(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 709 212**

(51) Int. CI.:

C12Q 1/68 (2008.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(86) Fecha de presentación y número de la solicitud internacional: **30.06.2014 PCT/US2014/044971**

(87) Fecha y número de publicación internacional: **08.01.2015 WO15002908**

(96) Fecha de presentación y número de la solicitud europea: **30.06.2014 E 14819680 (1)**

(97) Fecha y número de publicación de la concesión europea: **02.01.2019 EP 3017066**

(54) Título: **Análisis biomolecular a gran escala con marcadores de secuencia**

(30) Prioridad:

01.07.2013 US 201361841878 P
21.05.2014 US 201462001580 P

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
15.04.2019

(73) Titular/es:

ADAPTIVE BIOTECHNOLOGIES CORPORATION (100.0%)
1551 Eastlake Avenue East, Suite 200
Seattle, Washington 98102, US

(72) Inventor/es:

ASBURY, THOMAS;
HERVOLD, KIERAN;
KOTWALIWALE, CHITRA;
FAHAM, MALEK;
MOORHEAD, MARTIN;
WENG, LI;
WITTKOP, TOBIAS y
ZHENG, JIANBIAO

(74) Agente/Representante:

UNGRÍA LÓPEZ, Javier

DESCRIPCIÓN

Análisis biomolecular a gran escala con marcadores de secuencia

Antecedentes de la invención

La secuenciación de ADN a gran escala en aplicaciones diagnósticas y pronósticas se ha expandido rápidamente ya que su velocidad y conveniencia ha aumentado y su coste por base ha disminuido, por ejemplo, Ding y col., Nature, 481(7382): 506-510 (2012); Chiu y col., Brit. Med. J., 342: c7401 (2011); Ku y col., Annals of Neurology, 71(1): 5-14 (2012); y similares. En particular, perfiles de ácidos nucleicos que codifican moléculas inmunitarias, tales como receptores de linfocitos T o linfocitos B, o sus componentes, contienen una rica información sobre el estado de salud o enfermedad de un organismo, de modo que se ha propuesto el uso de tales perfiles como indicadores diagnósticos o pronósticas para una amplia variedad de afecciones, por ejemplo, Faham y Willis, patentes de Estados Unidos 8.236.503 y 8,628927; Freeman y col., Genome Research, 19: 1817-1824 (2009); Han y col, J. Immunol., 182 (1001):42.6 (2009); Boyd y col, Sci. Transl. Med., 1(12): 12ra23 (2009); He y col, Oncotarget (March 8, 2011).

Por ejemplo, pacientes tratados para muchos cánceres a menudo retienen una enfermedad mínima residual (EMR) relacionada con el cáncer. Es decir, aunque un paciente pueda tener mediante medidas clínicas una remisión completa de la enfermedad en respuesta al tratamiento, puede permanecer una pequeña fracción de las células cancerígenas que, por una razón u otra, haya escapado de su destrucción. El tipo y tamaño de esta población residual es un factor pronóstico importante para el tratamiento continuado del paciente, por ejemplo, Campana. Hematol. Oncol. Clin. North Am., 23(5): 1083-1098 (2009); Buccisano y col., Blood, 119(2): 332-341 (2012). Por consiguiente, se han desarrollado varias técnicas para evaluar esta población, incluidas técnicas basados en citometría de flujo, hibridación *in situ*, citogenética, ampliación de marcadores de ácidos nucleicos y similares, por ejemplo, Buccisano y col., Current Opinion in Oncology, 21: 582-588 (2009); van Dongen y col., Leukemia, 17(12): 2257-2317 (2003); y similares. La amplificación de ácidos nucleicos recombinantes que codifican segmentes de receptores inmunes (es decir, clonotipos) a partir de linfocitos T y/o linfocitos B han sido particularmente útiles en la evaluación de EMR en leucemias y linfomas, puesto que tales clonotipos tienen normalmente secuencias únicas que pueden servir como marcadores moleculares para sus células cancerígenas asociadas. Tales mediciones se realizan normalmente amplificantes y secuenciando ácidos nucleicos que codifican una única cadena receptora, en parte, debido a que tales amplificaciones están altamente multiplexadas y son complicadas de desarrollar. Según aumenta la escala de multiplexación, se encuentran varios problemas, incluida la probabilidad aumentada de amplificaciones engañosas debido a mis-hibridaciones, formación de cebador-dímero, tasas variables de amplificación que llevan a una representación de secuenciación sesgada y similares, por ejemplo, Elnifro y col., linical Microbiology Reviews, 13(4): 559-570 (2000). Además, la similitud de las secuencias diana y la incorporación de los marcadores de secuencia en secuencias amplificadas, o bien para el análisis secuencial, rastreo de muestra, detección de contaminación o similares, pueden empeorar las anteriores complicaciones asociadas con las amplificaciones a gran escala. Estos restos han evitado el desarrollo de amplificaciones de una reacción a gran escala de múltiples cadenas de receptores inmunitarios, que podrían ser altamente beneficioses para reducir la cantidad de ensayos por separado requeridos para medir secuencias de ácidos nucleicos correlacionadas con una enfermedad mínima.

El documento WO2013/155119 describe la detección y cuantificación de la contaminación de muestras en análisis de repertorio inmunitario.

A la vista de lo anterior, resultaría altamente ventajoso si hubiera disponibles métodos más eficaces para evaluar ácidos nucleicos seleccionados en una única reacción, tales como exones de genes cancerosos o clonotipos que codifican conjuntos de cadenas de receptores inmunitarios.

Sumario de la invención

La presente invención proporciona un método de determinación de un perfil de clonotipo de ácidos nucleicos recombinados que codifican una pluralidad de cadenas de receptor inmunitario en una muestra, comprendiendo el método las etapas de:

(a) unir marcadores de secuencia a moléculas de ácido nucleico recombinadas de genes de receptores de linfocitos T o genes de inmunoglobulina a partir de una muestra de un individuo que comprende linfocitos T y/o linfocitos B y/o ADN sin células para formar conjugados de marcador-ácido nucleico, en donde al menos un ácido nucleico recombinado o copias del mismo tienen distintos marcadores de secuencia unidos;
(b) amplificar los conjugados de marcador-ácido nucleico;
(c) secuenciar una muestra de los conjugados de marcador-ácido nucleico para proporcionar lecturas de secuencias teniendo cada una, una tasa de error y comprendiendo cada una, una secuencia de marcados y una secuencia de ácido nucleico recombinado;
(g) alinear las lecturas de secuencias como secuencias de marcadores para formar grupos de lecturas de secuencia que tienen los mismos marcadores de secuencia;
(e) fusionar lecturas de secuencia de grupos para determinar clonotipos, en donde los grupos de las lecturas de

secuencias de fusionan en distintas secuencias de ácidos nucleicos recombinados siempre que dichos grupos de lecturas de secuencias sean distintos con una probabilidad de al menos el noventa y cinco por ciento; y
(f) determinar el perfil de clonotipo de la muestra determinando los niveles de los clonotipos;

y en donde las etapas de unión y amplificación comprenden:

