exceso de tres veces de transiciones, y con tres cuartos de ellos que son cambios codificantes, lo que es consistente con mutaciones aleatorias. Esperamos que haya SNP de baja frecuencia en la población vírica de variantes raras, pero persistentes que se muestran de forma fortuita, y de la tasa de error intrínseca de replicación vírica (la tasa de error durante una ronda de replicación vírica representaría aproximadamente una mutación por 150 secuencias pro gen). Sin embargo, no podemos distinguir polimorfismos reales de la tasa de error de fondo inferida asociada con la primera y segunda rondas de síntesis de ADN in vitro. Por tanto, hemos limitado el análisis de diversidad de población a los SNP que aparecían al menos dos veces en el conjunto de datos (por ejemplo, ligados a al menos dos ID de cebador separados), ya sea en el mismo punto de tiempo o en múltiples puntos de tiempo en el conjunto de datos global (Tabla 1). No hemos corregido el conjunto de datos para los presuntos 13 SNP que aparecían dos veces que se espera que estén presentes debido a error incluso aunque esto representa el 33% de todos los SNP que aparecían dos veces (13 de 39). En conjunto, el 80% de los SNP (por ejemplo, cualquier cambio de secuencia del consenso que aparecía al menos una vez) en el conjunto de datos total de 72.160 lecturas de secuencia se eliminaron como error. Además, el 60-65% de las lecturas de secuencia se revelaron como remuestreo. Por último, se corrigió la distorsión alélica de hasta casi 100 veces (Fig. 7).

La secuenciación longitudinal del (pro) gen de proteasa de VIH-1 en un individuo sin tratar revela cambios dinámicos en la variación genética. Analizamos las secuencias de las poblaciones de pro genes para evaluar la frecuencia alélica en dos puntos de tiempo muestreados, separados por 6 meses y antes de la selección de fármaco ritonavir (37) (Fig. 4). La población de secuencias combinadas de los dos puntos de tiempo (T1 y T2) antes de la terapia consistía en 492 secuencias pro gen únicas con 155 SNP. Aproximadamente el 4% (por ejemplo, 21) de estas secuencias de gen únicas tenían por encima del 0,5% de abundancia, y estas 21 secuencias de gen únicas representaban el 67% de todos los genomas muestreados, el genoma representa la secuencia consenso global que comprende el 21% de la población total (Fig. 8A y 8B). El número relativamente pequeño de secuencias de gen únicas por encima del 0,5% de frecuencia en la población contenía solo el 7% de los 155 SNP detectados. Por tanto, una gran proporción de la diversidad de la población vírica estaba asociada con un gran número de secuencias pro gen que estaban presentes a baja abundancia (Fig. 8A y 8C); por el contrario, la mayoría de la población consistía en un pequeño número de SNP. De forma similar, la estadística D de Tajima para T1 y T2 en este individuo fueron -2,35 y -2,31, respectivamente (Tabla 2), indicativo de una estructura de población que tiene un exceso de polimorfismos de baja frecuencia. Este patrón es consistente con, pero más extremo que, el observado en un estudio intrahuésped superficial anterior en el que se propuso un modelo de metapoblación para explicar el patrón de la estadística D de Tajima (40). La figura 2A-2B muestra la variabilidad de aminoácidos codificados y variabilidad de nucleótidos sinónimos presente en dos o más genomas individuales a través de los 99 codones en el pro gen para estas muestras.

Variabilidad sinónima. Había 57 codones (con 63 variantes/SNP) que contenían diversidad sinónima que aparecía en ambos puntos preterapia, y 30 codones (con 31 variaciones) que aparecían solo en un punto de tiempo. En conjunto, 75 de los 99 codones contenían algún nivel de diversidad sinónima (figura 2A-2B y Tabla 1). De las 63 variantes que estaban presentes en ambos puntos de tiempo sin tratar, el 92% eran transiciones. De las 31 variantes que aparecían solo en uno de los puntos de tiempo, el 71% eran transiciones, que representan una fracción significativamente menor de transiciones que entre las variantes sinónimas que aparecían en ambos puntos de tiempo (P = 0,012; prueba exacta de Fisher). Esto sugiere que las transversiones sinónimas se seleccionan en contra a lo largo del tiempo.

Variabilidad no sinónima. Había 26 codones (28 variantes) que contenían variabilidad codificante que aparecían en ambos puntos de tiempo preterapia y 28 codones adicionales (33 variantes) con cambios no sinónimos encontrados solo en uno de los puntos de tiempo. En conjunto, 49 de los 99 codones contenían algún nivel de diversidad no sinónima (figura 2A-2B y Tabla 1). Para las 28 variantes no sinónimas detectadas en ambos puntos de tiempo, 22 eran transiciones, y estas representaban en su mayor parte cambios de aminoácidos conservadores. En el caso de mutaciones sinónimas dos tercios de las variantes estaban presentes en ambos puntos de tiempo, mientras que en el caso de mutaciones no sinónimas, menos de la mitad estaban presentes en ambos puntos de tiempo (P = 0,012; prueba exacta de Fisher). Esta observación sugiere que, a este nivel de muestreo de secuencia, somos capaces de ver una diferencia en estabilidad dentro de la población al comparar sustituciones sinónimas y no sinónimas.

Fluctuación genética. Comparamos la estabilidad de SNP minoritarios presentes tanto en T1 como en T2. Un total de 14 de los 91 SNP (sinónimos y no sinónimos que aparecían en ambos puntos de tiempo) tuvo cambios significativos en abundancia entre los dos puntos de tiempo (prueba de χ^2 con una tasa de descubrimiento falsa de 0,05). De los 14 SNP con cambios significativos en abundancia, 11 tenían una disminución en la abundancia, con una disminución media de aproximadamente 7,5 veces. Había tres SNP que tenían un aumento significativo en abundancia, todos los cuales eran sinónimos, que variaba de 4 a 47 veces de aumento. Aunque la mayoría los SNP que cambiaban en abundancia tenían una disminución en la frecuencia entre T1 y T2, a un nivel de población, no había un gran cambio en la diversidad entre los dos puntos de tiempo (T1 π = 0,0080, T2 π = 0,0079; Tabla 2). Sin embargo, la tendencia de abundancia aumentada en los tres sitios puede estar dirigida por selección de epítopos crípticos en un marco de lectura alternativo (véase, Discusión).

Significación de variantes raras. Observamos dos extremos en términos de relevancia biológica en la población sin tratar entre variantes detectadas como al menos dos secuencias independientes a través de los tres puntos de tiempo. En un extremo estaba la detección de genomas no viables en la forma de una variante codificante en la posición 25, que muta el sitio activo de la proteasa, y la detección de codones de terminación en las posiciones 42 y 61 (Tabla 1). En el otro extremo estaba la detección de las variantes L90M y V82A (en los puntos de tiempo 1 y 2, respectivamente) que se convirtieron en las poblaciones de resistencia principales después de que se iniciara la terapia de ritonavir (véase posteriormente; Fig. 3); además, se detectaron V82I y V82L en T2. Encontramos dos ejemplos más de mutaciones de resistencia primarias en baja abundancia, K20R en los tres puntos de tiempo y M46I en dos puntos de tiempo, pero estas no crecieron en presencia de ritonavir (Fig. 3 y Tabla 1). De forma similar, también se detectaron mutaciones compensatorias de adecuación a baja abundancia (L10F, M36I, L63P, A71T, y V77I), todas por debajo del 1% y solo L63P aumentó (modestamente) en abundancia después de la exposición a ritonavir. Más en general, de las 28 sustituciones más estrechamente asociadas con resistencia a fármaco inhibidor de proteasa, encontramos 10 tales variantes, la mitad de las cuales se detectaron en ambos puntos de tiempo de preterapia (Tabla 1).

Evaluación de desequilibrio de ligamiento (LD) en poblaciones de pro gen de VIH-1. Medimos el LD para las secuencias en las poblaciones T1 y T2. Identificamos muy pocos ejemplos de LD en estos dos puntos de tiempo usando la prueba exacta de Fisher con una corrección de Bonferroni. De los 103 sitios polimórficos en T1, solo tres pares estaban en LD significativo. De forma similar, en T2 con 118 sitios polimórficos, solo y antes de selección de fármaco de ritonavir (37) (Fig. 4). La población de secuencia combinada de los dos puntos de tiempo (T1 y T2) antes de la terapia consistía en 492 secuencia pro gen únicas con 155 SNP. Aproximadamente el 4% (por ejemplo 21) de estas secuencias de gen únicas estaban por encima del 0,5% de abundancia, y estas 21 secuencias de gen únicas representaban el 67% de todos los genomas muestreados, representando el genoma la secuencia consenso global que comprende el 21% de la población total (Fig. 8A y 8B). El número relativamente pequeño de secuencias de gen únicas por encima del 0,5% de frecuencia en la población contenía solo el 7% de los 155 SNP detectados. Por tanto, una gran proporción de la diversidad de la población vírica estaba asociada con un gran número de secuencias pro gen que estaban presentes a baja abundancia (Fig. 8A y 8C); por el contrario, la mayoría de la población consistía en un pequeño número de SNP. De forma similar la estadística D de Tajima para T1 y T2 en este individuo fueron -2,35 y -2,31, respectivamente (Tabla 2), indicativo de una estructura de población que tiene un exceso de polimorfismos de baja frecuencia. Este patrón es consistente con, pero más extremo que, el observado en un estudio intrahuésped superficial anterior en el que se propuso un modelo de metapoblación para explicar el patrón de la estadística D de Tajima (40). La figura 2A-2B muestra la variabilidad de aminoácidos codificados y variabilidad de nucleótidos sinónimos presente en dos o más genomas individuales a través de los 99 codones en el pro gen para estas muestras.

Variabilidad sinónima. Había 57 codones (con 63 variantes/SNP) que contenían diversidad sinónima que aparecía en ambos puntos preterapia, y 30 codones (con 31 variaciones) que aparecían solo en un punto de tiempo. En conjunto, 75 de los 99 codones contenían algún nivel de diversidad sinónima (Fig. 2A-2B y Tabla 1). De las 63 variantes que estaban presentes en ambos puntos de tiempo sin tratar, el 92% eran transiciones. De las 31 variantes que aparecían solo en uno de los puntos de tiempo, el 71% eran transiciones, que representan una fracción significativamente menor de transiciones que entre las variantes sinónimas que aparecían en ambos puntos de tiempo (P = 0,012; prueba exacta de Fisher). Esto sugiere que las transversiones sinónimas se seleccionan en contra a lo largo del tiempo.

Variabilidad no sinónima. Había 26 codones (28 variantes) que contenían variabilidad codificante que aparecían en ambos puntos de tiempo preterapia y 28 codones adicionales (33 variantes) con cambios no sinónimos encontrados solo en uno de los puntos de tiempo. En conjunto, 49 de los 99 codones contenían algún nivel de diversidad no sinónima (Fig. 2A-2B y Tabla 1). Para las 28 variantes no sinónimas detectadas en ambos puntos de tiempo, 22 eran transiciones, y estas representaban en su mayor parte cambios de aminoácidos conservadores. En el caso de mutaciones sinónimas dos tercios de las variantes estaban presentes en ambos puntos de tiempo, mientras que en el caso de mutaciones no sinónimas, menos de la mitad estaban presentes en ambos puntos de tiempo (P = 0,012; prueba exacta de Fisher). Esta observación sugiere que, a este nivel de muestreo de secuencia, somos capaces de ver una diferencia en estabilidad dentro de la población al comparar sustituciones sinónimas y no sinónimas.

Fluctuación genética. Comparamos la estabilidad de SNP minoritarios presentes tanto en T1 como en T2. Un total de 14 de los 91 SNP (sinónimos y no sinónimos que aparecían en ambos puntos de tiempo) tuvo cambios significativos en abundancia entre los dos puntos de tiempo (prueba de χ^2 con una tasa de descubrimiento falsa de 0,05). De los 14

SNP con cambios significativos en abundancia, 11 tenían una disminución en la abundancia, con una disminución media de aproximadamente 7,5 veces. Había tres SNP que tenían un aumento significativo en abundancia, todos los cuales eran sinónimos, que variaba de 4 a 47 veces de aumento. Aunque la mayoría los SNP que cambiaban en abundancia tenían una disminución en la frecuencia entre T1 y T2, a un nivel de población, no había un gran cambio en la diversidad entre los dos puntos de tiempo (T1 π = 0,0080, T2 π = 0,0079; Tabla 2). Sin embargo, la tendencia de abundancia aumentada en los tres sitios puede estar dirigida por selección de epítopos crípticos en un marco de lectura alternativo.

Significado de variantes raras. Observamos dos extremos en términos de relevancia biológica en la población sin tratar entre variantes detectadas como al menos dos secuencias independientes a través de los tres puntos de tiempo. En un extremo estaba la detección de genomas no viables en la forma de una variante codificante en la posición 25, que muta el sitio activo de la proteasa, y la detección de codones de terminación en las posiciones 42 y 61 (Tabla 1). En el otro extremo estaba la detección de las variantes L90M y V82A (en los puntos de tiempo 1 y 2, respectivamente) que se convirtieron en las poblaciones de resistencia principales después de que se iniciara la terapia de ritonavir (véase posteriormente; Fig. 3); además, se detectaron V82I y V82L en T2. Encontramos dos ejemplos más de mutaciones de resistencia primarias en baja abundancia, K20R en los tres puntos de tiempo y M46I en dos puntos de tiempo, pero estas no crecieron en presencia de ritonavir (Fig. 3 y Tabla 1). De forma similar, también se detectaron mutaciones compensatorias de adecuación a baja abundancia (L10F, M36I, L63P, A71T, y V77I), todas por debajo del 1% y solo L63P aumentó (modestamente) en abundancia después de la exposición a ritonavir. Más en general, de las 28 sustituciones más estrechamente asociadas con resistencia a fármaco inhibidor de proteasa, encontramos 10 tales variantes, la mitad de las cuales se detectaron en ambos puntos de tiempo de preterapia (Tabla 1).

Evaluación de desequilibrio de ligamiento (LD) en poblaciones de pro gen de VIH-1. Medimos el LD para las secuencias en las poblaciones T1 y T2. Identificamos muy pocos ejemplos de LD en estos dos puntos de tiempo usando la prueba exacta de Fisher con una corrección de Bonferroni. De los 103 sitios polimórficos en T1, solo tres pares estaban en LD significativo. De forma similar, en T2 con 118 sitios polimórficos, solo el número de genomas patógenos en la muestra es limitado, y el uso de PCR puede oscurecer la calidad del muestreo al crear una gran cantidad de ADN de un número relativamente pequeño de moldes iniciales. Esto puede crear homogeneidad artificial, inflar estimaciones de variación genética segregante, distorsionar la distribución de alelos en la población, e introducir diversidad artificial.

Hemos desarrollado una estrategia que permite que cada molde muestreado se etiquete con un ID único mediante un cebador que tiene una etiqueta de secuencia degenerada incorporada durante la síntesis de oligonucleótidos cebadores (Fig. 10). Esta etiqueta se puede seguir después mediante la PCR y el protocolo de secuenciación profunda para identificar cobertura excesiva de secuenciación (remuestreo) de los moldes víricos individuales. Puesto que el ID de cebador permite la identificación de cobertura excesiva, esto se puede usar entonces para crear una secuencia consenso para cada molde, evitando tanto errores relacionados con PCR como errores de secuenciación (Fig. 11). Además, el número de los ID de cebador diferentes refleja el número de moldes que realmente se muestrearon. Esto permite una evaluación realista de la profundidad del muestreo de la población y hace posible aplicar un análisis más riguroso de variantes minoritarias al corregir la distorsión alélica durante la PCR.

Ensayamos el enfoque de ID de cebador secuenciando el dominio codificante de la proteasa de VIH-1 en tres puntos de tiempo en un sujeto que se expuso de forma intermitente a un inhibidor de proteasa entre el segundo y tercer puntos de tiempo. Una característica clave de nuestro enfoque es la eliminación de errores fortuitos y que representan el remuestreo, que produce una reestructuración drástica del conjunto de datos original de 72.162 lecturas. Se han desarrollado otros enfoques que se basan en modelado estadístico para tratar el problema de altas tasas de errores de secuenciación asociadas con tecnologías de secuenciación profunda. El uso del ID de cebador para crear secuencias consenso produjo la eliminación del 80% de los polimorfismos de secuencia única (definido como un cambio en el consenso sin considerar la frecuencia de aparición) en el conjunto de datos. Similarmente, la distorsión alélica era drástica entre las secuencias muestreadas, en la mayoría de los casos variaba de 2 a 15 veces, pero subiendo hasta casi 100 veces. Aunque el ID de cebador revela tal distorsión y ayuda a corregirla, esta es claramente una característica mal controlada de las amplificaciones de PCR que puede afectar drásticamente la abundancia observada de poblaciones complejas, especialmente las variantes minoritarias. La distorsión alélica aún puede persistir si el cebador de ADNc o el cebador de PCR anterior se une diferencialmente entre los moldes, o si los ADNc entran la amplificación de PCR en rondas posteriores y se descartan porque no producen al menos tres lecturas para permitir que se forme una secuencia consenso. Además, los errores de mala incorporación residuales de RT y en la primera ronda de la síntesis por PCR aun limitan la interpretación de mutaciones que se producen en el intervalo del 0,01-0,1%. Este problema no se supera con mayores números de secuencias. Dada la baja diversidad en estas muestras, eliminamos todas las sustituciones que aparecían una vez porque su número se aproximaba al número esperado de errores de secuencia residuales, y esto produjo una sensibilidad de detección en el intervalo del 0,1% para los SNP que aparecían por encima de la frecuencia de la tasa de errores de secuencia residuales.

Al usar el enfoque de ID de cebador, pudimos describir un número de características de la población de secuencia de la proteasa. Primero, un análisis conjunto de dos puntos de tiempo separados seis meses mostró que las variantes presentes en más del 0,5% en abundancia constituían hasta dos tercios de la población total, pero representaban solo el 4% de secuencias de genoma únicas y contenían solo el 7% de los polimorfismos de secuencia única totales.