(i) combinar en una mezcla de reacción con condiciones de extensión de cebador un primer conjunto de cebadores con la muestra, en donde cada cebador del primer conjunto tiene una porción específica a receptor de modo que la porción específica a receptor se hibrida a un ácido nucleico recombinado distinta en un emplazamiento predeterminado y se extiende para formar un primer producto de extensión, y en donde cada cebador del primer conjunto tiene un extremo 5' no complementario que contiene un primer sitio de unión y un marcador de secuencia dispuestos entre la porción específica a receptor y el primer sitio de unión del cebador;
(ii) retirar de la mezcla de reacción cebadores no extendidos del primer conjunto;
(iii) añadir a la mezcla de reacción en condiciones de extensión de cebador un segundo conjunto de cebadores, en donde cada cebador del segundo conjunto tiene una porción específica a receptor de modo que la porción específica a receptor se hibrida al primer producto de extensión en un emplazamiento predeterminado y tiene un extremo 5' no complementario que contiene un segundo sitio de unión de cebador, y en donde cada cebador del segundo conjunto se extiende para formar un segundo producto de extensión, de modo que cada segundo producto de extensión comprende un primer sitio de unión de cebador, un segundo sitio de unión de cebador, al menos un marcador de secuencia y ácido nucleico recombinante que codifica una porción de una cadena de receptor de linfocitos T o una cadena de receptor de linfocitos B; y
(iv) llevar a cabo una reacción en cadena de la polimerasa en la mezcla de reacción para formar un amplicón, usando la reacción en cadena de la polimerasa cebadores directos específicos para el primer sitio de unión de cebador y cebadores indirectos específicos para el segundo sitió de unión de cebador.

La presente invención se dirige a métodos de amplificación a gran escala en una única reacción, en particular, mediante reacción en cadena de la polimerasa (PCR), de una población de polinucleótidos diana, tales como ácidos nucleicos recombinados que codifican cadenas de receptor inmunitario, seguidos por su identificación usando secuenciación de ADN a gran escala. Se describe en el presente documente la aplicación de los métodos anteriores para controlar la enfermedad mínima residual de un cáncer. La invención se ejemplifica en una cantidad de implementaciones y aplicaciones, algunas de las cuales se resumen a continuación y por toda la memoria descriptiva.

En el presente documento se describen métodos de generación de perfiles de ácidos nucleicos que codifican una población de biomoléculas de interés, tales como moléculas de receptor inmunitario. En un caso, los métodos comprenden unir marcadores de secuencia a una población seleccionada de ácidos nucleicos en una muestra para formar conjugados de marcador-ácido nucleico, amplificar los conjugados de marcador-ácido nucleico y secuenciar conjugados de marcador-ácido nucleico amplificados para proporcionar lecturas de secuencia comprendiendo cada una tanto una secuencia de marcador como una secuencia de ácido nucleico, para el cual se genera un perfil de los ácidos nucleicos. En algunas realizaciones, la unión de marcadores de secuencia se permite por una o más etapas sucesivas de extensión de cebador y retiración de cebador, después de la cual el producto resultante puede amplificarse adicionalmente sin sesgo mediante cebadores directos e indirectos comunes.

Se describen en el presente documento métodos para detectar y medir la contaminación, tal como contaminación de arrastre, en una muestra a partir de un material que proviene de una muestra distinta. En un caso, tal método para detectar la contaminación en un individuo que está siendo controlado para una enfermedad mínima residual puede comprender la siguientes etapas: (a) obtener de un individuo una muestra de tejido; (b) unir marcadores de secuencia a moléculas de gen cancerosas o ácidos nucleicos recombinados para formar conjugados de marcador-ácido nucleico, en donde al menos un ácido nucleico o copias del mismo tiene distintos marcadores de secuencia unidos y en donde las moléculas de gen cancerosas son características de un cáncer del individuo; (c) amplificar los conjugados de marcador-ácido nucleico; (d) secuenciar una muestra de los conjugados de marcador-ácido nucleico para proporcionar lecturas de secuencia que tienen tasas de error y que comprenden una secuencia de marcador y una secuencia de gen canceroso o secuencia de ácido nucleico recombinado; (e) comparar secuencias de marcador con respecto a secuencias de marcador determinadas por separado de otras muestras de tejido; y (f) determinar la presencia, ausencia y/o nivel de contaminación mediante la identidad de una o más secuencias de marcador con cualquier secuencia de marcador determinada por separado de otras muestras de tejido.

En otro aspecto, la invención se dirige a un método tal como se ha descrito anteriormente para generar perfiles de clonotipo basados en al menos dos cadenas de un receptor de linfocitos B, cuyo método comprende amplificar en una única reacción ácidos nucleicos diana que codifican dos o más cadenas de un receptor de linfocitos B. En otro aspecto, tales métodos se emplean para controlar enfermedad mínima residual en un cáncer de linfocitos B.

En otro aspecto, la invención se dirige a un método tal como se ha descrito anteriormente para generar perfiles de clonotipo basados en al menos dos cadenas de un receptor de linfocitos T, cuyo método comprende amplificar en una única reacción ácidos nucleicos diana que codifican dos o más cadenas de un receptor de linfocitos T. En otro aspecto, tales métodos se emplean para controlar enfermedad mínima residual en un cáncer de linfocitos T.

Estos aspectos anteriormente caracterizados, así como otros aspectos, de la presente invención se ejemplifican en una cantidad de implementaciones y aplicaciones ilustradas, algunas de las cuales se muestran en las figuras y se caracterizan en la sección de las reivindicaciones que sigue.

Breve descripción de los dibujos

Los rasgos novedosos de la invención se indican a continuación en particular, en las reivindicaciones adjuntas. Se obtiene una mejor comprensión de los rasgos y ventajas de la presente invención haciendo referencia a la siguiente descripción detallada que indica a continuación las realizaciones ilustrativas, en las que se utilizan los principios de la invención y los dibujos adjuntos de los cuales:

Las Fig. 1A a 1C ilustran diagramáticamente diversas realizaciones de la invención. La Fig. 1D ilustra un método de generación (con o sin marcadores de secuencia) moldes de ácidos nucleicos recombinados que tienen una longitud predeterminada.
Las Fig. 2A a 2G ilustran diversos métodos para unir marcadores de secuencia únicos a sustancialmente cada secuencia diana en una muestra.
Las Fig. 3A y 3B ilustran diagramáticamente un aspecto de la invención para generar perfiles de clonotipo a partir de secuencias de ácidos nucleicos que codifican cadenas lgH.
La Fig. 4A ilustra el uso de marcadores de secuencia para determinar secuencias de clonotipo a partir de lecturas de secuencia. La Fig. 4B ilustra el uso de marcadores de secuencia en realizaciones donde se unen múltiples marcadores de secuencia distintos al mismo polinucleótido diana o copias del mismo.
La Fig. 5A ilustra conceptos de clonotipos en espacio y distancias de secuencia entre clonotipos estrechamente relacionados. La Fig. 5B es un diagrama de flujo que ilustra una realización de un método para distinguir clonotipos genuinamente distintos de clonotipos que difieren únicamente por errores de secuenciación (que deben fusionarse).
La Fig. 5C ilustra la forma de una función numérica usada en una realización para determinar si fusionar o no clonotipos relacionados.
Las Fig. 5D y 5E ilustran el uso de árboles de secuencia en un método de fusión de lecturas de secuencia.