Aproximadamente el 60% de la diversidad era estable sobre ambos puntos de tiempo, con SNP sinónimos mantenidos en una proporción significativamente mayor en los dos puntos de tiempo que los SNP no sinónimos. Solo el 18% de la diversidad total representaba SNP no sinónimos que estaban presentes en ambos puntos de tiempo. Sin embargo, nuestra capacidad para evaluar la persistencia de estas secuencias está limitada por la profundidad de muestreo, aunque creemos que estamos llegando al límite práctico de muestreo con esta tecnología. Observamos sustituciones no viables y estimamos que la mayoría de los SNP que aparecían una vez eran el resultado de error de método restante. No encontramos patrón de ligamiento conservado entre estos SNP, consistente con los altos niveles de recombinación a través de la población.

Aunque la medida global de la diversidad (π) era similar entre los primeros dos puntos de tiempo, advertimos que los mayores cambios en la abundancia de SNP entre los dos puntos de tiempo estaban en tres posiciones de codones sinónimos (L24L, K70K y G73G). Estos aumentos dinámicos hacían estos SNP parte de un grupo mayor de SNP que representaban el 51% de las secuencias totales que eran de otra manera idénticas a la secuencia consenso (Q18Q; L19I, L24L, K70K, G73G, y Q81Q/L19I/L24L'). Estos SNP también solapaban con los SNP principales que definían subgrupos de las variantes resistentes (L19I; L19V; G16G/L19V). Consideramos la posibilidad de que hubiera una característica unificadora de estos SNP. Encontramos tal característica en que todos estos SNP, tanto codificantes como no codificantes, producen cambios en dos ORF alternativas relativamente grandes que están en los extremos 5' y 3' del pro gen. Se han sugerido marcos de lectura alternativos para generar epítopos de LTC crípticos. En este escenario, estos SNP abundantes representarían varios mutantes de escape. Tales presiones selectivas podrían explicar el comportamiento dinámico de varios de estos SNP entre los dos primeros puntos de tiempo.

Después de exposición intermitente al inhibidor de proteasa ritonavir, pudimos identificar seis linajes independientes de mutaciones resistentes a fármaco. Con la exposición intermitente en este sujeto particular, fue posible ver el linaje principal V82A con la mayor frecuencia vista con resistencia a ritonavir, pero también poblaciones significativas de I84V y L90M. También vimos poblaciones minoritarias de V82I, V82L, y V82F. Esta población mezclada de linajes resistentes probablemente representa las fases tempranas de la evolución de resistencia, una conclusión apoyada por la aparición minoritaria de la mutación compensatoria L63P y la ausencia completa de I54V, que es una mutación compensatoria vista con frecuencia para V82A. Vimos pocos ejemplos de genomas con múltiples mutaciones de resistencia, aunque estas se esperarían después de selección más extensa. Nosotros y otros hemos examinado previamente secuencias víricas que se han recogido en grandes bases de datos. Típicamente, estas secuencias representan la secuencia predominante única en un individuo, y el uso de estas secuencias permite la evaluación de diversidad interpersonal. En el futuro, será un ejercicio interesante comparar las conclusiones alcanzadas al examinar diversidad vírica en una persona con diversidad vírica entre personas; sin embargo, se necesita medir más diversidad intrapersonal a este nivel de detalle para permitir la comparación de diversidad inter- frente a intrapersonal.

La presencia de variantes resistentes a fármaco preexistentes y su papel en el fracaso de la terapia es de gran interés, y el muestreo profundo, preciso de una población vírica puede añadir significativamente a nuestro entendimiento de esta cuestión. Pudimos detectar varios ejemplos de mutaciones resistentes a fármaco, pero a un nivel muy bajo. Nuestra capacidad para detectar de forma fiable estas mutaciones está limitada a las que aparecen a una frecuencia del 0,1-0,2%, limitada en parte por la baja diversidad global en la población. Pudimos ver ejemplos de mutaciones que típicamente se ven solo en presencia de selección con fármaco. Sin embargo, la detección era habitualmente como un genoma en dos puntos de tiempo o dos genomas en un punto de tiempo. Este también era el nivel de detección de mutaciones de sitio activo en la proteasa y de codones de terminación, que deben representar o bien genomas víricos transitorios o errores de mala incorporación residuales. En dos casos, pudimos observar la mutación de resistencia (V82A y L90M) en puntos de tiempo preterapia ligados a los mismos polimorfismos que estaban presentes en la variante que creció durante la exposición al fármaco. Por tanto, aunque es probable que estemos detectando variantes resistentes a fármaco preexistentes relevantes, estas están en el límite de detección y, si se mantienen a un nivel estacionario, está bien por debajo del 0,5% de abundancia.

La mayoría de los protocolos de tecnologías de secuenciación de alto rendimiento aun requiere una cantidad inicial de ADN que necesita una etapa de PCR por adelantado para muchas aplicaciones. El uso de un ID de cebador ayudará a clarificar los productos de secuenciación en cualquier estrategia que use una etapa de PCR inicial con su tasa relacionada de error, recombinación, y remuestreo. En un esfuerzo independiente Kinde y col. han descrito un enfoque análogo en otra secuenciación profunda de moldes individuales antes de PCR y posterior análisis de las secuencias será esencial para entender la verdadera complejidad y diversidad de poblaciones genéticamente dinámicas.

Materiales y Métodos

El ARN vírico se aisló de plasma sanguíneo usando el kit de ARN vírico QIAmp (Qiagen). El ADNc se generó usando transcriptasa inversa SuperScript III (Invitrogen) usando el cebador (con ID de cebador) como se ha descrito. Después de la reacción, el ARN en el híbrido se eliminó por tratamiento con RNasa H (Invitrogen). El cebador de ADNc no incorporado se eliminó, y el producto de ADNc se amplificó por PCR. La secuenciación se hizo usando la plataforma 454 (Roche).

Extracción de ARNv y síntesis de ADNc. El ARN vírico se extrajo de tres muestras de plasma tomadas longitudinalmente de un individuo infectado con el subtipo B de VIH-1 que participaba en un ensayo de eficacia de inhibidor de proteasa (M94-247). Dos muestras se recogieron a ~6 meses antes e inmediatamente antes de la adición del inhibidor de proteasa ritonavir a una pauta de terapia fracasada (cargas víricas en plasma de 285.360 copias de ARN vírico/ml y 321.100 copias de ARN vírico/ml, respectivamente), y una muestra se recogió durante la terapia con ritonavir (aproximadamente 2 meses en terapia, 349.920 copias de ARN vírico/ml), pero durante un tiempo de cumplimiento intermitente aparente. Para cada muestra de plasma, el ARNv se extrajo de partículas víricas precipitadas (25.000 × g durante 2 h) usando el kit de ARN vírico QIAmp (Qiagen). Aproximadamente 10.000 copias de ARN vírico de cada muestra estaban presentes en la reacción de síntesis de ADNc como se describe. El cebador de etiquetado usado fue, 5'-GCCTTGCCAGCACGCTCAGGCCTTGCA(CÓDIGO DE BARRAS)CGNNNNNNNNTCCTGGCTTTAATTTTACTGGTACAGT-3'. (SEQ ID NO. 2). El código de barras representaba TCA, GTA y TAT para los días de estudio 58, 248 y 303, respectivamente. El extremo 3' de cebador de etiquetado se dirigía a después del dominio codificante de la proteasa (HXB2 2568-2594). Los oligonucleótidos se compraron de IDT y se purificaron por desalado estándar.

Amplificación de secuencias etiquetadas. El ADNc monocatenario se purificó en columna usando el kit de purificación de PCR PureLink (Invitrogen), usando tampón de unión HC (alto valor de corte) y tres lavados para eliminar el cebador de ADNc. La eliminación del cebador se verificó por análisis de electroferograma usando un chip microfluídico de ARN Experion HighSense (Bio-Rad Laboratories). Las muestras se amplificaron por PCR anidada usando los cebadores anteriores 5'-GAGAGACAGGCTAATTTTTTAGG-3' (HXB2 2071-2093) (SEQ ID NO. 3) y 5'-ATAGACAAGGAACTGTATCC-3' (HXB2 2224-2243) (SEQ ID NO. 4); los cebadores posteriores dirigidos a la porción 5' del cebador de etiquetado del ADNc 5-GCCTTGCCAGCACGCTCAGGC-3' (SEQ ID NO. 5) después 5'-CCAGCACGCTCAGGCCTTGCA-3' (SEQ ID NO. 6). La PCR se hizo usando ADN polimerasa Platinmun Taq High Fidelity (Invitrogen). Cada reacción contenía 1 × tampón de PCR High Fidelity, 0,2 mM de cada dNTP, MgCl2 2 mM, 0,2 µM de cada cebador, 1,5 unidades de ADN polimerasa Platinum Taq. El molde de ADNc purificado se separó en 2 × 50 µl para la primera ronda de PCR, y se usó 1 µl del producto de primera ronda purificado para la PCR anidada. Las muestras se desnaturalizaron a 94ºC durante 2 min, seguido por 30 ciclos de 94ºC durante 15 s, 55ºC durante 30 s, 68ºC durante 1 min, y una extensión final a 68ºC durante 5 min.

Las muestras se purificaron en columna después de la primera ronda de PCR usando el kit de purificación de PCR MiniElute (Qiagen), y se eluyeron en 30 µl de tampón EB. El producto de PCR de la segunda ronda se purificó en gel usando un gel de agarosa al 2% y el kit de extracción de gel QIAquick (Qiagen), con incubación del tampón de solubilización a temperatura ambiente. El ADN se cuantificó por fluorómetro Qubit usando el ensayo dsDNA High Sense (Invitrogen). La generación de producto, calidad, y eliminación de cebador para ambas rondas de PCR se verificó usando un chip microfluídico de ADN Experion (Bio-Rad).

Pirosecuenciación en 454. Las muestras etiquetadas de los tres puntos de tiempo se combinaron y secuenciaron en la plataforma 454 GS FLX con química de secuenciación XLR70 Titanium según las instrucciones del fabricante (Roche), pero con bolas poco cargadas para minimizar la interferencia de señales. Las secuencias se procesaron de dos carreras de 454 GS FLX Titanium independientes (1/8 de una placa cada una).

Tubería bioinformática para procesamiento de secuencias en bruto. Se escribió un paquete de programas para filtrar y analizar lecturas de secuenciación 454 en bruto. Brevemente, primero, cada secuencia se colocó en la orientación correcta comparada con una secuencia pro gen de referencia. Este alineamiento se usó después para identificar inserciones o deleciones causadas por la secuenciación 454 de homopolímeros. Cuando había una inserción, la base extraña se cortó de la secuencia. Las deleciones retenidas se resolvían en gran parte en la construcción de la secuencia consenso. Segundo, se evaluó la presencia de la cola 5' de cebador de ADNc, con la información codificada (código de barras e ID de cebador) exactamente separada. Tercero, las secuencias individuales se archivaron por sus códigos de barras, y después por ID de cebador. Cuarto, las secuencias se recortaron al dominio codificante de la proteasa (pro gen). En un archivo de código de barras, cuando tres secuencias contenían un ID de cebador idéntico, se asignó una secuencia consenso por la regla de la mayoría. Se usaron designaciones de nucleótidos ambiguas cuando había un vínculo (Fig. 6B). Las secuencias están disponibles con los números de registro de GenBank JN820319-JN824997.

Análisis de población. Se usó una prueba χ^2 para ensayar cambios de significación en la frecuencia alélica entre los dos puntos de tiempo no tratados. Para controlar ensayos múltiples, la evaluación colectiva de la significación se basó en análisis de tasa de descubrimiento falsa (FDR = 0,05). Las pruebas para el desequilibrio de ligamiento se computaron mediante DnaSP v.5.10.01 (4). Estas pruebas se hicieron en poblaciones filtradas desprovistas de secuencias que contienen ambigüedades o huecos. Las pruebas para neutralidad se computaron mediante DnaSP y R (5) en poblaciones filtradas desprovistas de secuencias que contienen ambigüedades. Los huecos y alelos representados por una secuencia única se revirtieron al consenso. Se calcularon valores P beta contra la hipótesis nula que D = 0, asumiendo que D sigue una distribución beta después de reajuste en [0, 1].

La diversidad a través y dentro de las poblaciones se computó mediante paquetes bioinformáticos a medida. Se usaron secuencias no filtradas en el análisis, y ambigüedades, huecos y alelos representados por una única secuencia se eliminaron de la tabulación final (Fig. 2A-2B y Tabla 1).

Los SNP se mostraron gráficamente mediante la herramienta Highlighter (www.hiv.lanl.gov).

Resolución filogénica de secuencias. La filogenia para la población de secuencias consenso de los tres puntos de tiempo se resolvió usando dos métodos alternativos y en poblaciones desprovistas de secuencias que contienen huecos o ambigüedades. Cuando solo estaba presente un ejemplo de un SNP a través de todas las secuencias, se convirtió al consenso en la suposición de que estaba probablemente generada por error de método residual. Primero, se construyó el árbol de unión de vecinos usando la traducción de Kimura para distancia por pares y un método de muestreo repetido (Bootstrap) de 100 iteraciones con QuickTree v.1.1.

Segundo, se infirió la filogenia de máxima probabilidad usando el paquete PHYLIP, versión 3.69, y la filogenia calculada está disponible bajo demanda. Se usó el programa PHYLIP seqboot para crear 100 muestreos repetidos. Los muestreos repetidos resultantes se sometieron al programa PHYLIP dnamlk para inferencia de máxima probabilidad sujeto a un reloj molecular estricto. El árbol consenso de todos los resultados de muestreo repetido se construyó usando el consenso del programa PHYLIP.

Ambos árboles filogenéticos se visualizaron por una modificación personalizada de Figtree v.1.3.1.

Consideraciones adicionales. Síntesis de base degenerada en el cebador de ADNc. Las bases degeneradas (ID de cebador) en el cebador de síntesis de ADNc se aleatorizaron usando mezclado a máquina durante la síntesis de oligonucleótidos. Las cuatro bases monómeros de fosforamidita de ADN se introducen a la columna al mismo tiempo, pero debido a ligeras diferencias en la unión o administración, una proporción equimolar estricta de dA, dT, dC y dG puede no realizarse, dando un sesgo del ID de cebador (Fig. 10). Cuando hay un sesgo del ID de cebador, hay una probabilidad aumentada de que un ID de cebador particular etiquete múltiples moldes porque las etiquetas de secuencia con nucleótidos sobrerrepresentados serán más abundantes que las etiquetas de secuencia con nucleótidos subrepresentados. Puesto que el sesgo se amplifica a lo largo de la longitud del ID de cebador la distorsión puede ser significativa. Observamos un sesgo de ~40% dC en una de las síntesis de ID de cebador, y en el extremo dC_8 estaría presente en un exceso de 40 veces sobre la frecuencia de secuencia esperada si todos los nucleótidos estuvieran presentes a una concentración igual. De forma similar, observamos el 15% de dA en una síntesis que produciría una disminución de 60 veces en la frecuencia esperada de dA_8. Esto parece ser el resultado en variación en la síntesis de cebador porque el sesgo varió en los diferentes archivos de códigos de barras y por tanto no era una característica constante de la etapa de síntesis de ADNc. Sin embargo, este fenómeno se mitiga de alguna manera cuando se forma una secuencia consenso, ya que cualquier molde se remuestreó a mayor grado en una población de ID de cebador mezclada se registraría.

Mutaciones de cambio de marco de lectura. La pirosecuenciación comúnmente asigna mal homopolímeros, lo que produce mutaciones de cambio de marco de lectura o bien por asignar demasiado pocos o demasiados nucleótidos en la carrera del homopolímero. El pro gen de VIH-1 contiene varios tramos homopoliméricos. Tomamos ventaja de una longitud conocida (conservada en una región codificante) para alinear lecturas individuales frente a una secuencia de referencia. Dado este sesgo eliminamos las inserciones para retener la longitud correcta de la carrera del homopolímero. Las deleciones se retuvieron. Mediante la generación de la secuencia consenso, la base delecionada con frecuencia se recuperó cuando las otras lecturas remuestreadas contenían la base que falta. Aunque la generación de la secuencia consenso redujo la expansión y frecuencia de deleciones en las lecturas consenso finales, resueltas, no eliminó deleciones del todo (Fig. 11).

Ejemplo 2. Muestreo preciso y secuenciación profunda de proteasa de VIH-1 usando un ID de cebador

Los virus pueden crear poblaciones genéticas complejas dentro de un huésped, y las tecnologías de secuenciación profunda ofrecen la oportunidad de muestrear extensamente estas poblaciones. Sin embargo, características de estas técnicas limitan su aplicación, en particular cuando una etapa de reacción en cadena de la polimerasa (PCR) precede al protocolo de secuenciación.

Típicamente, se utilizan un número desconocido de moldes en iniciar la amplificación por PCR y esto puede producir remuestreo de secuencia no reconocido. La recombinación mediada por PCR puede crear ligamiento artificial y desorganizar el ligamiento real. Por último, la mala incorporación durante la PCR y los errores durante el protocolo de secuenciación pueden crear diversidad artefactual.

Hemos resuelto esto incluyendo una etiqueta de secuencia aleatoria en el cebador inicial de modo que cada molde recibe un ID de cebador. Después de secuenciar, la identificación repetitiva del ID de cebador revela remuestreo de secuencia, que se puede usar entonces para crear una secuencia consenso exacta para cada molde. La población resultante de secuencias consenso directamente identifica los moldes muestreados iniciales. El uso de los ID de cebador puede corregir directamente el remuestreo de secuencia no reconocido, recombinación mediada por PCR, amplificación de molde diferencial, mala incorporación de nucleótido por la polimerasa, y error de secuenciación.

Aplicamos este enfoque al (pro) gen de la proteasa de VIH-1 para ver la distribución de la variación de secuencia en una población compleja. Identificamos polimorfismos principales y secundarios en posiciones codificantes y no

codificantes. Además, observamos cambios dinámicos a través de la población durante exposición intermitente a fármaco, incluyendo la aparición de alelos resistentes.

Métodos: Se diseñó una población de cebadores de síntesis de ADNc para que contuvieran una cadena de ocho nucleótidos degenerados (65.536 combinaciones de secuencias distintas, o ID de cebador), y un código de barras de tres nucleótidos seleccionado *a priori* (Fig. 12A). Después de la síntesis de ADNc, el extremo 5' no específico del cebador se usó para enriquecimiento de secuencias etiquetadas por PCR anidada.

Se extrajo el ARN de VIH-1 de plasma sanguíneo. La proteasa de dos pre-terapia de ritonavir y post-terapia intermitente de ritonavir se etiquetó y secuenció (Fig. 12B). Aproximadamente 10.000 copias de ARN vírico de VIH-1 se etiquetaron después de proteasas, se amplificaron y secuenciaron en el 454 GS FLX Titanium. La figura 13B muestra un resumen de las secuencias resueltas.