Descripción detallada de la invención

La práctica de la presente invención puede emplear, salvo que se indique lo contrario, técnicas convencionales y descripciones de biología molecular (incluidas técnicas recombinantes), bioinformática, biología celular y bioquímica, que se encuentran en el conocimiento de un técnico de la materia. Tales técnicas convencionales incluyen, pero sin limitación, muestreo y análisis de glóbulos rojos, secuenciación y análisis de ácido nucleico y similares. Ilustraciones específicas de la técnica adecuada pueden obtenerse por referencia al ejemplo en el presente documento a continuación. Sin embargo, otros procedimientos convencionales equivalentes pueden, por supuesto, también usarse. Tales técnicas convencionales y descripciones se pueden encontrar en manuales de laboratorio estándar tales como Genome Analysis: A Laboratory Manual Series (Vols. I-IV); PCR Primer: A Laboratory Manual; *and* Molecular Cloning: A Laboratory Manual (all from Cold Spring Harbor Laboratory Press) y similares.

En el presente documento se describen métodos de producción de perfiles de clonotipos de múltiples cadenas de receptores inmunitarios mediante amplificación de multiplexación a gran escala de ácidos nucleicos que codifican tales cadenas seguidos por secuenciación de alto rendimiento del producto de amplificación o amplicón. En algunas realizaciones, la invención supera con inconvenientes comunes de la amplificación de multiplexación incluyendo etapas sucesivas de extensión de cebador, retiración de cebadores no extendidos o no incorporados, y adición de nuevos cebadores o bien para su amplificación (por ejemplo, mediante PCR) o para extensiones de cebador adicionales. Tales etapas también permiten el uso de marcadores de secuencia que, de otro modo, contribuirían a amplificaciones no específicas o engañosas. En otro aspecto, los marcadores de secuencia se emplean en realizaciones con aplicaciones clínicas, en particular, análisis de enfermedad mínima residual (EMR), por ejemplo, de ejemplos de un paciente que está siendo tratado de cáncer. Marcadores de secuencia incorporados en lecturas de secuencia proporcionan un medio eficaz para determinar clonotipos y, al mismo tiempo, proporcionar un medio conveniente para detectar la contaminación por arrastre detectando la presencia o ausencia de marcadores de secuencia a partir de ensayos previos, o bien a partir de muestras del mismo paciente o a partir de muestras de un paciente distinto que se sometieron a ensayo en el mismo laboratorio. Son de interés particular métodos para generar perfiles de clonotipos a base de secuencias de ácidos nucleicos recombinados que codifican una pluralidad de cadenas de receptores de linfocitos B (RLB) usando una única reacción de amplificación seguida por secuenciación de siguiente generación de alto rendimiento. También son de interés particular métodos para generar perfiles de clonotipos a base de secuencias de ácidos nucleicos recombinados que codifican una pluralidad de cadenas de receptor de linfocitos T (RLT) usando una única reacción de amplificación seguida por secuenciación de siguiente generación de alto rendimiento. Los métodos de la invención también se pueden aplicar a otros de amplificación y secuenciación a gran escala de otros conjuntos de ácidos nucleicos de interés, que incluyen, por ejemplo, conjuntos de exones de genes cancerosos. En estos aspectos, los marcadores de secuencia permiten tanto el control de contaminación por arrastre y la determinación más sensible de secuencias de nucleótidos de polinucleótidos diana en vista de metodologías de secuenciación proclive a errores. También en estos aspecto, un conjunto de marcadores de secuencia (tal como se describe en más detalle a continuación) es normalmente más

grande que el número de polinucleótidos diana en una muestra y la diferencia secuencial entre marcadores de secuencia unidos a polinucleótidos diana en lo suficientemente grande de modo que eficazmente una secuencia de un marcador no podría transformarse en otro mediante error de secuenciación.

Un ejemplo se ilustra en la Fig. 1A. En una mezcla de reacción, los cebadores (22) de un primer conjunto (cada cebador del primer conjunto teniendo una porción específica a receptor (16) y una porción de extremo 5' no complementaria (15) que comprende un primer sitio de unión de cebador) se hibridan a un extremo de los polinucleótidos diana (10) (después de fusionar el polinucleótido diana (10)) y los cebadores (24) a partir de un segundo conjunto (cada cebador del segundo conjunto teniendo una porción específica a receptor (20) y una porción de extremo 5' no complementaria que comprende un marcador de secuencia (14) y un segundo sitió de unión de cebador (12)) se hibridan a otro extremo de polinucleótidos diana (10). En algunas realizaciones, tal como se señala a continuación, una porción no complementaria (15) de cebador (22) también puede comprender un marcador de secuencia. En alguna circunstancia, dos marcadores de secuencia más cortos pueden ser más ventajosos que un único marcador de secuencia más largo de diversidad equivalente. Por lo tanto, por ejemplo, dos marcadores de secuencia de nucleótidos aleatorios de 8 meros pueden ser menos probable que provoquen una sensibilización engañosa, dímeros de cebador y similares, que un único marcador de secuencia de nucleótidos aleatorios de 16 meros. Los polinucleótidos diana (10) son normalmente ácidos nucleicos somáticamente recombinados de linfocitos T o linfocitos B que codificaban cadenas o porcionas de cadenas de receptores de linfocitos T (RLT) o receptores de linfocitos B (por ejemplo, porciones de cadenas IgH o cadenas IgK). Por lo tanto, en algunas realizaciones, las porciones específicas a receptor de cebadores (22) y (24) pueden ser específicas para secuencias de región V y secuencias de región J, respectivamente o, en otras realizaciones, viceversa.

En algunas realizaciones, los polinucleótidos diana (10) pueden comprender mezclas complejas de ácidos nucleicos cuyos perfiles de secuencia se desean, incluyendo, aunque no de forma limitativa, ácidos nucleicos recombinados que codifican porciones de moléculas de receptor inmunitario, ADNr 16S de comunidades microbianas, amplificaciones metagenómicas de genes que codifican proteínas de importancia industrial o medicinal (tales como, enzimas), genes de seres humanos o animales y/o exones relacionados con enfermedades específicas, tales como cáncer, enfermedades infecciosas o similares. En realizaciones que se refieren a ácidos nucleicos recombinados que codifican receptores inmunitarios, normalmente al menos porciones de una región V, D o J están presentes entre los dos emplazamientos de unión del primer y segundo conjunto de cebadores. En algunas realizaciones, entre los dos emplazamientos de unión del primer y segundo conjunto de cebadores hay al menos una porción de una reorganización VDJ de IgH, una reorganización DJ de IgH, una reorganización VJ de IgK, una reorganización VJ de IgL, una reorganización VDJ de RLT β, una reorganización DJ de RLT β, una reorganización VJ de RLT α, una reorganización VJ de RLT γ, una reorganización VDJ de RLT δ, o una reorganización VD de RLT δ. En algunas realizaciones, entre los dos emplazamientos de unión del primer y segundo conjunto de cebadores hay al menos una porción de una reorganización VDJ de IgH, una reorganización DJ de IgH, una reorganización VJ de IgK o una reorganización VJ de IgL. En algunas realizaciones, entre los dos emplazamientos de unión del primer y segundo conjunto de cebadores hay al menos una porción de una reorganización VDJ de RLT β, una reorganización DJ de RLT β, una reorganización VJ de RLT α, una reorganización VJ de RLT γ, una reorganización VDJ de RLT δ, o una reorganización VD de RLT δ. En todavía otras realizaciones, entre los dos emplazamientos de unión del primer y segundo conjunto de cebadores hay al menos una porción de una reorganización VDJ de IgH, una reorganización DJ de IgH y una reorganización VJ de IgK. Y en otras realizaciones, entre los dos emplazamientos de unión del primer y segundo conjunto de cebadores hay al menos una porción de una reorganización VDJ de RLT β, una reorganización VJ de RLT γ, y una reorganización VDJ de RLT δ o una reorganización VD de RLT δ. En algunas realizaciones, al menos una porción de una reorganización VDJ comprende la porción de D o NDN completa y partes de los segmentos V y J suficientes para su identificación. En algunas realizaciones, al menos una porción de una reorganización VDJ comprende al menos un segmento de 50 nucleótidos que comprende la porción de D o NDN completa y partes de los segmentos V y J. En algunas realizaciones, al menos una porción de una reorganización VDJ comprende al menos un segmento de 70 nucleótidos que comprende la porción de D o NDN completa y partes de los segmentos V y J.

En algunas realizaciones, un primer conjunto comprende uno o más cebadores que son cada uno específico para un segmento J o un segmento C. Los cebadores de tal primer conjunto se hibridan a sus secuencias diana y se extienden, después de lo cual se retiran los cebadores no extendidos del primer conjunto. Los cebadores de un segundo conjunto que son cada uno específico para un segmento V se hibridan a sus secuencias diana y se extienden. En otras realizaciones, un primer conjunto comprende cebadores que son cada uno específico para un segmento V y los cebadores de tal primer conjunto se hibridan a sus secuencias diana y se extienden, después de lo cual se retiran los cebadores no extendidos del primer conjunto, los cebadores de un segundo conjunto que son cada uno específico para un segmento J o un segmento C se hibridan a sus secuencias diana y se extienden. En alternativas de ambas de estas realizaciones, el primer y segundo conjunto pueden cada uno contener una pluralidad de cebadores y cada cebador puede ser específico para un segmento de receptor inmunitario distinto.

Volviendo a la Fig. 1A, en algunos casos, los cebadores del primer y segundo conjunto se extienden (5) llevando a cabo en realizaciones alternativas, 1-10, o 2-10, o 3-10, o 4-10, o 5-10 ciclos de fusión, hibridación y extensión, después de lo cual se retiran los cebadores no extendidos de la mezcla de reacción usando técnicas convencionales. En otros casos, los cebadores del primer y segundo conjunto se extienden (5) llevando a cabo en

realizaciones alternativas, 2-5, o 3-5, o 4-5 ciclos de fusión, hibridación y extensión, después de lo cual se retiran los cebadores no extendidos de la mezcla de reacción usando técnicas convencionales. Aún en otro caso, los cebadores del primer y segundo conjunto se extienden llevando a cabo dos ciclos de fusión, hibridación y extensión. Por ejemplo, los cebadores no extendidos se pueden retirar mediante digestión de exonucleasa, hibridación a secuencias complementarias sobre perlas magnéticas, cromatografía de exclusión por tamaño, columnas de centrifugación disponibles en el mercado (por ejemplo, Qiagen QIAquick PCR Purification Kit), o similares. En una realización, los cebadores no extendidos o no incorporados, se retiran, por ejemplo, mediante digestión con una exonucleasa I. ADN bicatenarios (18) que son productos de extensiones (5) tienen primer y segundo sitios de unión de cebador en cada extremo, a los que (en algunas realizaciones) cebadores directos e indirectos, con secuencias complementarias (6 y 11) puede añadirse para la última generación de clústeres mediante PCR de puente. En algunos casos, el ADN bicatenario también tiene un marcador de secuencia (19) y el cebador directo o indirecto puede incluir un marcador de muestra (2) para identificar o rastrear o asociar ADN (18) con una muestra o paciente. En algunos casos, el marcador de secuencia (19) es sustancialmente único para cada ácido nucleico recombinado distinto en una muestra. Como se explica más completamente a continuación, el marcador de secuencia (19) puede usarse para fusionar lecturas de secuencia en clonotipos bien como se ha usado para detectar y rastrear la contaminación de muestra. Los cebadores directos e indirectos también pueden incluir sitios de unión de cebador (4) y (8) para implementar (13) PCR de puente para determinados protocoles de secuenciación, por ejemplo, sobre un Analizador de Genomas (Illumina, San Diego)(17). En otras realizaciones, en las que se lleva a cabo más de una extensión con cebadores que contienen marcador, cada ácido nucleico recombinado distintos en una muestra puede tener copias con distintos marcadores de secuencia unidos; por tanto, por ejemplo, si se llevan a cabo cuatro ciclos por separado de fusión, hibridación y extensión sobre polinucleótidos diana de acuerdo con la realización de la Fig. 1A, y la muestra contiene ácido nucleico recombinado, S_1, entonces, cuando se finaliza la amplificación (13) con cebadores comunes, las copias de S_1 tendrán hasta cuatro marcadores de secuencia distintos. Por lo tanto, las lecturas de secuencia de S_1 tendrán hasta cuatro marcadores de secuencia distintos. Como se explica más completamente a continuación, en tales realizaciones, se pueden determinar clonotipos mediante una combinación de alineación de los marcadores de secuencia y fusión de las lecturas de secuencia dentro de cada subconjunto definido mediante un marcador de secuencia común.

En otra realización, al menos dos extensiones y dos etapas para retirar cebadores no incorporados se implementan antes de la PCR con cebadores comunes. Tal como se ilustra en la Fig. 1B, los cebadores (101) se hibrida a un extremo de polinucleótidos diana (100) tal como ácidos nucleicos recombinados que codifican cadenas de receptor inmunitario, y se extienden, por ejemplo, con una polimerasa de ADN. Los cebadores (101) pueden cada uno incluir una porción específica a receptor (103) y una porción de extremo 5' no complementaria (105) que, a su vez, comprende un marcador de secuencia (104) y un primer sitio de unión de cebador (102). Después de la extensión y retiración de cebadores no incorporados (130), tal como se ha descrito anteriormente, al primer producto de extensión (109) en la mezcla de reacción se añade (a) cebadores (125), en los que cada cebador comprende una porción específica a receptor (106) y una porción de extremo 5' no complementaria (115) (que contiene un sitio de unión de cebador) y (b) cebadores (127) que comprenden una porción (108) específica para un primer sitio de unión de cebador (102) y una porción de extremo 5' no complementaria (117). Después de que los cebadores (125) y (127) se hayan hibridado a sus sitió de unión de cebador, se extienden (107) para formar un segundo producto de extensión (118) después de lo cual se retiran los cebadores no extendidos. Al segundo producto de extensión (118) se añaden cebadores directos (112) y cebadores indirectos (110) comunes y se implementa (111) una PCR, después de la cual el amplicón resultante se secuencia (120). Como anteriormente en el ejemplo de la Fig. 1A, cuando se lleve a cabo más de una etapa de extensión en presencia de cebadores que contienen marcadores de secuencia (tales como (101)), pueden marcarse copias del mismo polinucleótido diana (100) con una pluralidad de distintos marcadores de secuencia.