Se desarrolló una tubería bioinformática para evaluar las secuencias en bruto para proteasa de longitud completa etiquetada (Fig. 14), y cuando tres o más secuencias contenían un ID de cebador idéntico, se generó una secuencia consenso (Fig. 15). Después de corrección de errores directa por filtración y procesamiento de los ID de cebador, se evaluó la frecuencia alélica de las poblaciones de pro gen (Fig. 13A).

Discusión: Hemos desarrollado una estrategia que permite que cada molde muestreado se etiquete con un ID de secuencia mediante un cebador que tiene una etiqueta de secuencia degenerada incorporada durante la síntesis. Esta etiqueta se puede seguir después a través del protocolo de secuenciación profunda para identificar cobertura excesiva de secuenciación de los moldes individuales. La cobertura excesiva se puede usar para crear una secuencia consenso para cada molde, evitando tanto errores relacionados con PCR como errores de secuenciación. Además, el número de diferentes ID de cebador refleja el número de moldes que se muestrearon realmente, lo que permite una evaluación realista de la calidad del muestreo que hará posible aplicar un análisis más riguroso de variantes minoritarias. En muchos marcos, especialmente cuando se trabaja con agentes patógenos en muestras clínicas, el número de moldes puede ser limitante y el uso de PCR puede oscurecer la calidad limitada del muestreo. Este problema se resuelve etiquetando cada molde como la primera etapa y después simplemente contando el número de moldes que se usaron realmente como parte de la última etapa.

Resolver de forma precisa los moldes víricos tiene un coste de profundidad de secuenciación, ya que la profundidad es una función directa del número de moldes, pero hemos documentado la naturaleza y grado de fluctuación alélica en toda la proteasa hasta el nivel del 0,1% de resolución. Mostramos fluctuación sinónima y no sinónima a lo largo del tiempo en un medio sin tratar, y cuando en un medio de selección intermitente del inhibidor de proteasa, Ritonavir. Detectamos el alelo resistente a fármaco, V82A, como un alelo minoritario en la población sin tratar, y detectamos esta variante exacta sobre un cuarto de la población que repunta (Fig. 16). Estos resultados demuestran la utilidad de aplicar este enfoque a la detección de variantes minoritarias en el contexto de tratamiento de VIH-1, y más en general a la pregunta de variantes minoritarias en el contexto de una población genéticamente muy compleja.

Tabla 1. Frecuencia de codones no consenso por posición

AApos[a]	AAc[b]	Cc[c]	Cm[d]	Am[e]	T1[f]	T2[g]	T3[h]	T3s[i]	T3r[j]	Cm[k]	T1[l]	T2[m]	T3[n]	T3s[o]	T3r[p]
4	T	ACT	GCT	A		0.06	0.05		0.09						
5	L	CTT	CCT	P	0.12		0.05	0.14							
7	Q	CAA								CAG	0.35	0.12	0.09	0.14	0.09
8	R	CGA								CGG	0.12		0.05	0.14	
9	P	CCC													
10	L	CTC	TTC	F		0.19				CTT		0.19			
11	V	GTC	ATC	I	0.23	0.25				GTT		0.12			
14	K	AAG	AGG	R		0.12				AAA	1.17	0.19	0.59	0.29	0.72
15	I	ATA	GTA	V	1.17	0.12	0.14	0.14	0.18	ATC			0.09		0.18
16	G	GGG	AGG	R		0.06	0.05		0.09	GGA	2.22	3.54	38.86	17.70	45.97
17	G	GGG	AGG	R			0.09	0.29		GGA	0.35	0.19	0.18	0.43	0.09
18	Q	CAA	GAA	E	0.23	0.12				CAG	18.55	21.75	6.46	12.81	3.53
19	L	CTA	ACA	T	0.47										
			ATA	I	19.25	19.83	20.42	19.28	24.98	TTA	0.12	0.19	0.09	0.29	
			GTA	V	3.38	5.66	46.00	25.61	52.76						
20	K	AAG	AGG	R	0.12	0.12	0.05		0.09	AAA		0.31	0.86	0.29	1.27
21	E	GAA								GAG	0.12	0.06	0.05	0.14	
22	A	GCT								GCC	0.47	0.44	0.27	0.58	0.18
										GCG	0.23				
23	L	CTA								CTG		0.19			
24	L	TTA								CTA	0.35	5.72	1.31	2.16	0.63
										TTG	12.49	0.81	0.59	1.01	0.27
25	D	GAT	GGT	G	0.12	0.12				GAC	0.23	0.93	0.05	0.14	
26	T	ACA	GCA	A		0.12									
27	G	GGA								GGG	0.12	0.06			
28	A	GCA								GCG	0.12		0.09	0.14	
29	D	GAT	AAT	N	0.12		0.05		0.09	GAC	0.23	0.19			
30	D	GAT								GAC		0.06	0.09	0.14	0.09
31	T	ACA								ACG		0.12			
32	V	GTA								GTG		0.25			
33	L	TTA	GTA	V	0.47	0.06				CTA		0.25	0.14	0.29	0.09
										TTG	0.35	0.12	0.14	0.43	
34	E	GAA	GGA	G		0.12	0.05		0.09	GAG	0.12		0.05	0.14	
			CAA				0.09								
35	E	GAA	AAA	K	0.12	0.06	0.09	0.14							

AApos[a]	AAc[b]	Cc[c]	Cm[d]	Am[e]	T1[f]	T2[g]	T3[h]	T3s[i]	T3r[j]	Cm[k]	T1[l]	T2[m]	T3[n]	T3s[o]	T3r[p]
36	M	ATG	ATA	I	0.82	0.81	0.27	0.43	0.27						
37	N	AAT	AGT	S		0.19	0.05			AAC		0.06	0.05	0.14	
			GAT	D	2.33	2.30	0.95	0.86	1.27						
38	L	TTG								TTA	0.23	0.62	0.05		0.09
39	P	CCA								CCT	0.23				
40	G	GGA								GGG	0.12	0.12			
41	K	AAA	AGA	R		0.06	0.18	0.14	0.27	AAG	4.08	1.43	0.50	1.15	0.27
42	W	TGG	CGG	R	0.12	0.06									
			TAG		0.12		0.05		0.09						
			TGA				0.14		0.27						
43	K	AAA	AGA	R		0.06	0.05		0.09	AAG	0.35		0.14	0.14	0.18
44	P	CCA								CCG		0.06	0.23	0.43	0.18
45	K	AAA	AGA	R	0.12	0.12	0.05		0.09	AAG	0.58	0.99	0.41	1.29	
46	M	ATG	ATA	I		0.12	0.09	0.14	0.09						
48	G	GGA	GAA	E			0.14	0.14	0.18	GGG	0.35	0.19			
49	G	GGA	GAA	E	0.12	0.06	0.05		0.09	GGG	0.23	0.12			
50	I	ATT								ATC	0.12	0.12			
51	G	GGA								GGG	0.12	0.06			
52	G	GGT	AGT	S	0.12	0.06	0.05	0.14		GGA		0.06	0.05	0.14	
										GGC	0.12	0.31	0.09	0.14	0.09
										GGG			0.14	0.43	
53	F	TTT								TTC	0.70		0.05	0.14	
54	I	ATC	ACC	T	0.12	0.06	0.05		0.09	ATT	0.35	0.06	0.14	0.14	
55	K	AAA	AGA	R	0.12		0.05		0.09	AAG	0.12	0.06			
56	V	GTA	ATA	I	0.12		0.05	0.14		GTG		0.75	0.14	0.14	0.18
57	R	AGA	AAA	K	0.23					AGG	0.23	0.87	0.14	0.14	0.18
58	Q	CAG	TAG				0.05		0.09	CAA	0.93	0.50	0.23	0.29	0.27
60	D	GAT	AAT	N		0.12									
			GGT	G		0.12									
61	Q	CAA	CGA	R	0.12	0.06	0.05	0.14		CAG		0.19	0.23	0.58	
			TAA		0.12	0.06	0.05		0.09						
62	I	ATA	GTA	V	0.35	0.06									
63	L	CTC	CCC	P	0.12		0.41	0.58	0.36	CTT	11.32	5.41	1.27	2.88	0.45
64	I	ATA	GTA	V	1.05	0.06	0.09		0.18						
			ATG	M	0.23		0.05	0.14							
65	E	GAA	AAA	K			0.09	0.14	0.09	GAG	0.35	0.06	0.05		
66	I	ATC								ATA		0.25	0.18	0.58	
										ATT	1.98	0.19			
67	C	TGT								TGC	0.35	0.12	0.05	0.14	

AApos[a]	AAc[b]	Cc[c]	Cm[d]	Am[e]	T1[f]	T2[g]	T3[h]	T3s[i]	T3r[j]	Cm[k]	T1[l]	T2[m]	T3[n]	T3s[o]	T3r[p]
68	G	GGA								GGG	0.23	0.12	0.05	0.14	
69	H	CAT	TAT	Y	0.23	0.06	0.09	0.14		CAC	0.82	0.31	0.14	0.29	0.09
70	K	AAA	CAA	Q	0.47	0.12	0.41	1.29		AAG	3.27	10.88	15.27	6.62	25.34
71	A	GCT	ACT	T		0.12	0.09								
72	I	ATA	GTA	V	0.12	0.12									
73	G	GGT								GGC	0.47	18.09	7.05	15.68	3.62
74	T	ACA								ACG	0.23	0.12			
75	V	GTA	ATA	I	0.23	0.06	0.05			GTG	1.87	0.99	0.27	0.43	0.27
			GCA	A			0.09		0.18						
76	L	TTA								CTA		0.12	0.09		0.18
										TTG	0.93	0.62	0.27	0.43	0.18
77	V	GTA	ATA	I	0.23	0.56	0.72	2.01	0.18	GTG	0.82	0.62	0.23	0.58	
			CTA	L			0.14								
78	G	GGA								GGG	1.17	1.24	0.09	0.14	
79	P	CCT								CCC	1.17	0.31	0.54	1.29	0.18
81	P	CCT								CCC	0.12	0.19			
										CCG	1.52	0.44			
82	V	GTC	ATC	I		0.06	1.27	3.60		GTA	0.35	0.31	0.05		
			CTC	L		0.06	1.08	3.45		GTT	1.05	0.75	0.41	1.01	
			GCC	A		0.12	49.89		99.91						
			TTC	F			0.14	0.43							
83	N	AAC	AGC	S	0.12		0.05		0.09	AAT	8.17	6.40	3.62	4.75	4.16
84	I	ATA	GTA	V			5.15								
85	I	ATT								ATA		0.12	0.05	0.14	
										ATC	0.12	0.12	0.05		
86	G	GGA								GGG		0.12			
										GGT	0.12	0.06			
87	R	AGA	AAA	K	0.12	0.06	0.05		0.09	AGG	0.58	0.37	0.05	0.14	
			GGA	G		0.06	0.09	0.14	0.09						
88	N	AAT								AAC	0.35	0.93			
89	L	CTA	ATA	I		0.12				CTG	1.17	0.68	1.36	1.87	1.54
										TTA	1.98	0.56	1.27	0.14	2.44
90	L	TTG	ATG	M	0.12		13.56		0.09	CTG	0.47		0.09	0.14	0.09
			TCG	S	0.12		0.05		0.09	TTA	0.47	0.19	0.14	0.43	
91	T	ACT	GCT	A		0.06	0.05		0.09	ACC	0.12	0.06	0.09	0.14	0.09
										ACG	0.12	0.12	0.77		1.54
92	Q	CAG								CAA	0.23	0.19	0.14		
93	I	ATT	CTT	L	0.12	0.06				ATC	0.23		0.09	0.14	0.09

AApos[a]	AAc[b]	Cc[c]	Cm[d]	Am[e]	T1[f]	T2[g]	T3[h]	T3s[i]	T3r[j]	Cm[k]	T1[l]	T2[m]	T3[n]	T3s[o]	T3r[p]
94	G	GGT	GAT	D	0.12	0.06				GGA	0.23				
										GGC	1.28	0.25	0.50	1.29	0.18
										GGG	0.23	0.06	0.09	0.14	
95	C	TGC								TGT	0.70	0.12	0.14		0.27
96	T	ACT								ACA	0.12		0.09	0.14	0.09
										ACC	0.70	0.12	0.23	0.43	0.09
										ACG		0.06	0.05	0.14	
97	L	TTA								CTA	0.58		0.05	0.14	
										TTG	0.12	0.25	0.27	0.43	0.27

Solo se muestran las posiciones de diversidad y SNP que estaban representados por más de 1 secuencia.

Consenso
AApos[a] Posición de aminoácido, proteasa.
AAc[b] Aminoácido consenso en población sin tratar.
Cc[c] Codón consenso en población sin tratar.

No sinónimo
Cm[d] Aminoácido no consenso codificante.
AAm[e] Codón no consenso codificante.
T1[f] Frecuencia de SNP en el primer punto de tiempo sin tratar.
T2[g] Frecuencia de SNP en el segundo punto de tiempo sin tratar.
T3[h] Frecuencia de SNP en el tercer punto de tiempo, tratado.
T3[i] Frecuencia de SNP en el tercer punto de tiempo, tratado, población susceptible (no V82A, I84V, L90M).
T3r[j] Frecuencia de SNP en el tercer punto de tiempo, tratado, población que contiene la variante a ritonavir principal V82A.

Sinónimo
Cm[k] Codón no consenso silencioso
T1[l] Frecuencia de SNP en el primer punto de tiempo sin tratar.
T2[m] Frecuencia de SNP en el segundo punto de tiempo sin tratar.
T3[n] Frecuencia de SNP en el tercer punto de tiempo, tratado.
T3s[o] Frecuencia de SNP en el tercer punto de tiempo, tratado, población susceptible (no V82A, I84V, L90M).
T3r[p] Frecuencia de SNP en el tercer punto de tiempo, tratado, población que contiene la variante a ritonavir principal V82A.

Tabla 2. Resumen de variación de nucleótidos en puntos de tiempo muestreados

Variables	T1	T2	T3	T3s	T3r
No. se secuencias	810	1449	1925	547	970
No. de polimórficos	104	115	110	71	69
sitios (segregantes)					
Número total de mutaciones	115	129	121	75	73
Número medio de diferencias de nt, k	2,38809	2,33683	3,08838	2,43819	2,05962
Diversidad de nucleótidos, π	0,00804	0,00787	0,01040	0,00821	0,00693
Theta (por secuencia)	14,29822	14,63943	13,51412	10,31864	9,25678
Theta (por sitio)	0,04814	0,04929	0,04550	0,03474	0,03117
D de Tajima	-2,3541	-2,3164	-2,0937	-2,1606	-2,1209
Valor P beta	0,0013	0,0014	0,0070	0,0065	0,0071

T1 y T2 son poblaciones sin tratar, y T3 es una población intermitentemente expuesta a monoterapia con ritonavir. Dentro de T3, T3s representa la porción sensible (no V82A, I84V o L90M) de la población. T3r representa el clado de resistencia a fármaco principal V82A.

Lista de secuencias

<110> LA UNIVERSIDAD DE CAROLINA DEL NORTE EN CHAPEL HILL

<120> MÉTODOS Y USOS PARA ETIQUETAS MOLECUALRES

<130> UNC12003WO

<150> 61/603.909
<151> 27-02-2012

<160> 6

<170> PatentIn versión 3.5

<210> 1
<211> 11
<212> ADN
<213> Secuencia artificial

<220>
<223> .

<400> 1
cataatacta g 11

<210> 2
<211> 67
<212> ADN
<213> Secuencia artificial

<220>
<223> .

<220>
<221> característica_misc
<222> (28)..(30)
<223> n es a, c, g o t

<220>
<221> característica_misc
<222> (33)..(40)
<223> n es a, c, g o t

<400>
gccttgccag cacgctcagg ccttgcannn cgnnnnnnnn tcctggcttt aattttactg 60

gtacagt 67

<210> 3
<211> 23
<212> ADN
<213> Secuencia artificial

<220>
<223> .

<400> 3
gagagacagg ctaatttttt agg 23

<210> 4
<211> 20
<212> ADN
<213> Secuencia artificial

<220>
<223> .

<400> 4
atagacaagg aactgtatcc 20

<210> 5
<211> 19
<212> ADN
<213> Secuencia artificial

<220>
<223> .

<400> 5
gaagtactgc tcgtaggag 19

<210> 6
<211> 21
<212> ADN
<213> Secuencia artificial

<220>
<223> .

<400> 6
ccagcacgct caggccttgc a 21

REIVINDICACIONES

1. Un método para analizar una pluralidad de moléculas de ácido nucleico que comprende:

(a) unir una pluralidad de cebadores que comprenden un ID de cebador a una pluralidad de moléculas de ácido nucleico en una muestra para generar moldes de ácido nucleico etiquetados, en donde
   (i) la pluralidad de moléculas de ácido nucleico comprende 10 o más moldes de ácido nucleico, y
   (ii) cada molde de ácido nucleico etiquetado está unido a un ID de cebador único;
(b) amplificar los moldes de ácido nucleico etiquetados para producir amplicones etiquetados;
(c) detectar las amplicones etiquetados, analizando mediante ello la pluralidad de moléculas de ácido nucleico; y
(d) determinar un sesgo de amplificación de la reacción de amplificación basado en la detección de las moléculas de ácido nucleico etiquetado, en donde determinar el sesgo de amplificación se basa en la comparación de dos o más proporciones, en donde la comparación de las dos o más proporciones comprende comparar una primera proporción de la cuantificación de diferentes ID de cebador asociados con dos o más tipos de moléculas de ácido nucleico a una segunda proporción de la cuantificación del número total de amplicones de dos o más tipos de moléculas de ácido nucleico,
en donde la primera proporción se basa en una cantidad de diferentes ID de cebador que se asocian con un primer tipo de molécula de ácido nucleico y una cantidad de diferentes ID de cebador asociados con un segundo tipo de molécula de ácido nucleico,
en donde la segunda proporción se basa en una cantidad de amplicones totales que están asociados con el primer tipo de moléculas de ácido nucleico y una cantidad de amplicones totales que están asociados con el segundo tipo de moléculas de ácido nucleico y
en donde el sesgo de amplificación se revela por la diferencia en la primera proporción y la segunda proporción.