La Fig. 1C ilustra otra realización con regiones V, D y J que se muestran explícitamente. En una mezcla de reacción con condiciones de hibridación de cebadores, a ácidos nucleicos recombinados (1200) que codifican receptores inmunitarios, tales como RLT, cebadores (1212) se añade un primer conjunto de cebadores específicos a la región V (1226). Cada cebador del primer conjunto (1212) incluye una porción específica a receptor y una porción de extremo 5' no complementaria que, a su vez, comprende un marcador de secuencia y un primer sitio de unión de cebador (por ejemplo, 102, 103 y 104 en Fig. 1B). Los cebadores del primer conjunto (1212) se hibridan a regiones V (1226) de ácidos nucleicos recombinados (1200) y los cebadores del primer conjunto (1212) se extienden (1202) a través de la región D (1224) en al menos la región J (1222) y opcionalmente a la región C (1220) para formar primeros productos de extensión (1216) que incluyen un marcador de secuencial opcional (1228) y un primer sitio de unión de cebador (1230). Después de retirar los cebadores no extendidos del primer conjunto (1212), los cebadores del segundo conjunto (1240) se añaden a la mezcla de reacción en condiciones de hibridación de modo que se hibridan a sus respectivas regiones J diana (1222), después de lo cual se extienden (1204) para formar segundos productos de extensión (1232), cada uno de los cuales comprende un marcador de secuencia (1236) (opcional) y un segundo sitio de unión de cebador (1234). Los segundos productos de extensión (1232) pueden comprender un único marcador de secuencia emplazado, por ejemplo, adyacente a regiones V (1226), tal como se muestra mediante el marcador de secuencia (1228) o adyacente a regiones J (1222), tal como se muestra mediante el marcador de secuencia (1236) o los segundos productos de extensión (1232) pueden comprender dos marcadores de secuencia emplazados en ambas posiciones. En una realización, los segundos productos de extensión (1232) comprenden un único marcador de secuencia (1228) adyacente a las regiones V (1226). En otra realización, segundos productos de

extensión (1232) comprenden un único marcador de secuencia (1236) adyacente a regiones J (1222). En algunas realizaciones, los marcadores de secuencia (1228) y/o (1236) son marcadores en mosaico descritos a continuación. Después de haber retirado los cebadores no extendidos del segundo conjunto (1240), se añaden cebadores directos e indirectos comunes que son específicos para primer y segundo sitios de unión de cebador (1230) y (1234), respectivamente, y se lleva a cabo una PCR (1206). Se secuencia (1208) una muestra del amplicón resultante para generar lecturas de secuencia para construir clonotipos y perfiles de clonotipo.

La Fig. 1D ilustra un método de generación de modelos de una longitud definida y para unir uno o dos marcadores de secuencia al mismo. La realización de la Fig. 1D muestra ARN mensajero (ARNm) como el material de partida, pero el método puede usarse o bien con muestras de ADN o ARN. A ARNm (1300) que contiene una región VDJ, se hibridan uno o más cebadores (1312) específicos para la región C (1308) ("cebadores C") a ARNm (1300). Normalmente solo se usa un único cebador C. Como alternativa, se puede usar uno o más cebadores (que tienen una estructura similar) específicos para la región J. El cebador C (1312) comprende un segmento específico diana (1313), un segmento de marcador de secuencia (1314) y un sitio de unión de cebador común (1315). También hay hibridados a ARNm diana (1300) bloqueadores de la polimerasa (1310) que pueden ser oligonucleótidos específicos para regiones V (1302). En algunas realizaciones, los bloqueadores (1310) pueden ser un oligonucleótido natural siempre y cuando se use una polimerasa para extender el cebador (1312) que no tenga ni actividad de desplazamiento de cadena ni actividad de exonucleasa 5'→3' y siempre y cuando el oligonucleótido no sea extensible, por ejemplo, tiene un 3'-dideoxinucleótido. Normalmente, los bloqueadores (1310) son análogos de oligonucleótidos con actividad de unión potenciada y resistencia de nucleasa, tal como compuestos antisentido. En algunas realizaciones, los bloqueadores (1310) pueden ser ácidos nucleicos bloqueados (LNA) o ácidos nucleicos peptídicos (PNA) o ácidos nucleicos unidos (BNA), que se desvelan en las siguientes referencias, Wengel y col., patentes de los EE.UU. 6.794.499; 7.572.582; Vester y col., Biochemistry, 43(42): 13233-13241 (2004); y similares, y Kazuyuki y col., Chem. Comm., 3765-3767 (2007); Nielson y col., Chem. Soc. Rev., 26: 73-78 (1997); y similares. Se seleccionan las secuencias de bloqueadores (1310) de modo que la extensión de cebador(es) (1312) se detienen en un emplazamiento predeterminado sobre la región V (1302). En algunas realizaciones, los bloqueadores (1310) están diseñados de modo que solo se copia lo suficiente de la región V (1302) en la etapa de extensión de modo que la región V puede identificarse a partir de la secuencia copiada. En algunas realizaciones, la obtención de bloqueadores (1310) específicos para cada región V es innecesaria, ya que las secuencias de consenso se pueden seleccionar de modo que permitan algunos desajustes, siempre y cuando la progresión de una polimerasa se detenga. Las longitudes de los bloqueadores (1310) puede variar ampliamente dependiendo del tipo de oligonucleótido o análogo usado. En algunas realizaciones, los bloqueadores (1310) tienen longitudes en el intervalo de 10 a 25 monómeros. En algunas realizaciones, los bloqueadores (1310) pueden hibridarse a distintos emplazamientos sobre distintas secuencias de región V.

Volviendo a la Fig. ID, los cebadores (1312) se extienden a los bloqueadores (1310) haciendo una copia de ADNc de una porción de la región VDJ de diana (1300) que tiene una longitud predeterminada. En algunas realizaciones, la longitud predeterminada (o de forma equivalente, los sitios de unión de los bloqueadores (1310)) se seleccionan de modo que una porción deseada de la región VDJ puede cubrirse por una o más lecturas de secuencia de la técnica de secuencia usada en el método. Después de finalizar la extensión, se digiere (1325) un modelo de ARN (1300) usando técnicas convencionales, por ejemplo, digestión con una ARNasa, tal como ARNasa y/o ARNasa A, para proporcionar un ADNc unicatenario (1326). A este ADNC se añade una cola de extremo 3' de mononucleótido, tal como una cola poliC, usando desoxinucleotidil transferasa terminal (TdT) en un protocolo convencional. A ADNc con cola (1331), adaptador (1336) que tiene un saliente complementario a la cola del mononucleótido de ADNc (1331), después de lo cual se extiende para producir ADN bicatenario (1340), que puede amplificarse, por ejemplo, mediante PCR (1337) y el amplicón resultante secuenciado (1338).

Los ácidos nucleicos recombinados que se someten a hipermutación, tal como ácidos nucleicos que codifican IgH, pueden amplificarse usando conjuntos de cebadores que incluyen cebadores que se unen a distintos sitios de unión de cebador sobre el mismo ácido nucleico recombinado; es decir, tales conjuntos pueden incluir cebadores que se unen a uno o más sitios de unión de cebador que no se solapan sobre el mismo ácido nucleico recombinado que codifica una cadena receptora. Tal conjunto puede comprender uno o ambos primer y segundo conjunto de cebadores. En algunas realizaciones, los ácidos nucleicos recombinados sometidos a hipermutación se amplifican con un primer conjunto de cebadores y un segundo conjunto de cebadores en donde al menos uno de los dos conjuntos comprende cebadores específicos para una pluralidad de sitios de unión de cebador que no se solapan, por ejemplo, un conjunto puede contener para cada segmento V distinto una pluralidad de cebadores cada uno específico para un cebador que no se solapa distintos que se une sobre los distintos segmentos V. Una realización aplicable a la amplificación de ácidos nucleicos recombinados que se someten a hipermutación se ilustra en las Fig. 3A-3B, en donde se emplean conjuntos anidados de cebadores para asegurar la amplificación de cada ácido nucleico recombinado en una muestra en condiciones, por ejemplo, de hipermutación somática, evolución clonal o similares. Los ácidos nucleicos recombinados, por ejemplo, moléculas que codifican IgH, se combinan en una mezcla de reacción con condiciones de hibridación con un primer conjunto anidado (302) de cebadores, que comprende en este ejemplo los grupos (304), (306) y (308) de cebadores específicos para distintos sitios junto con la región B (316) de ácidos nucleicos recombinados (300). En esta realización, el primer conjunto anidado comprende una pluralidad de grupos de cebadores, cada uno específico para un sitio o emplazamiento distinto de la región V, en donde los distintos miembros de un grupo son específicos para distintas variantes de la región V en el sitio. En

algunas realizaciones, la pluralidad de grupos se encuentra en el intervalo de 2-4; en otras realizaciones, la pluralidad es de 2 o 3. En algunas realizaciones, cada cebador del primer conjunto anidado (302) puede tener un único marcador de secuencia (314) y un primer sitio de unión de cebador (312) en una cola de extremo 5' no complementaria. Los cebadores del primer conjunto anidado (302) se hibridan a sus ácidos nucleicos recombinados y se extienden a través de la región D (318) y al menos una porción de la región J (320) para formar un primer amplicón (323) que comprende tres componentes (330), (332) y (334) que se corresponden con los tres subconjuntos de cebadores (304), (306) y (308), respectivamente. Cada miembro del primer amplicón (323) incorpora un marcador de secuencia (324) y un sitio de unión de cebador (326).

Después de haber retirado los cebadores no extendidos (322), se añade el segundo conjunto de cebadores anidados (340) a la mezcla de reacción en condiciones de hibridación. Tal como se ilustra en la Fig. 3A, los cebadores del segundo conjunto anidado (340) comprende subconjuntos (336) y (338) de cebadores que se hibridan en distintas posiciones que no se solapan sobre la región J (320) de miembros de un primer amplicón (323). En algunas realizaciones, el segundo conjunto anidado de cebadores puede contener un único grupo de cebadores. Los cebadores del segundo conjunto anidado (340) se extienden para formar un segundo producto de extensión (360) que comprende subconjuntos (350), (352) y (354) que, a su vez, cada uno comprende otros dos subconjuntos (subsubconjuntos) que se corresponden con los cebadores (336) y (338). En algunas realizaciones, el segundo conjunto anidado de cebadores (340) contiene cebadores específicos a solo un único sitio de unión de cebador y primer conjunto anidado de cebadores (302) contiene cebadores específicos a al menos dos sitios de unión de cebador que no se solapan. Después de retirar los cebadores no extendidos (342), pueden añadirse cebadores directos e indirectos comunes para llevar a cabo la PCR (356) y una muestra del amplicón resultante se puede secuenciar (358). En diversas realizaciones, los cebadores de tanto el primer conjunto anidado y el segundo conjunto anidado pueden incluir marcadores de secuencia (339); los cebadores del primer conjunto anidado pero no el segundo conjunto anidado pueden incluir marcadores de secuencia; y los cebadores del segundo conjunto anidado pero no el primer conjunto anidado pueden incluir marcadores de secuencia. En algunas realizaciones, los cebadores del primer conjunto anidado se extienden, en primer lugar, después de lo cual se retiran o destruyen los cebadores no extendidos y los cebadores del segundo conjunto anidado se hibridan y extienden (tal como se ilustra en las Fig. 3A-3B). En otras realizaciones, el orden de las etapas de hibridación, extensión y retiración se invierte; es decir, los cebadores del segundo conjunto anidado se extienden, en primer lugar, después de lo cual se retiran o destruyen los cebadores no extendidos y los cebadores del primer conjunto anidado se hibridan y extienden.

En algunas realizaciones del método anterior, puede implementarse más de una etapa de extensión, o bien (322) o (342), por ejemplo, para unir marcadores de secuencia a una fracción superior de polinucleótidos diana en una muestra. En dichas realizaciones, más de un marcador de secuencia diferente puede unirse al polinucleótido diana y/o copias del mismo. Es decir, una pluralidad de distintos marcadores de secuencia puede unirse a un polinucleótido diana y su progenie a partir de una reacción de amplificación, tal como PCR; por tanto, pueden marcarse copias de un polinucleótido diana original con más de un marcador de secuencia. Como se explica más completamente a continuación, tales pluralidades de marcadores de secuencia aún son útiles en el rastreo de la contaminación por arrastre y en permitir una determinación más sensible de secuencias de polinucleótidos diana.

Algunos de los casos descritos anteriormente pueden llevarse a cabo con las siguientes etapas. Por ejemplo, un método de generación de perfiles de clonotipo a partir de múltiples, o una pluralidad de, cadenas de receptores de linfocitos T puede comprender las etapas de: (a) combinar en una mezcla de reacción con condiciones de extensión de cebador un primer conjunto de cebadores con una muestra de ácidos nucleicos recombinados de linfocitos T, en donde cada cebador del primer conjunto tiene una porción específica a receptor con una longitud de modo que la porción específica a receptor se hibrida a un ácido nucleico recombinado distinta en un emplazamiento o sitio predeterminado sobre el ácido nucleico recombinado diana y se extiende para formar un primer producto de extensión, y en donde cada cebador del primer conjunto tiene un extremo 5' no complementario que contiene un primer sitio de unión del cebador; (b) retirar de la mezcla de reacción cebadores no extendidos del primer conjunto; (c) añadir a la mezcla de reacción en condiciones de extensión de cebador un segundo conjunto de cebadores, en donde cada cebador del segundo conjunto tiene una porción específica a receptor de modo que la porción específica a receptor se hibrida al primer producto de extensión en un emplazamiento o sitio predeterminado y tiene un extremo 5' no complementario que contiene un segundo sitio de unión de cebador, los cebadores del primer conjunto y/o los cebadores del segundo conjunto que comprenden un marcador de secuencia dispuestos entre la porción específica a receptor y el primer o segundo sitio de unión de cebador, respectivamente, y en donde cada cebador del segundo conjunto se extiende para formar un segundo producto de extensión, de modo que cada segundo producto de extensión comprende un primer sitio de unión de cebador, un segundo sitio de unión de cebador, al menos un marcador de secuencia y o (i) una porción de un segmento Vβ y una porción de un segmento Jβ de una cadena de receptor de linfocitos T, (ii) una porción de un segmento Vδ y una porción de un segmento J δ de una cadena de receptores de linfocitos T o (iii) una porción de un segmento Vγ y una porción de un segmento Jy de una cadena de receptores de linfocitos T; (d) llevar a cabo una reacción en cadena de la polimerasa en la mezcla de reacción para formar un amplicón, usando la reacción en cadena de la polimerasa cebadores directos específicos para el primer sitio de unión de cebador y cebadores indirectos específicos para el segundo sitió de unión de cebador; y (e) secuenciar los ácidos nucleicos del amplicón para formar un perfil de clonotipo de múltiples cadenas de receptores de linfocitos T. Tal como se utiliza en el presente documento, "condiciones de extensión de cebador" en una mezcla de reacción incluye condiciones en las que sustancialmente todos los sitios de unión de cebador se encuentran en