2. El método de la reivindicación 1, en donde el molde de ácido nucleico comprende un molde de ADN.

3. El método de la reivindicación 1, en donde el molde de ácido nucleico comprende un molde de ARN.

4. El método de las reivindicaciones 1-3, en donde el molde de ácido nucleico comprende una secuencia de ácido nucleico vírico.

5. El método de las reivindicaciones 1-4, en donde el cebador que comprende el ID de cebador comprende además una secuencia diana específica complementaria al molde de ácido nucleico.

Fig. 1A

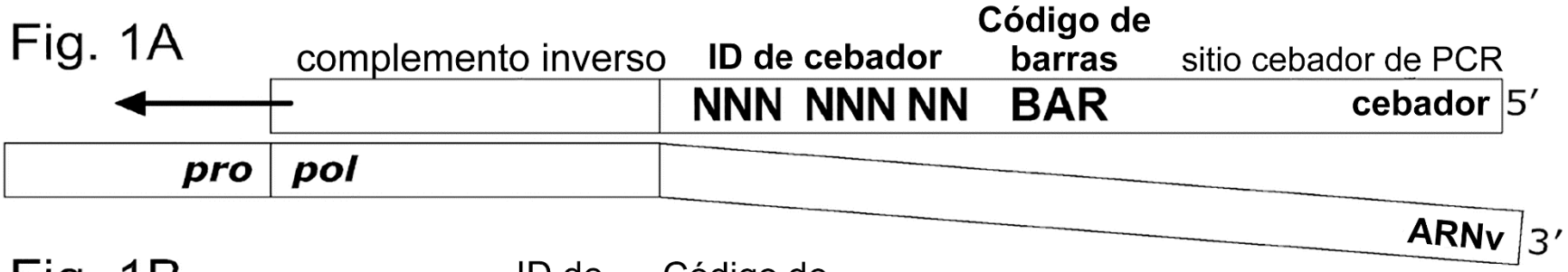


Fig. 1B

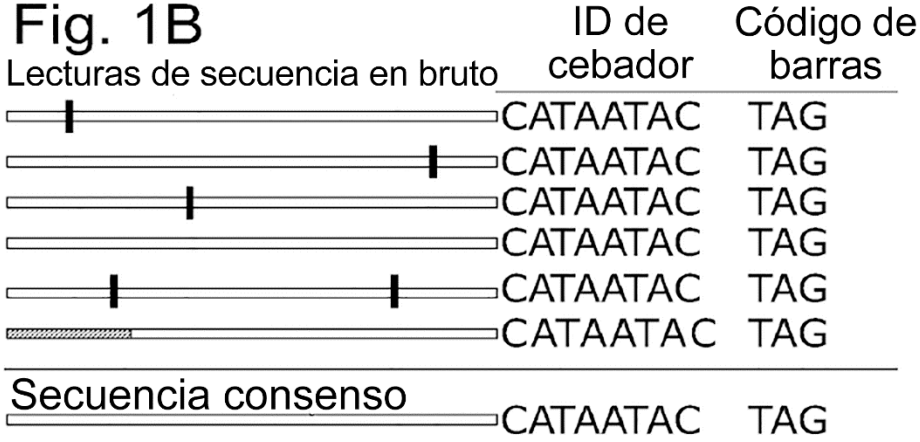


Fig. 1C

Muestra	**T1**	**T2**	**T3**
Ritonavir	-	-	+
Lecturas totales	20.429	24.658	27.075
Secuencias consenso	857	1.609	2.213

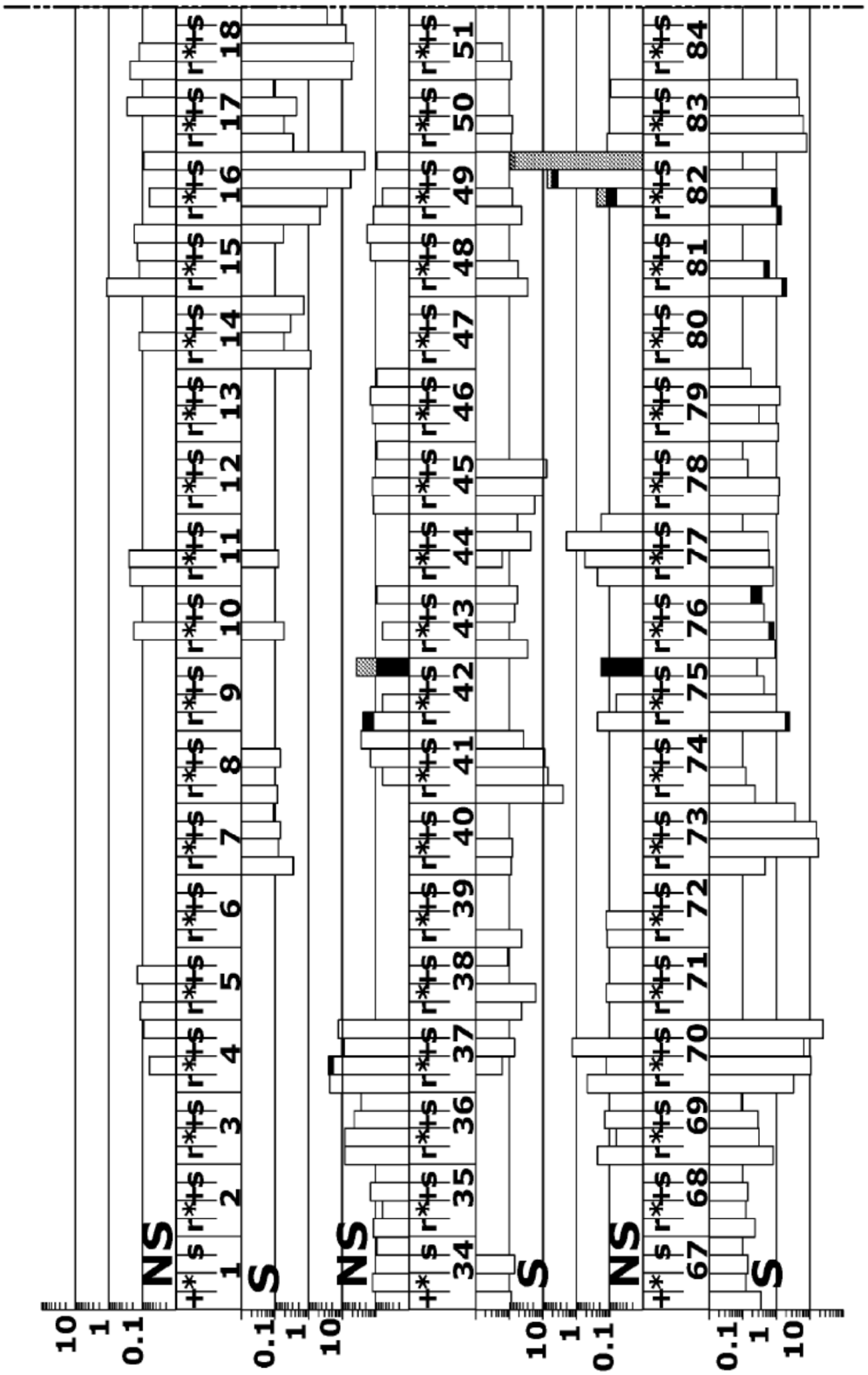


Fig. 2A

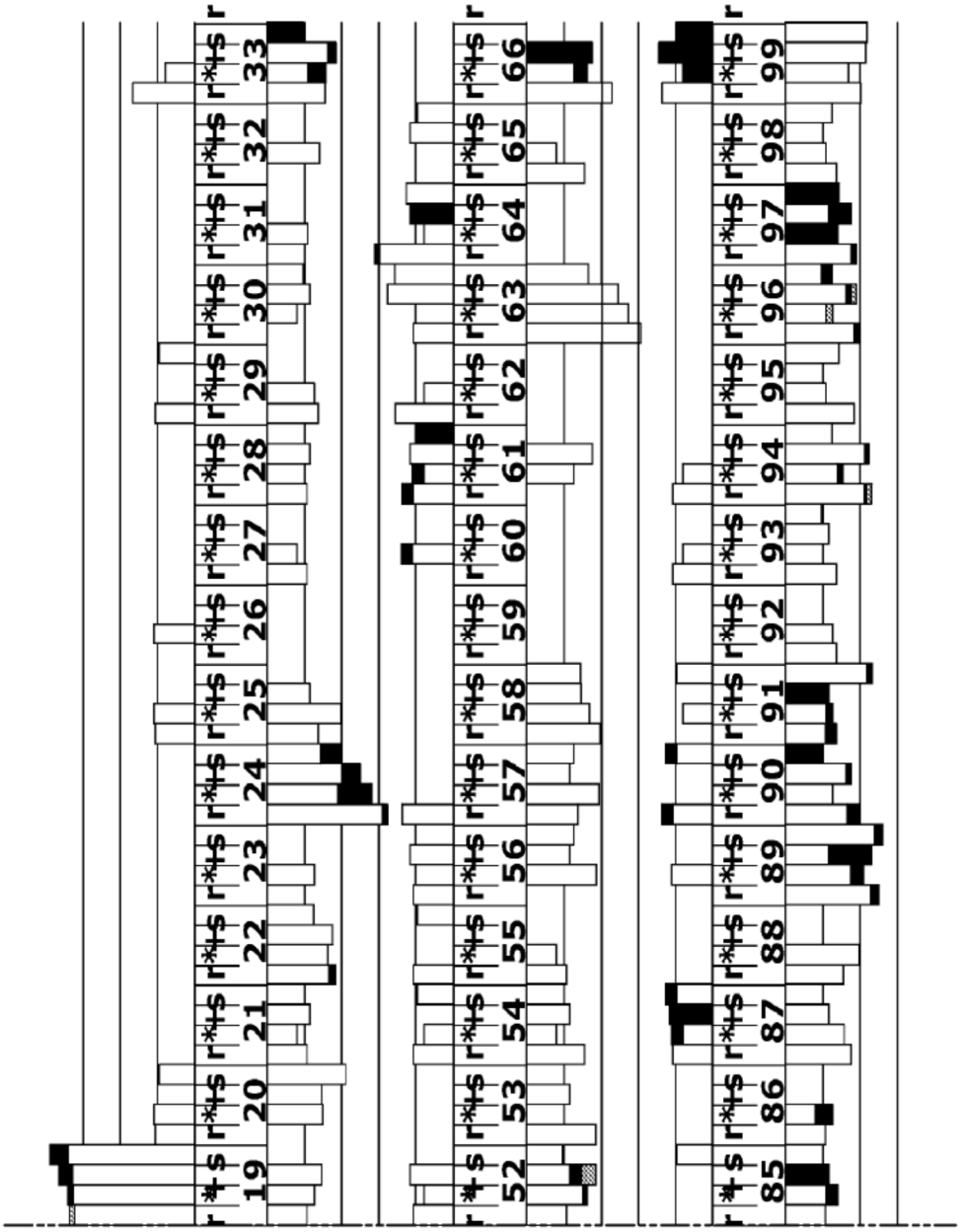


Fig. 2B

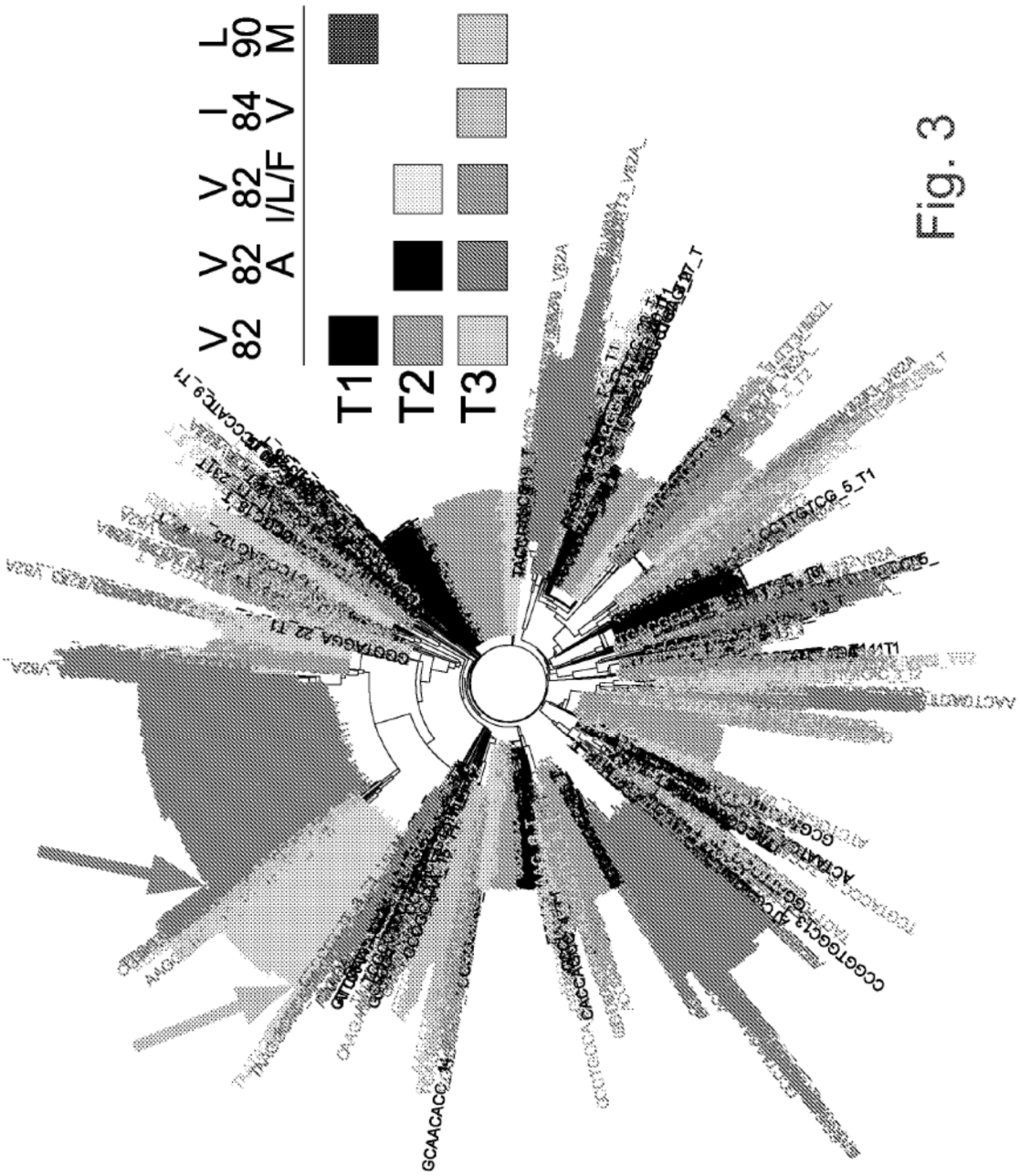


Fig. 3

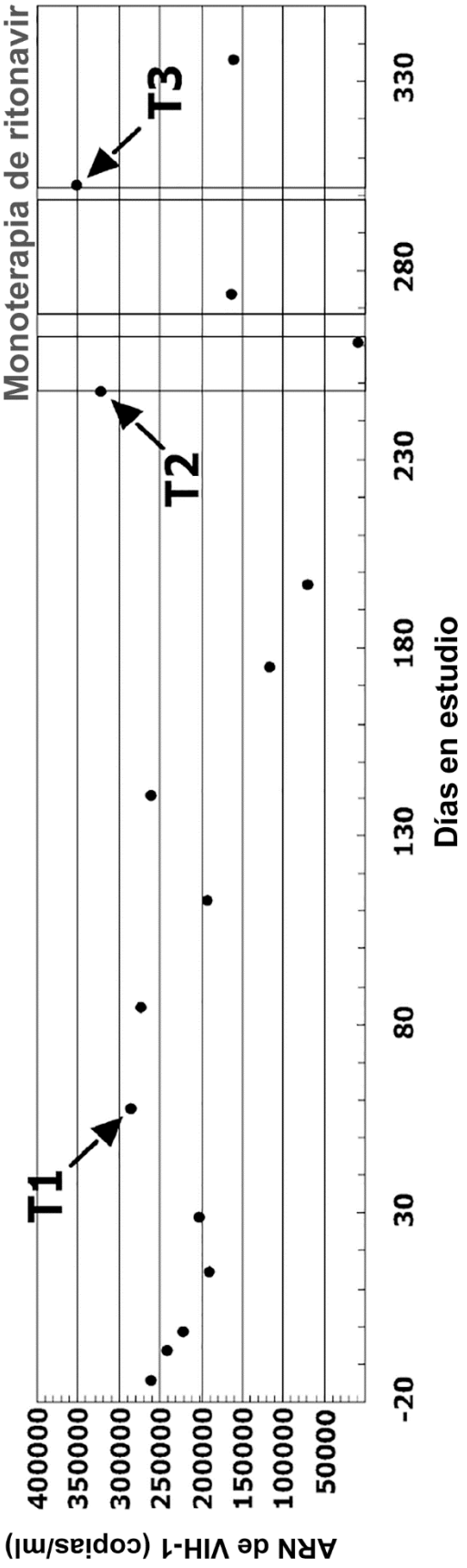

Monoterapia de ritonavir
T1
T2
T3
ARN de VIH-1 (copias/ml)
400000
350000
300000
250000
200000
150000
100000
50000
-20
30
80
130
180
230
280
330
Días en estudio