un estado unicatenario. En algunas realizaciones, tales condiciones se obtienen fundiendo ácidos nucleicos diana bicatenarios de modo que los sitios de unión de cebador se encuentran en forma unicatenaria de modo que los cebadores pueden hibridarse a ellos para formar sustratos para la extensión por polimerasa.

Los emplazamientos o sitios predeterminaros en los que los cebadores del primer y segundo conjuntos se unen pueden determinarse mediante métodos convencionales conocidos por los expertos en la técnica de amplificación de ácidos nucleicos por multiplexación, tal como PCR por multiplexación, tal como se ejemplificada en las referencias que se citan a continuación. Por ejemplo, en el caso de polinucleótidos diana que son ácidos nucleicos recombinados que codifican moléculas de receptores inmunitarios, Faham y Willis (citados anteriormente), Van Dongen y col., Leukemia, 17: 2257-2317 (2003), y referencias similares proporcionan directrices para seleccionar sitios de unión de cebador para amplificación por multiplexación de tales polinucleótidos diana. En algunas realizaciones, la selección de tales emplazamientos o sitios predeterminados depende de varios factores que incluyen (i) su efecto sobre la eficacia de amplificación (es deseable que las frecuencias de distintas copias en un amplicón representen fielmente las frecuencias de los polinucleótidos diana en una muestra), (ii) su efecto sobre las longitudes de las copias en un amplicón se corresponde con los requisitos de la química de secuenciación de ADN que se está empleando, (iii) si los cebadores seleccionados atraviesan una porción de los ácidos nucleicos recombinados con una diversidad deseada, por ejemplo, una región VDJ y similares. En relación con este aspecto, en parte de la invención se incluye una apreciación y reconocimiento de que la reactividad cruzada de los cebadores con distintos polinucleótidos diana no afecta los resultados de los métodos de la invención (en comparación con, por ejemplo, en los métodos basados únicamente en lecturas análogas de amplificaciones por PCR, espectrotipación y similares), ya que un conjunto de secuencia es la lectura en lugar de una señal de análogo.

En algunas realizaciones, la etapa de secuenciación incluye las siguientes etapas: (i) proporcionar una pluralidad de lecturas de secuencia teniendo cada una, una tasa de error y comprendiendo cada una, una secuencia de nucleótidos y una secuencia de marcador y (ii) alinear grupos de lecturas de secuencia que tienen secuencias de marcador similares, después de lo cual se realizan identificaciones de nucleótidos basándose en lecturas de secuencia dentro de tales grupos para determinar la secuencia de nucleótidos. Tales secuencias de nucleótidos de nivel de grupo pueden, a continuación, fusionarse en los mismos o distintos clonotipos como se describe a continuación. En algunas realizaciones, en las etapas de PCR, las longitudes de las porciones específicas a receptor de los cebadores del primer y segundo conjunto se seleccionan de modo que niveles relativos de distintos ácidos nucleicos recombinados en el amplicón son sustancialmente los mismos que los de los ácidos nucleicos recombinados en la muestra. En la implementación de tal selección de cebadores, las posiciones y longitudes de los sitios de unión de los cebadores sobre sus respectivos polinucleótidos diana puede variarse. En algunas realizaciones, los marcadores de secuencia se seleccionan a partir de un conjunto de marcadores de secuencia que es más grande que el número de distintos polinucleótidos diana en una muestra, de modo que sustancialmente cada polinucleótido diana diferentes en la muestra y copias del mismo tendrán un marcador de secuencia distintos (por ejemplo, de acuerdo con la metodología de "labeling by sampling" que se describe en Brenner, patente de los EE.UU. 7.537.897). En algunas realizaciones, el número de marcadores de secuencia en tal conjunto es al menos 100 veces el tamaño de la población de polinucleótidos diana en una muestra. Además, en algunas realizaciones en las que sustancialmente cada polinucleótido diana original y copias del mismo se marcan con el mismo marcador de secuencia único, la etapa de secuenciación incluye la generación de lecturas de secuencia de ácidos nucleicos del amplicón y la alineación de las lecturas de secuencia que tienen los mismos marcadores de secuencia para determinar lecturas de secuencia que se corresponden con los mismos clonotipos de la muestra. Además, en algunas realizaciones, la etapa de alineación incluye adicionalmente determinar una secuencia de nucleótidos de cada clonotipo determinando una mayoría de nucleótidos en cada posición de nucleótido de las lecturas de secuencia que tienen el mismo marcador de secuencia. Además, en algunas realizaciones, las etapas de retirar los cebadores no extendido se pueden llevar a cabo digiriendo ácidos nucleicos unicatenarios en la mezcla de reacción usando una nucleasa que tiene actividad de exonucleasa unicatenaria 3'→5' (que puede proporcionarse mediante, por ejemplo, exonucleasa I de *E. coli*, que puede inactivarse convenientemente por calor). En realizaciones adicionales, los anteriores métodos pueden usarse para generar perfiles de clonotipo para diagnosticar y/o controlar enfermedad mínima residual de un paciente con cáncer, tal como un paciente con mieloma, linfoma o leucemia. Tal diagnóstico y/o control puede implementarse con la siguiente etapa adicional después de las anteriores etapas del método: determinar a partir del perfil de clonotipo una presencia, ausencia y/o nivel de uno o más clonotipos específicos a paciente correlacionados con el cáncer. Los métodos de esta realización pueden incluir además etapas o determinación de secuencias de cada uno o más de los marcadores de secuencia y la comparación de tales secuencias con secuencias de marcadores de secuencia de perfiles de clonotipo previamente determinados para determinar una presencia, ausencia y/o nivel de secuencias de contaminación. En algunas realizaciones, tal etapa de comparación incluye la comparación de las secuencias de uno o más marcadores de secuencia con respecto a marcadores de secuencia de una base de datos de clonotipos que contiene clonotipos de al menos un individuo distinto al paciente.