Fig. 4

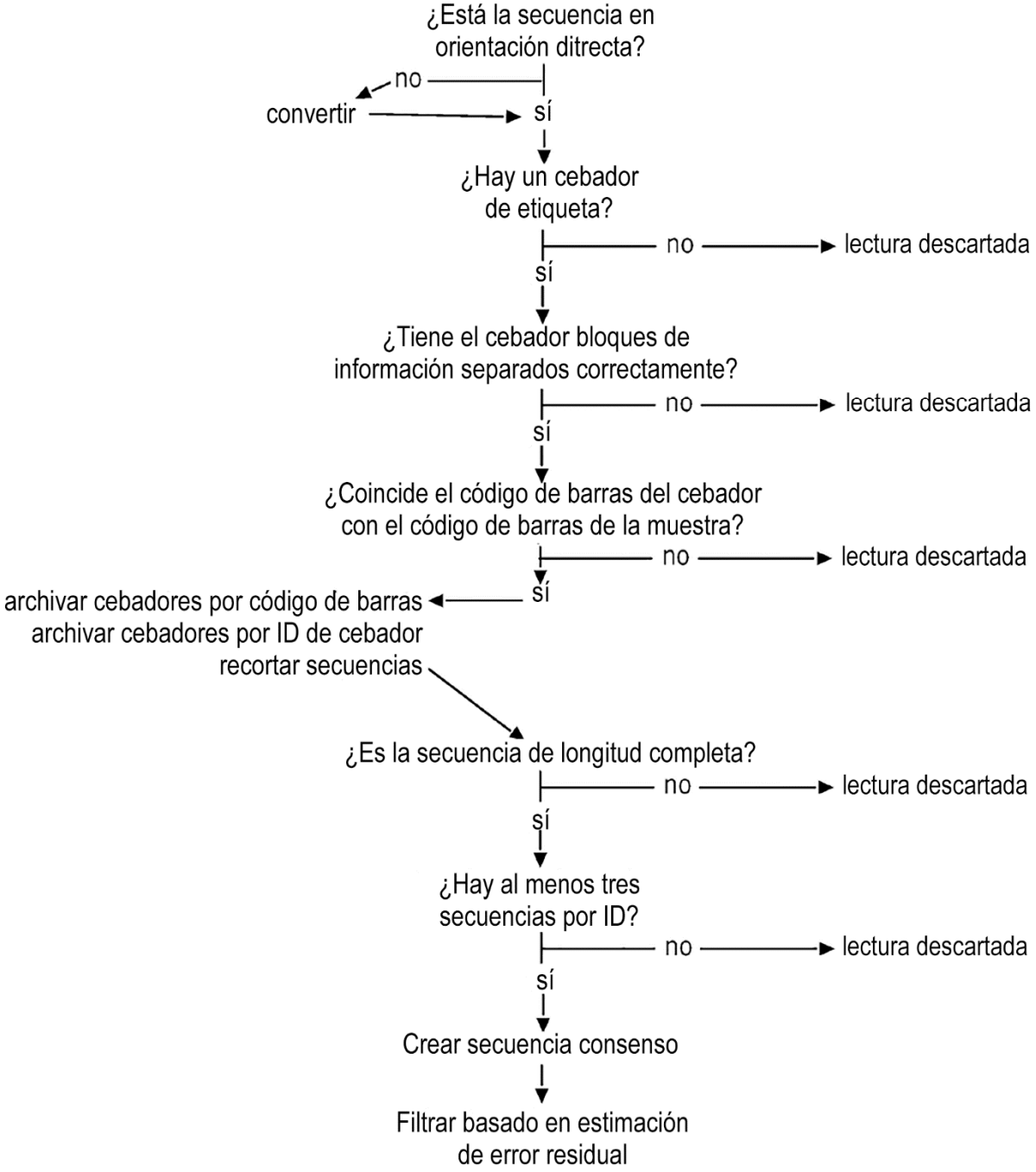

¿Está la secuencia en orientación ditrecta?
no
convertir
sí
¿Hay un cebador de etiqueta?
no
lectura descartada
sí
¿Tiene el cebador bloques de información separados correctamente?
no
lectura descartada
sí
¿Coincide el código de barras del cebador con el código de barras de la muestra?
no
lectura descartada
sí
archivar cebadores por código de barras
archivar cebadores por ID de cebador
recortar secuencias
¿Es la secuencia de longitud completa?
no
lectura descartada
sí
¿Hay al menos tres secuencias por ID?
no
lectura descartada
sí
Crear secuencia consenso
Filtrar basado en estimación de error residual

Fig. 5

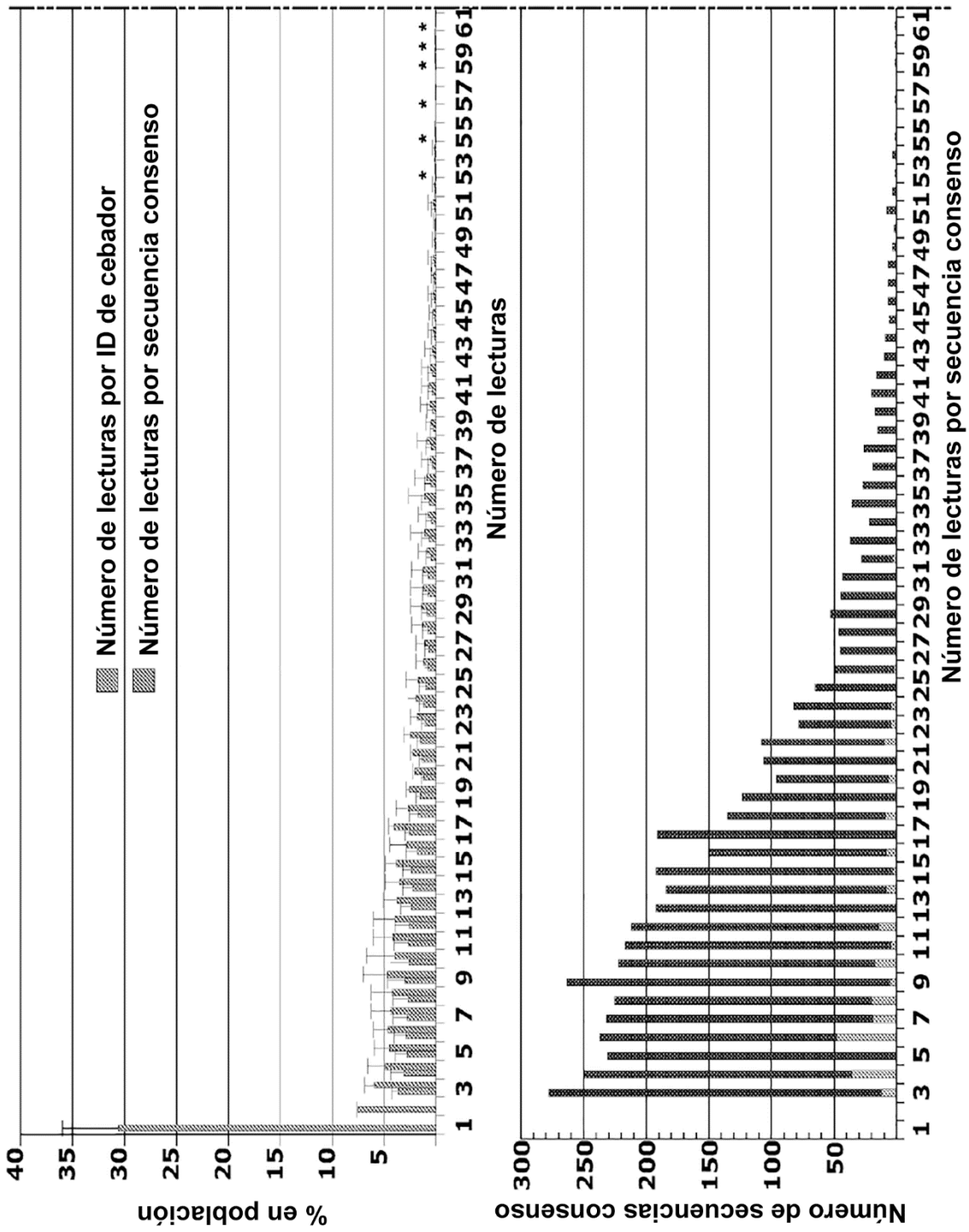

Número de lecturas por ID de cebador
Número de lecturas por secuencia consenso
% en población
Número de lecturas
Número de secuencias consenso
Número de lecturas por secuencia consenso

Fig. 6A

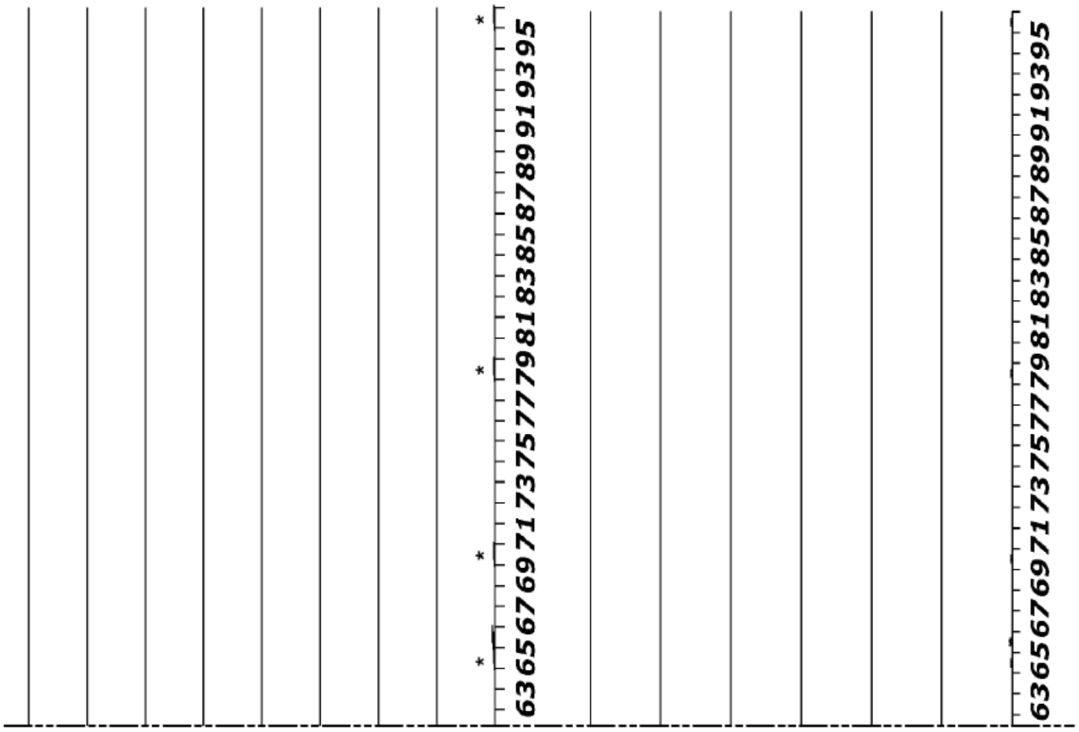

63656769717375777981838587899193 95
63656769717375777981838587899193 95

Fig. 6B

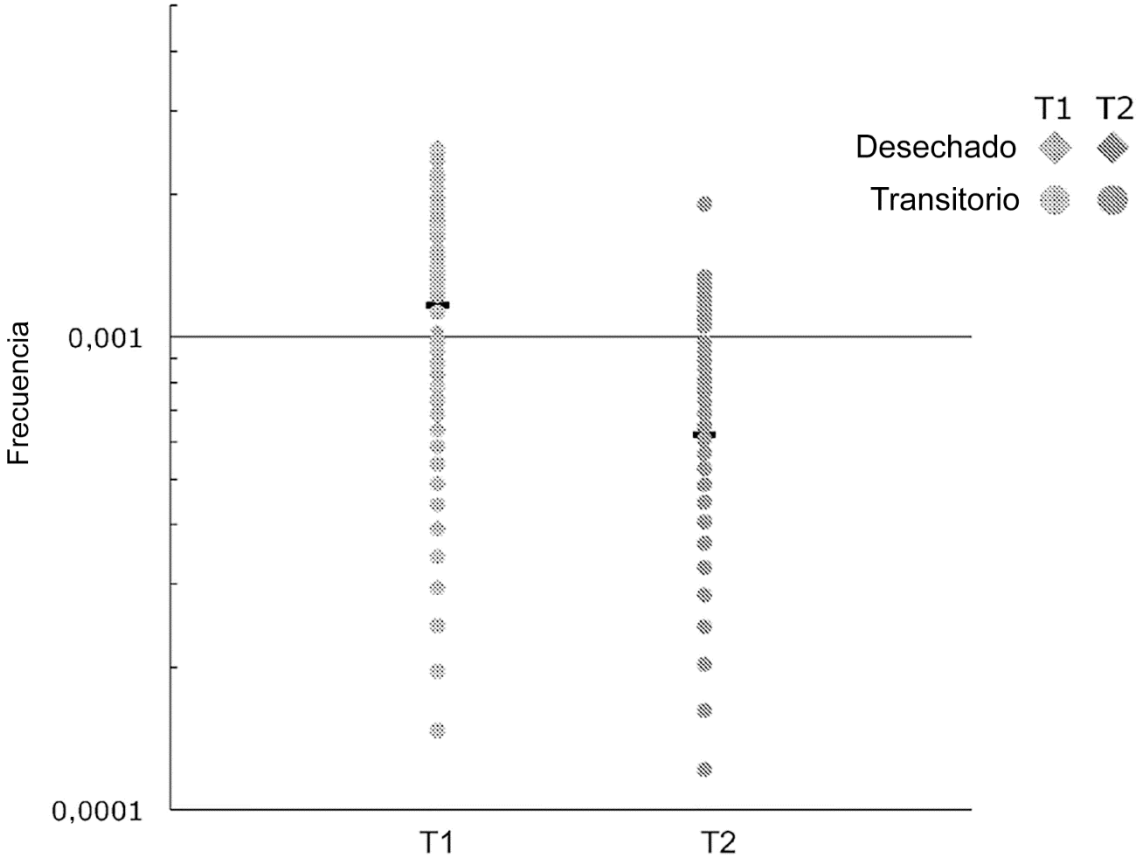


Fig. 7

Fig. 8A

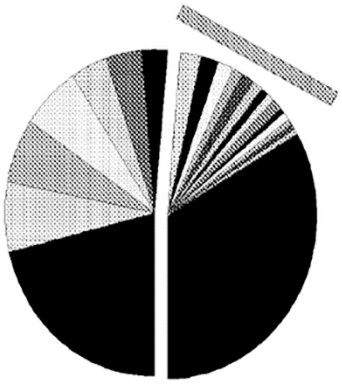


Fig. 8B

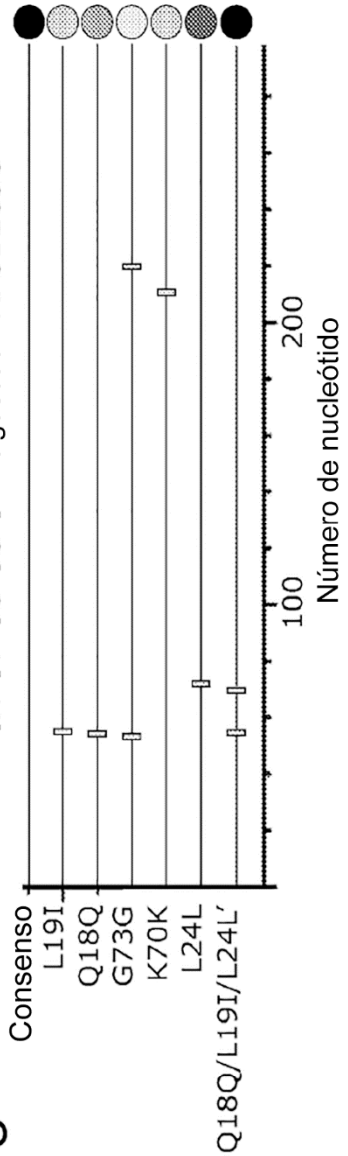
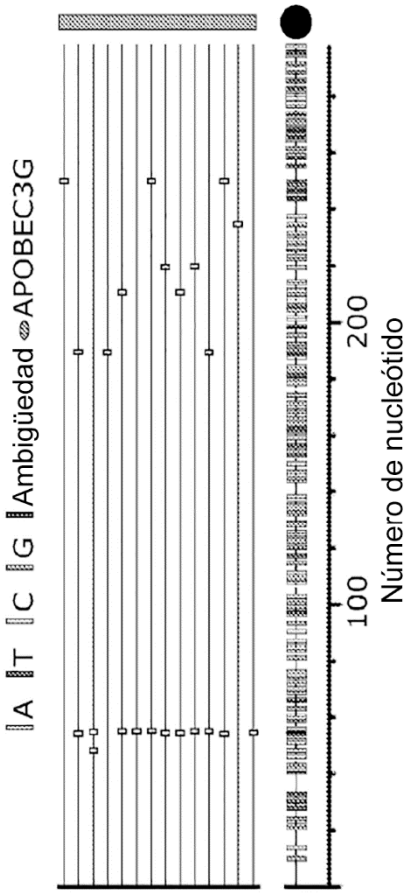
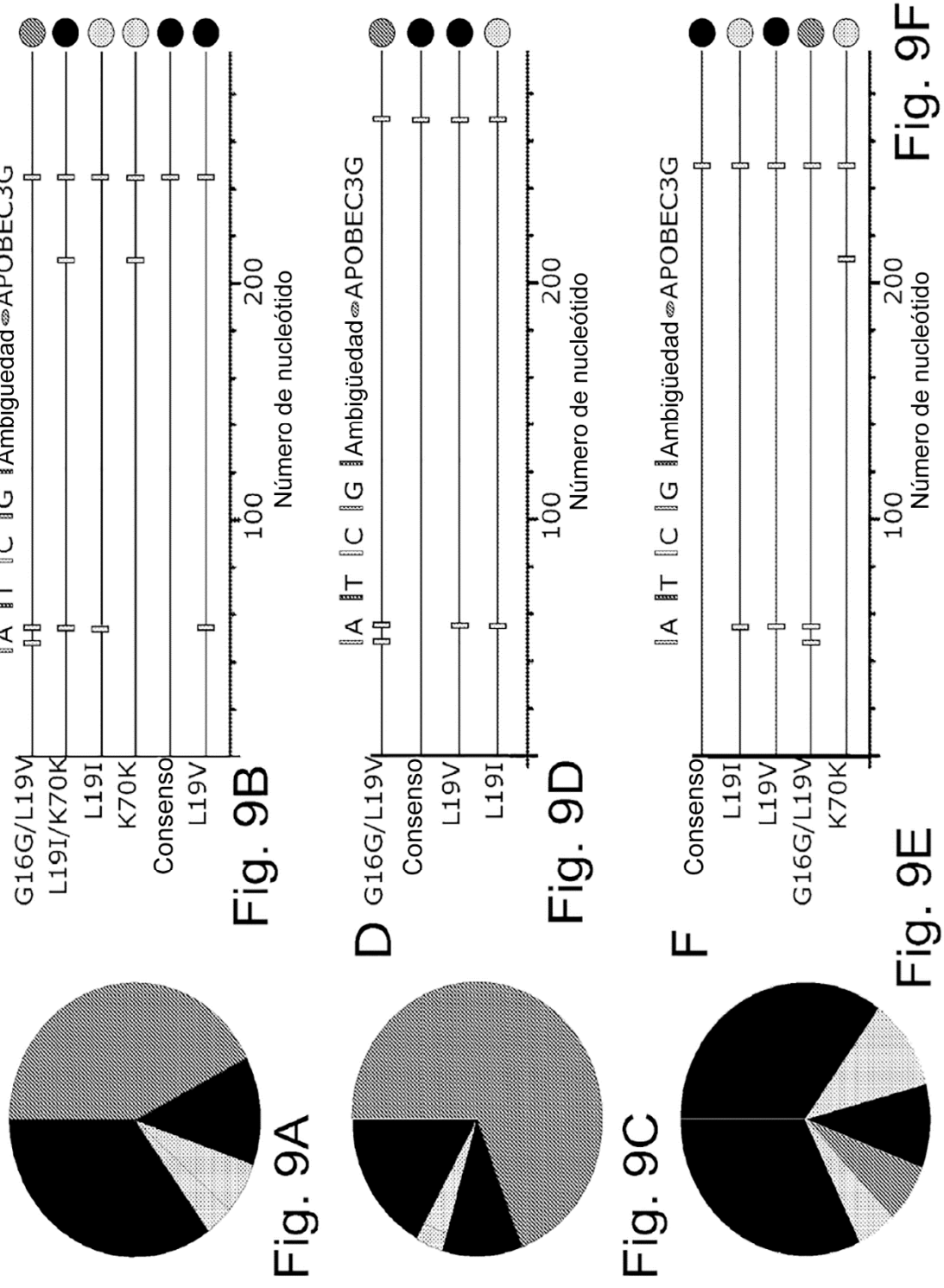


Fig. 8C




A T C G Ambigüedad APOBEC3G
G16G/L19V
L19I/K70K
L19I
K70K
Consenso
L19V
100
200
Número de nucleótido
Fig. 9B
Fig. 9A
D
G16G/L19V
Consenso
L19V
L19I
Fig. 9D
Fig. 9C
F
Consenso
L19I
L19V
G16G/L19V
K70K
Fig. 9E
Fig. 9F

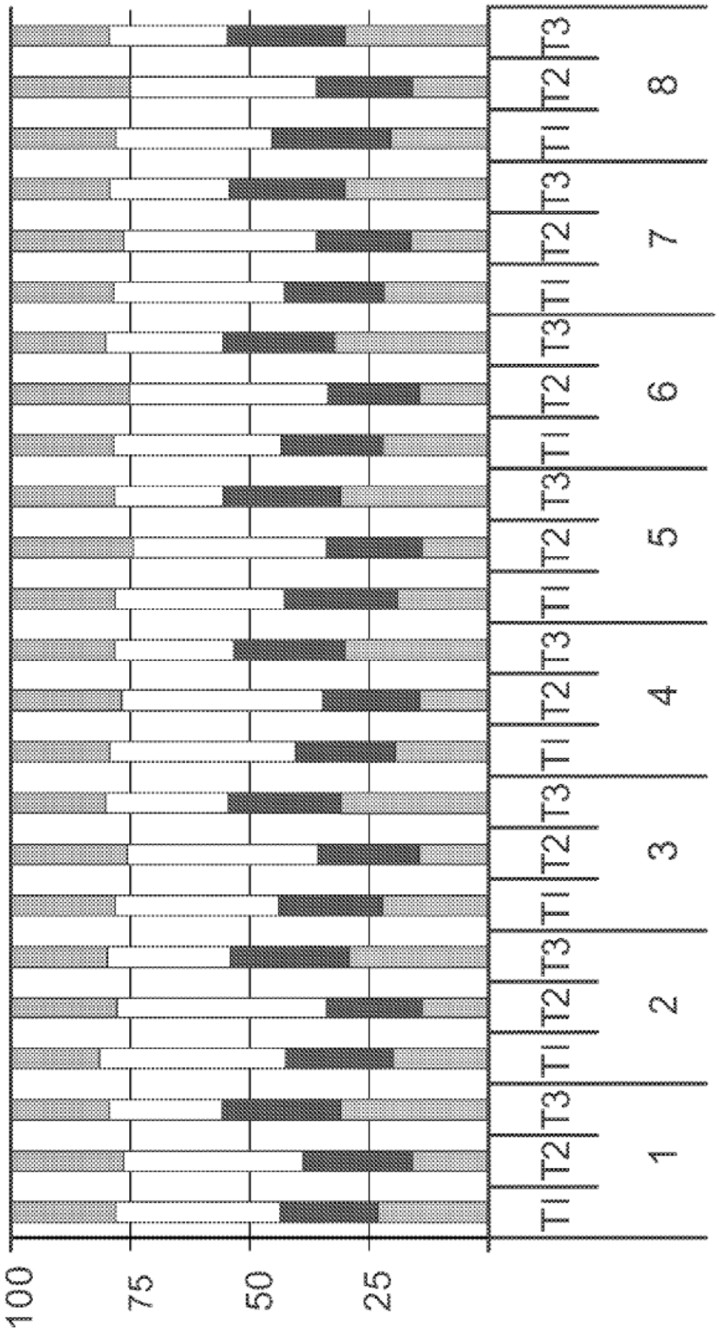

100
75
50
25
T1
T2
T3
1
T1
T2
T3
2
T1
T2
T3
3
T1
T2
T3
4
T1
T2
T3
5
T1
T2
T3
6
T1
T2
T3
7
T1
T2
T3
8

Fig. 10

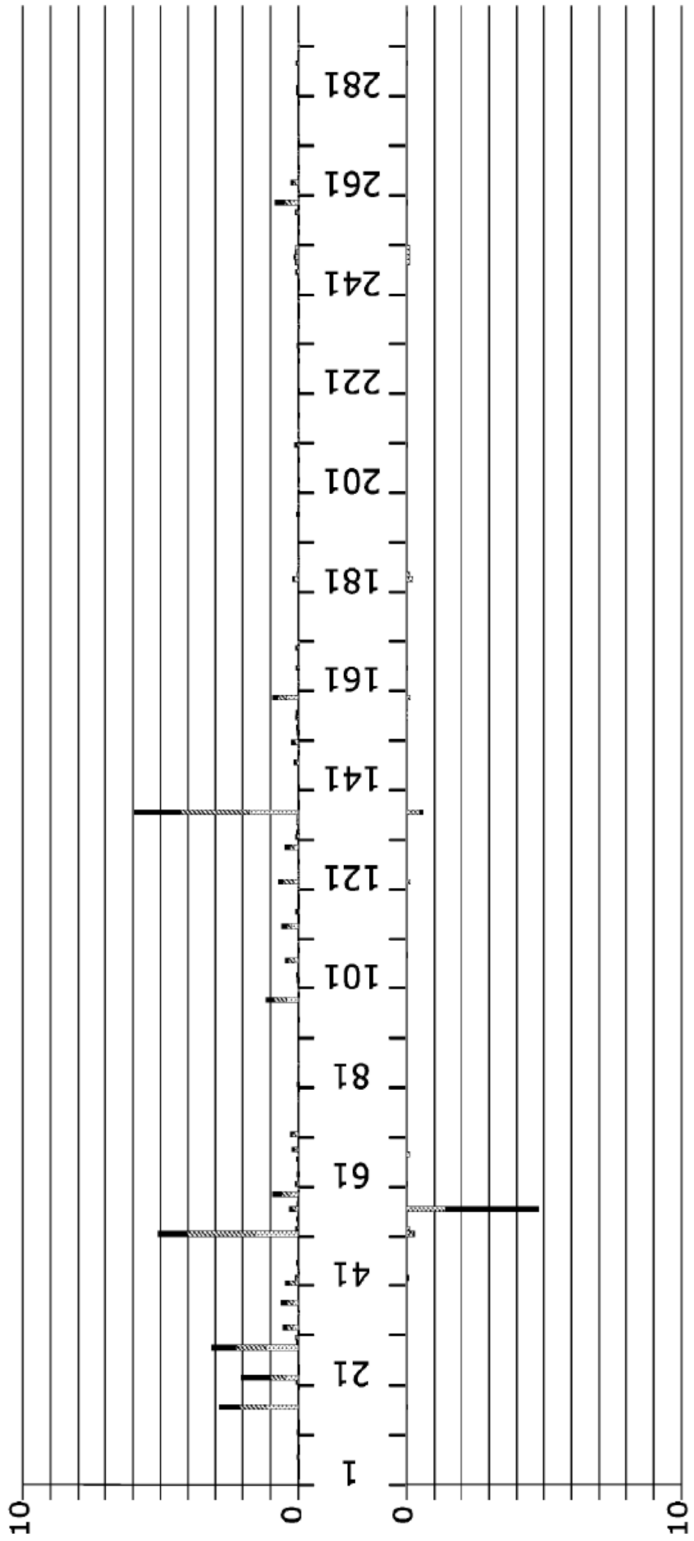


Fig. 11

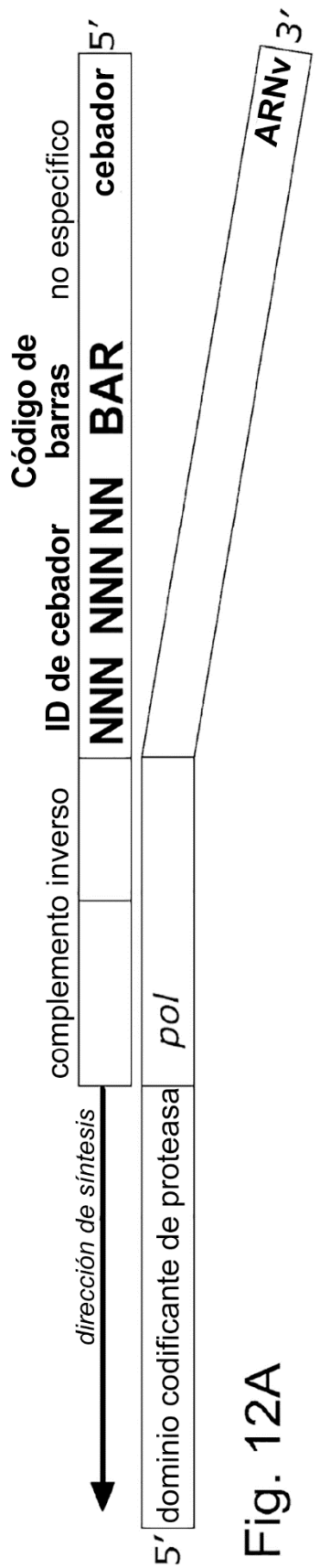


Fig. 12A

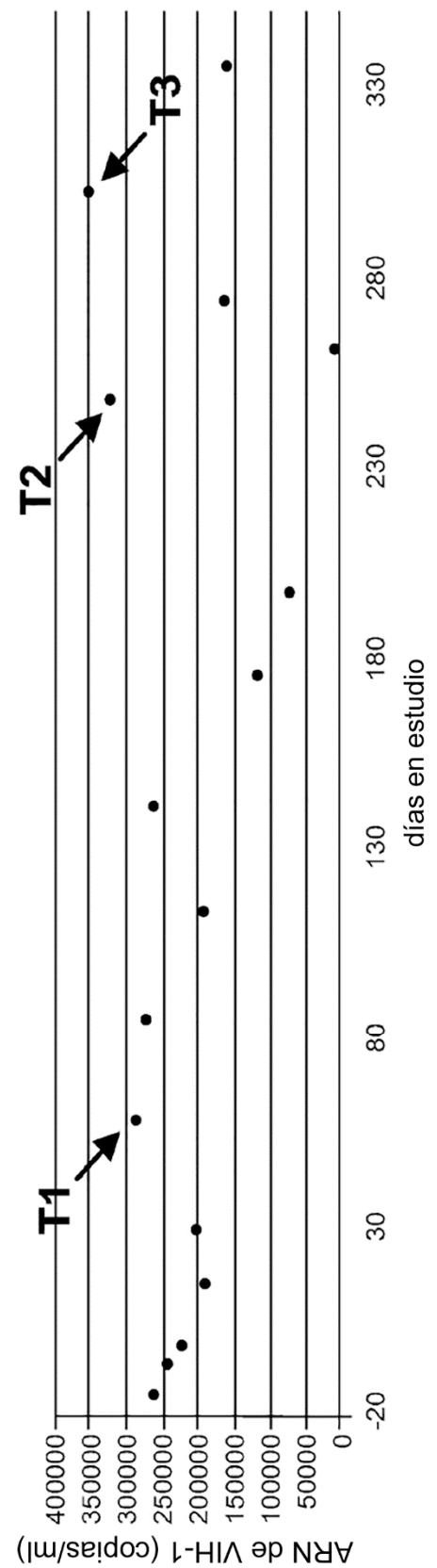


Fig. 12B

Fig. 13A-1

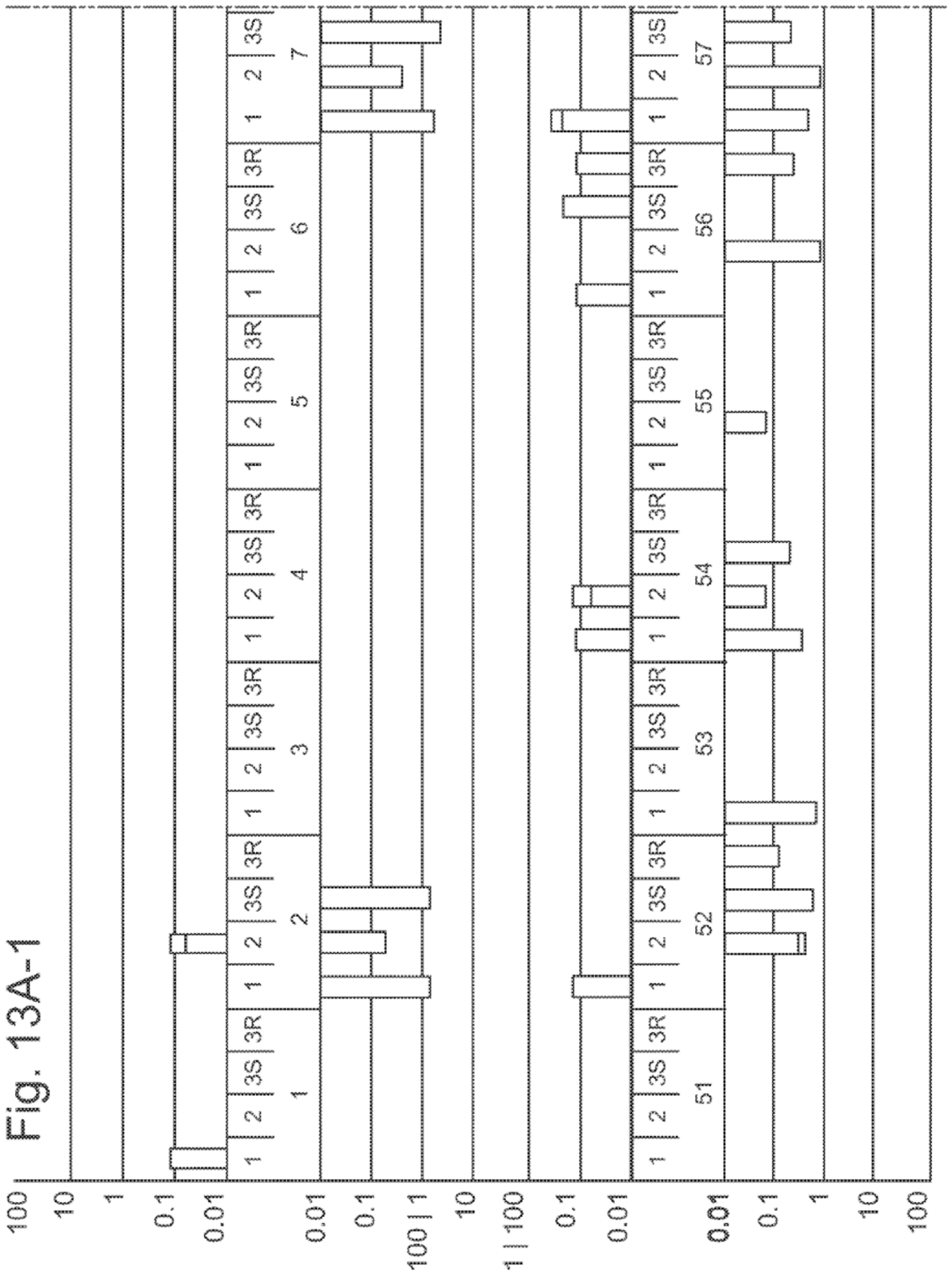
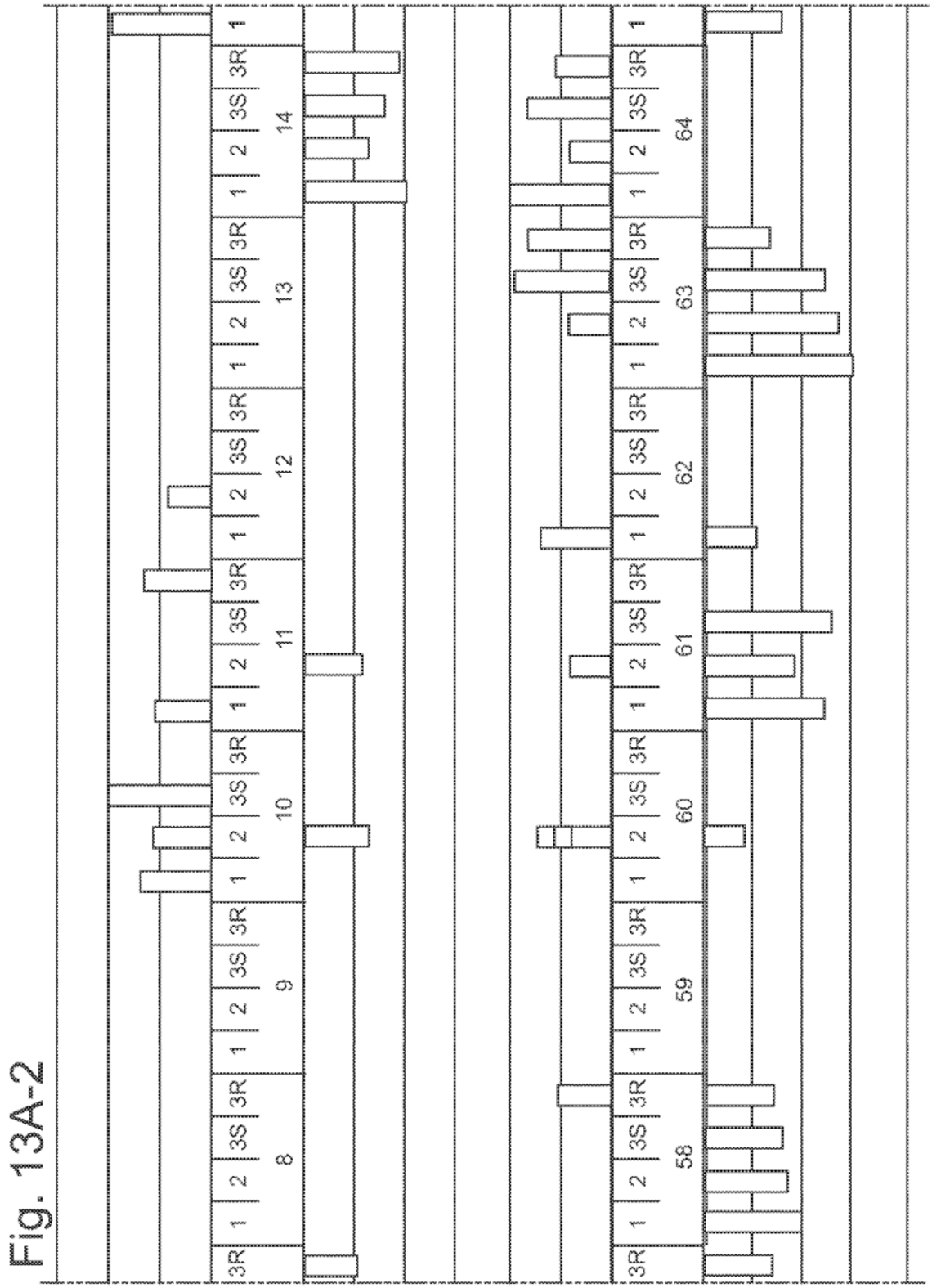

100
10
1
0.1
0.01
1 2 3S 3R 1 2 3S 3R 1 2 3S 3R 1 2 3S 3R 1 2 3S 3R 1 2 3S 3R 1 2 3S
1
2
3
4
5
6
7
0.01
0.1
1
10
100
1 2 3S 3R 1 2 3S 3R 1 2 3S 3R 1 2 3S 3R 1 2 3S 3R 1 2 3S 3R 1 2 3S
51
52
53
54
55
56
57
0.01
0.1
1
10
100

Fig. 13A-2


3R 1 2 3S 8
3R 1 2 3S 9
3R 1 2 3S 10
3R 1 2 3S 11
3R 1 2 3S 12
3R 1 2 3S 13
3R 1 2 3S 14
3R 1 2 3S 58
3R 1 2 3S 59
3R 1 2 3S 60
3R 1 2 3S 61
3R 1 2 3S 62
3R 1 2 3S 63
3R 1 2 3S 64

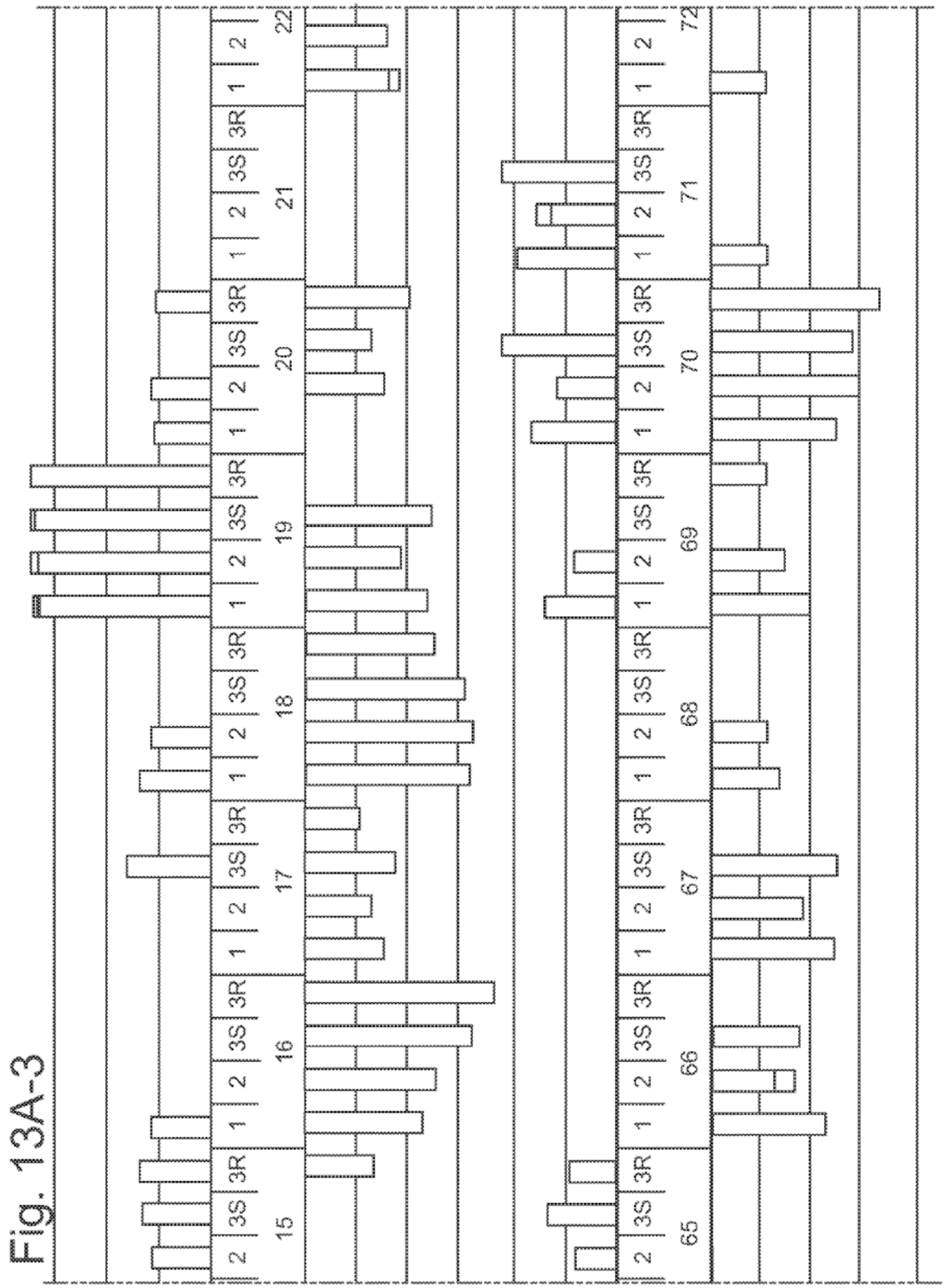


Fig. 13A-3

Fig. 13A-4

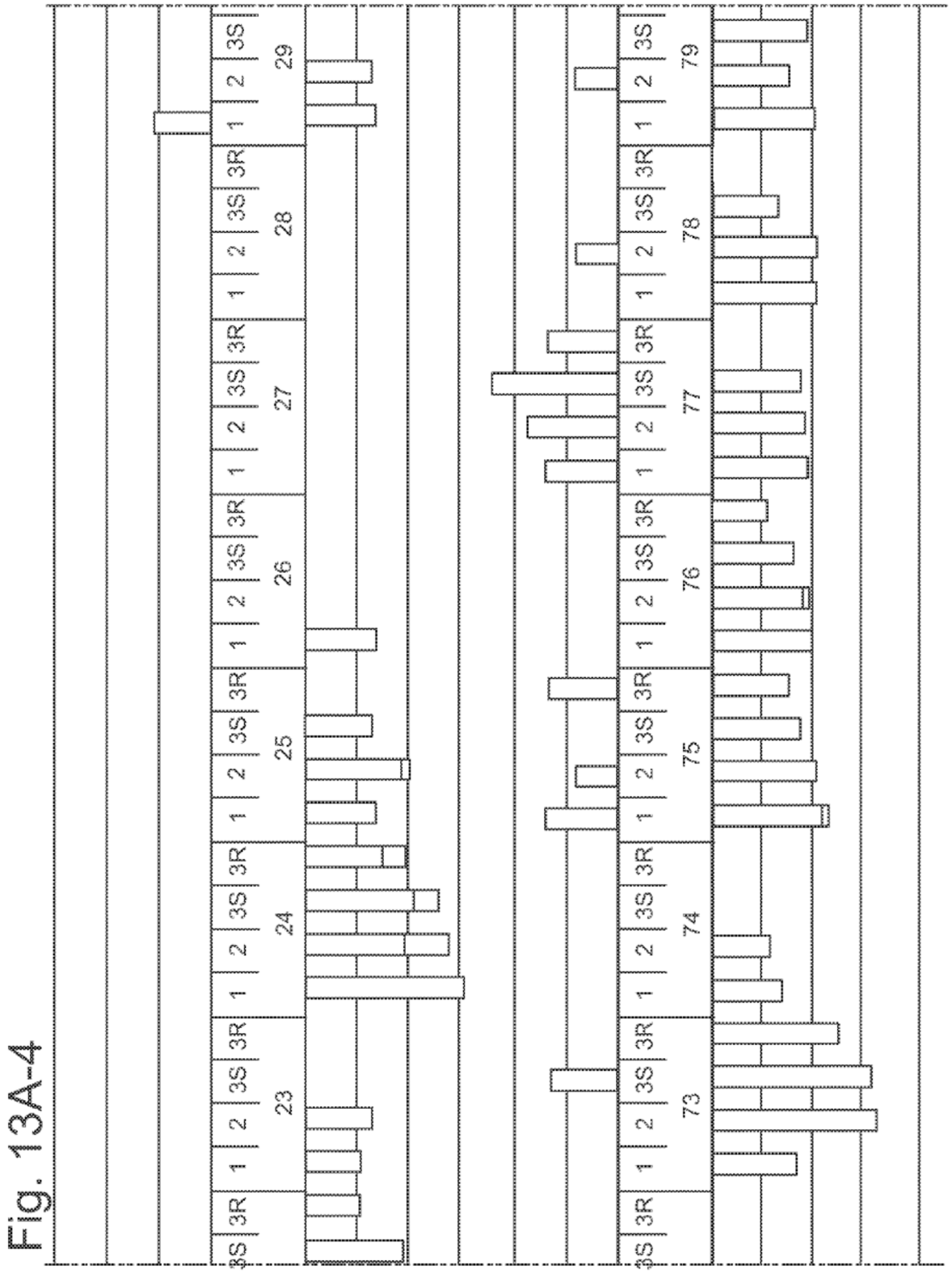

3S
3R
1
2
3S
23
3R
1
2
3S
24
3R
1
2
3S
25
3R
1
2
3S
26
3R
1
2
3S
27
3R
1
2
3S
28
3R
1
2
3S
29
3S
3R
1
2
3S
73
3R
1
2
3S
74
3R
1
2
3S
75
3R
1
2
3S
76
3R
1
2
3S
77
3R
1
2
3S
78
3R
1
2
3S
79

Fig. 13A-5

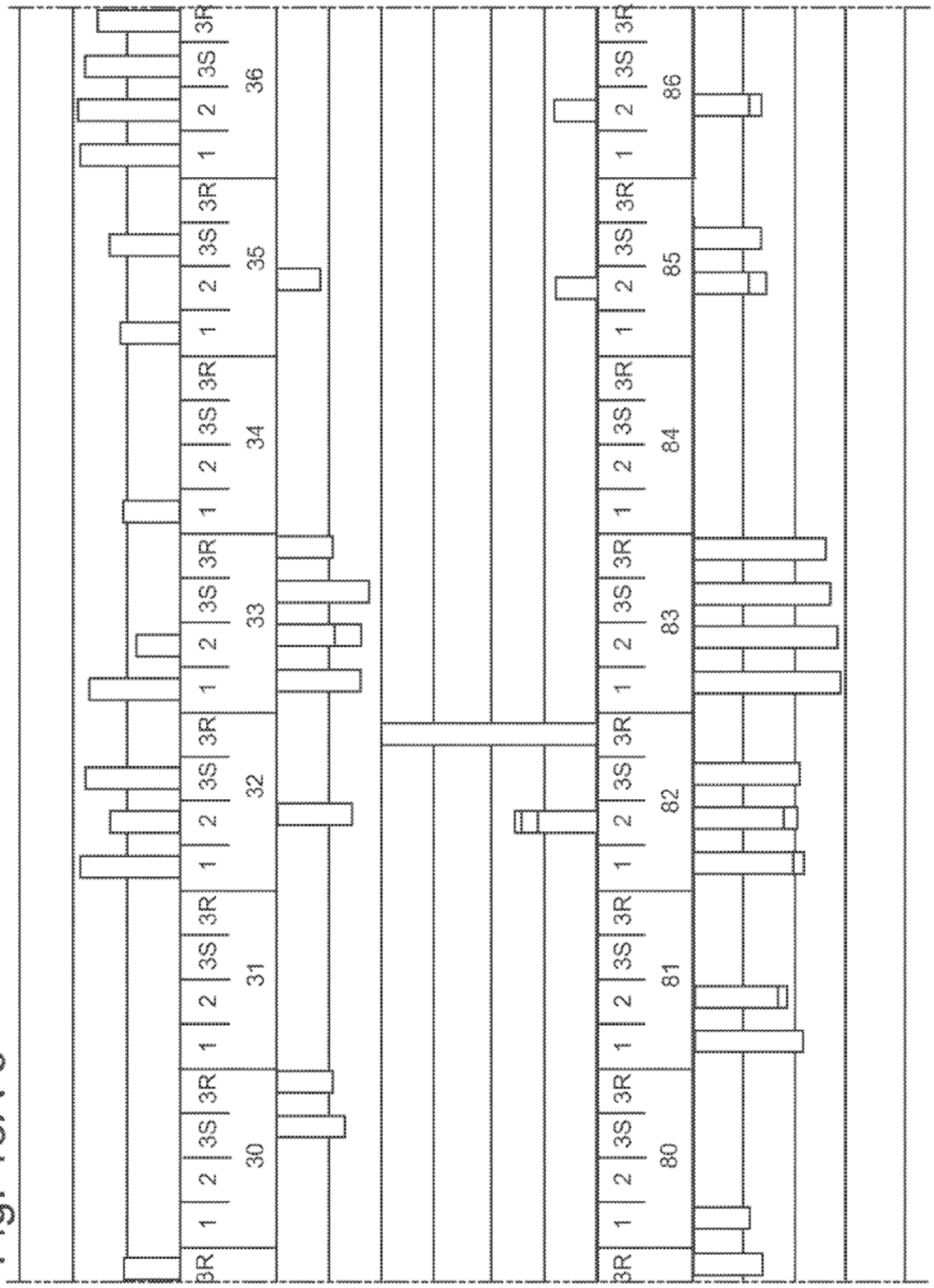

3R
1
2
3S
3R
30
3R
1
2
3S
3R
31
3R
1
2
3S
3R
32
3R
1
2
3S
3R
33
3R
1
2
3S
3R
34
3R
1
2
3S
3R
35
3R
1
2
3S
3R
36
3R
1
2
3S
3R
80
3R
1
2
3S
3R
81
3R
1
2
3S
3R
82
3R
1
2
3S
3R
83
3R
1
2
3S
3R
84
3R
1
2
3S
3R
85
3R
1
2
3S
3R
86

Fig. 13A-6

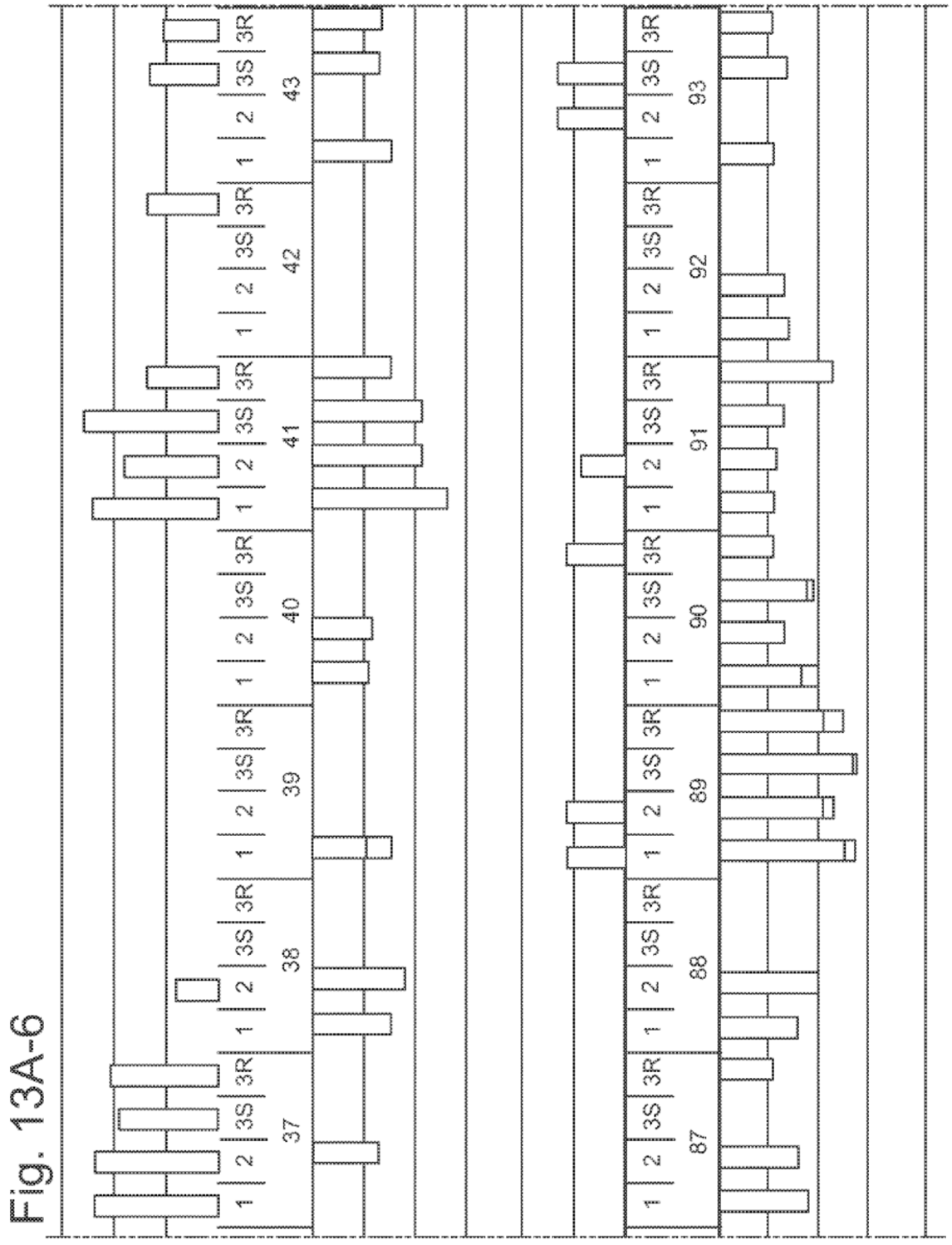
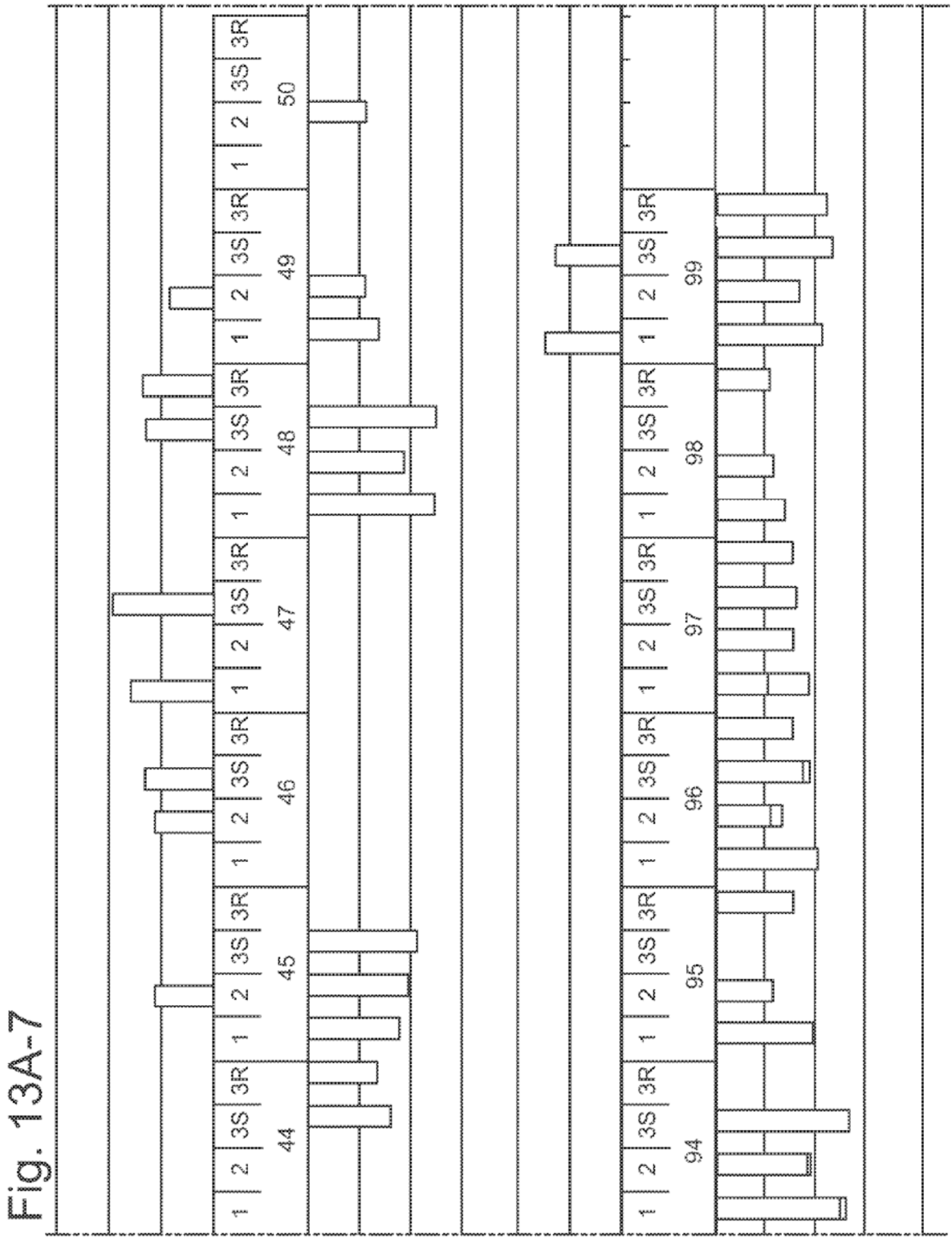
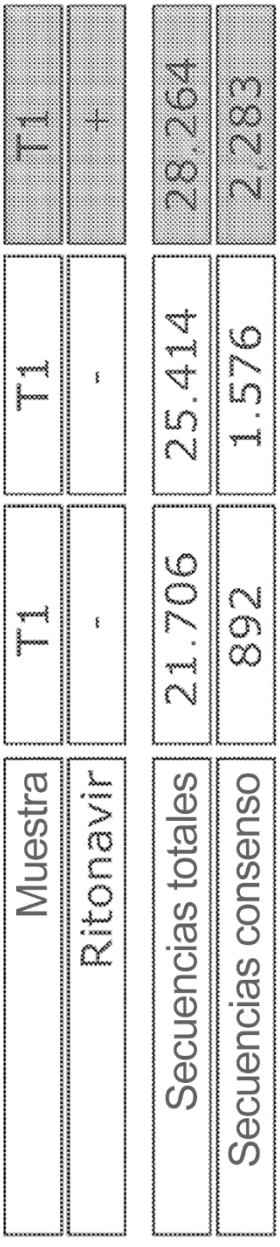


Fig. 13A-7





Muestra	T1	T1	T1
Ritonavir	-	-	+
Secuencias totales	21.706	25.414	28.264
Secuencias consenso	892	1.576	2.283

Fig. 13B

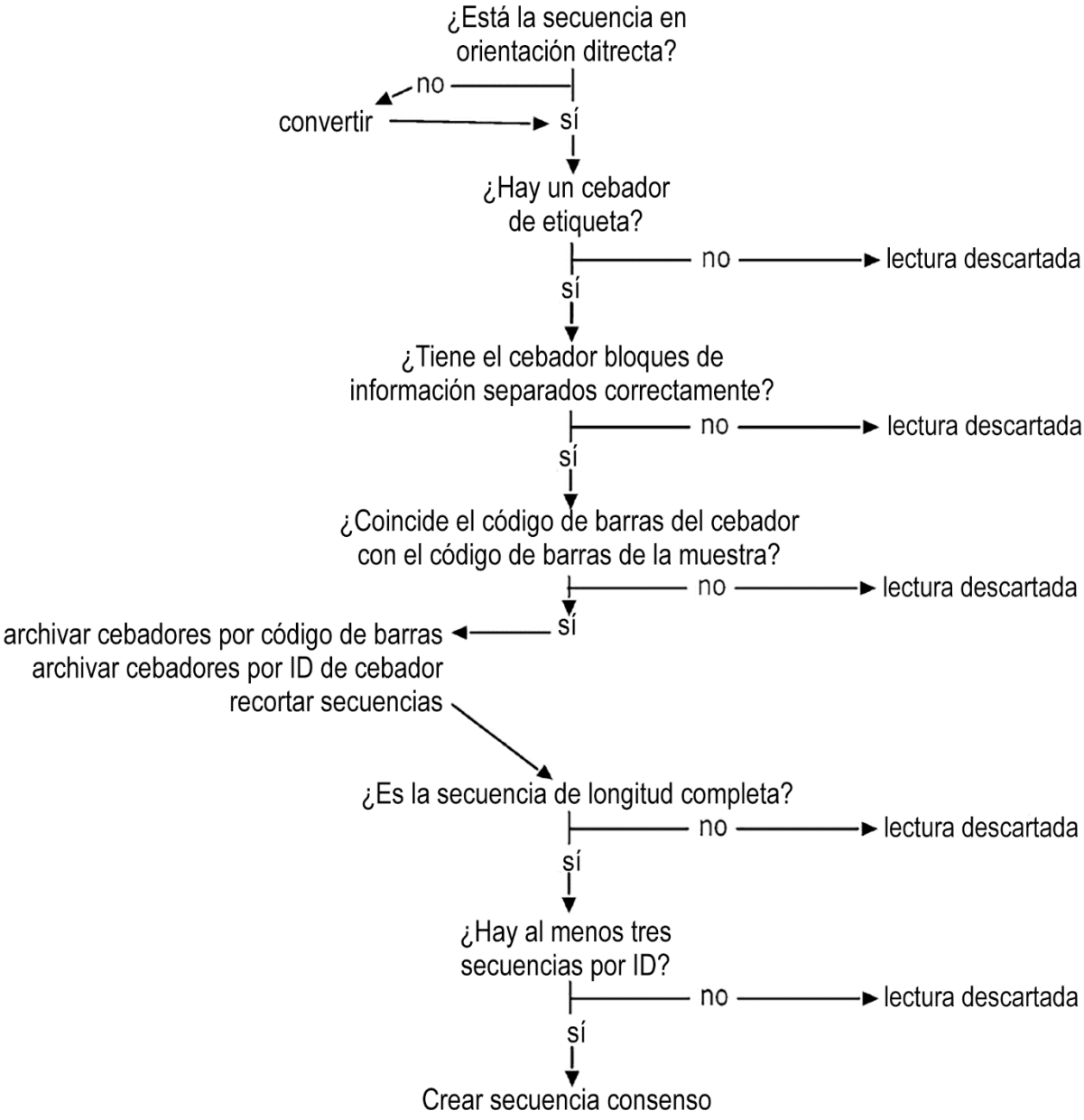

¿Está la secuencia en orientación ditrecta?
no
convertir
sí
¿Hay un cebador de etiqueta?
no
lectura descartada
sí
¿Tiene el cebador bloques de información separados correctamente?
no
lectura descartada
sí
¿Coincide el código de barras del cebador con el código de barras de la muestra?
no
lectura descartada
sí
archivar cebadores por código de barras
archivar cebadores por ID de cebador
recortar secuencias
¿Es la secuencia de longitud completa?
no
lectura descartada
sí
¿Hay al menos tres secuencias por ID?
no
lectura descartada
sí
Crear secuencia consenso

Fig. 14

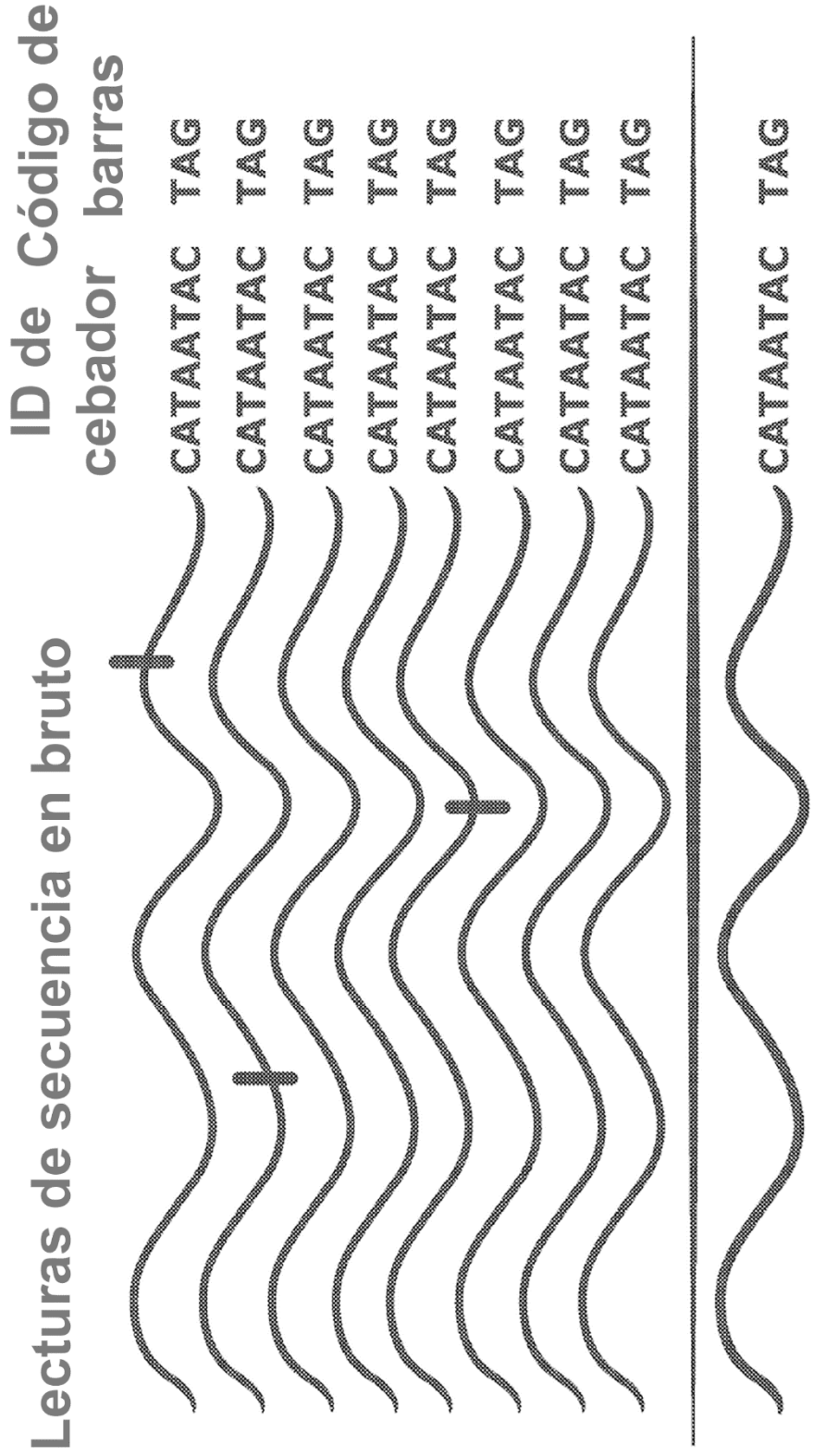

Lecturas de secuencia en bruto
ID de cebador
Código de barras
CATAATAC TAG
CATAATAC TAG
CATAATAC TAG
CATAATAC TAG
CATAATAC TAG
CATAATAC TAG
CATAATAC TAG
CATAATAC TAG
CATAATAC TAG
Secuencia consenso
CATAATAC TAG

Fig. 15

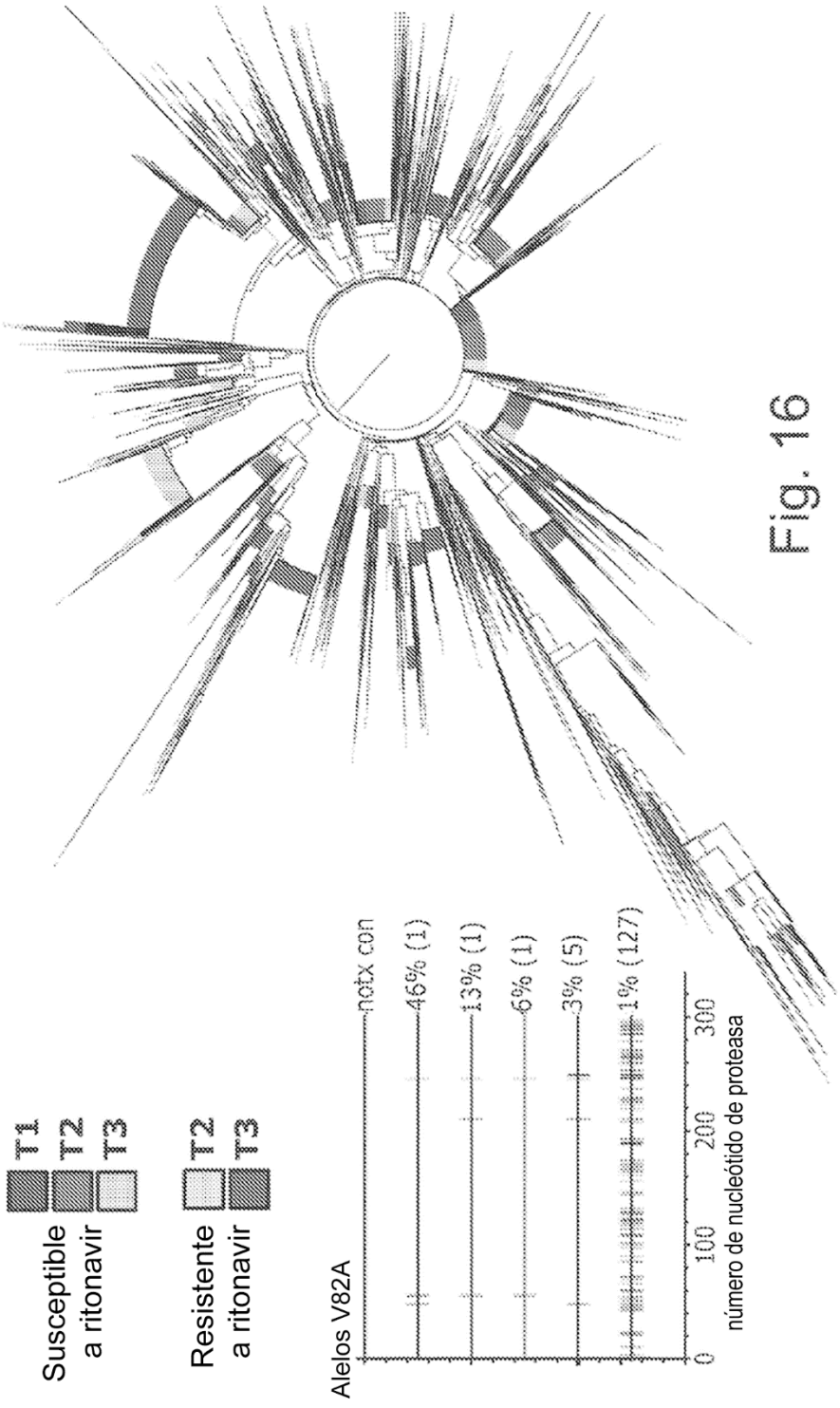

T1
T2
T3
Susceptible a ritonavir
T2
T3
Resistente a ritonavir
Alelos V82A
notx con
46% (1)
13% (1)
6% (1)
3% (5)
1% (127)
0
100
200
300
número de nucleótido de proteasa

Fig. 16