En un ejemplo, un método de amplificación en una reacción una pluralidad de ácidos nucleicos recombinados que codifican componentes de receptor de linfocitos T β, δ y γ puede comprender las etapas de: (a) combinar en una mezcla de reacción con condiciones de extensión de cebador un primer conjunto de cebadores con una muestra de ácidos nucleicos recombinados de linfocitos T, en donde cada uno de los ácidos nucleicos recombinados comprende un primer extremo en al menos una porción de un segmento jβ, Jδ o Jγ de un receptor de linfocitos T, y en donde

cada cebador del primer conjunto cada uno tiene una porción específica a receptor con una longitud, cuya porción específica a receptor se hibrida al primer extremo de un ácido nucleico recombinado distinto y se extiende para formar un primer producto de extensión y en donde cada cebador del primer conjunto tiene un extremo 5' no complementario que contiene en un orden 3'→5' un marcador de secuencia y un primer sitio de unión de cebador, siendo el marcador de secuencia distinto para sustancialmente cada cebador del primer conjunto; (b) retirar de la mezcla de reacción cebadores no extendidos del primer conjunto; (c) añadir a la mezcla de reacción en condiciones de extensión de cebador un segundo conjunto de cebadores, teniendo cada cebador del segundo conjunto una porción específica a receptor con una longitud, que se hibrida al primer producto de extensión y se extiende para formar un segundo producto de extensión, en donde cada segundo producto de extensión comprende la menos una porción de un segmento Vβ, Vδ o Vγ de un receptor de linfocitos T, y en donde cada cebador del segundo conjunto tiene un extremo 5' no complementario que contiene un segundo sitio de unión de cebador; y (d) llevar cabo una reacción en cadena de la polimerasa en la mezcla de reacción para formar un amplicón, usando la reacción en cadena de la polimerasa un cebador directo específico para el primer sitio de unión de cebador y un cebador indirecto específico para el segundo sitio de unión de cebador. El método anterior puede incluir adicionalmente una etapa de secuenciación de una muestra de secuencias del amplicón. Normalmente, tal muestra es una "muestra representativa" en que es lo suficientemente grande con respecto a los distintos clonotipos presentes en la muestra en aproximadamente las mismas frecuencias que en la muestra original de material biológico. En algunas realizaciones, la etapa de secuenciación incluye proporcionar una pluralidad de lecturas de secuencia teniendo cada una, una tasa de error y comprendiendo cada una, una secuencia de nucleótidos y una secuencia de marcador y alinear lecturas secuencia que tengan secuencias de marcador similares para determinar lecturas de secuencia que se corresponden con el mismo clonotipo. Tales lecturas de secuencia pueden procesar en una etapa adicional de fusión, tal como se explica más completamente a continuación, siempre que haya múltiples marcadores de secuencia unidos a los polinucleótidos diana originales o copias de los mismos.

En otro ejemplo, un método de generación de perfiles de clonotipo a partir de múltiples cadenas de receptores de linfocitos T puede comprender las etapas de: (a) combinar en una mezcla de reacción con condiciones de extensión de cebador un primer conjunto de cebadores con una muestra de ácidos nucleicos recombinados de linfocitos T, en donde cada cebador del primer conjunto tiene una porción específica a receptor con una longitud de modo que la porción específica a receptor se hibrida a un ácido nucleico recombinado distinta en un emplazamiento predeterminado y se extiende para formar un primer producto de extensión, y en donde cada cebador del primer conjunto tiene un extremo 5' no complementario que contiene un primer sitio de unión del cebador; (b) retirar de la mezcla de reacción cebadores no extendidos del primer conjunto; (c) añadir a la mezcla de reacción un segundo conjunto de cebadores, en donde cada cebador del segundo conjunto tiene una porción específica a receptor con una longitud, siendo la porción específica a receptor específica para el primer producto de extensión en un emplazamiento predeterminado y que tiene un extremo 5' no complementario que contiene un segundo sitio de unión de cebador, los cebadores del primer conjunto y/o los cebadores del segundo conjunto que comprenden un marcador de secuencia dispuestos entre la porción específica a receptor y el primer o segundo sitio de unión de cebador, respectivamente; (d) realizar una primera reacción en cadena de la polimerasa para formar un primer amplicón, usando la primera reacción en cadena de la polimerasa cebadores directos específicos para el primer sitio de unión de cebador y cebadores del segundo conjunto, en donde cada secuencia de nucleótidos del primer amplicón comprende un primer sitio de unión de cebador, un segundo sitio de unión de cebador, al menos un marcador de secuencia y o una porción de un segmento Vβ y una porción de un segmento Jβ de una cadena de receptor de linfocitos T, un porción de un segmento Vδ y una porción de un segmento J δ de una cadena de receptores de linfocitos T, o una porción de un segmento Vγ y una porción de un segmento Jy de una cadena de receptores de linfocitos T y en donde las longitudes de las porciones específicas a receptor de los cebadores del primer y segundo conjuntos se seleccionan de modo que los niveles relativos de distintos ácidos nucleicos recombinados en el amplicón son sustancialmente los mismos que los de los ácidos nucleicos recombinados en la muestra; (e) añadir cebadores indirectos específicos para el segundo sitio de unión de cebador; (f) llevar a cabo una segunda reacción en cadena de la polimerasa en la mezcla de reacción para formar un segundo amplicón, usando la reacción en cadena de la polimerasa cebadores directos específicos para el primer sitio de unión de cebador y cebadores indirectos específicos para el segundo sitió de unión de cebador; (g) secuenciar los ácidos nucleicos del segundo amplicón para formar un perfil de clonotipo de múltiples cadenas de receptores celulares. En algunas realizaciones, la etapa de secuenciación incluye proporcionar una pluralidad de lecturas de secuencia teniendo cada una, una tasa de error y comprendiendo cada una, una secuencia de nucleótidos y una secuencia de marcador y alinear lecturas secuencia que tengan secuencias de marcador similares para determinar lecturas de secuencia que se corresponden con el mismo clonotipo. En realizaciones adicionales en las que polinucleótidos diana y/o copias del mismo se marcan con más de un marcador de secuencia, después de alinear marcadores de secuencia iguales, se pueden procesar las lecturas de secuencia en una etapa adicional de fusión, tal como se explica más completamente a continuación.

En otro ejemplo, un método de generación de perfiles de clonotipo a partir de múltiples cadenas de receptores de linfocitos B puede llevarse a cabo mediante las etapas de: (a) combinar en una mezcla de reacción con condiciones de extensión de cebador un primer conjunto anidado de cebadores con una muestra de ácidos nucleicos recombinados de linfocitos B, comprendiendo el primer conjunto anidado uno o más grupos de cebadores, en donde cada cebador de cada grupo tiene una porción específica a receptor con una longitud de modo que la porción específica a receptor de cada cebador a partir de un grupo distinto se hibrida a un ácido nucleico recombinado

distinta en un sitio predeterminado que no se solapa con un sitio predeterminado de cualquier otro cebador del primer conjunto anidado y en donde cada cebador de cada grupo tiene un extremo 5' no complementario que contiene un primer sitio de unión del cebador; (b) extender los cebadores del primer conjunto anidado para formar un primer producto de extensión; (c) retirar de la mezcla de reacción cebadores no extendidos del primer conjunto anidado; (d) añadir a la mezcla de reacción en condiciones de extensión de cebador un segundo conjunto anidado de cebadores, comprendiendo el segundo conjunto anidado uno o más grupos de cebadores, en donde cada cebador de cada grupo tiene una porción específica a receptor con una longitud de modo que la porción específica a receptor de cada cebador a partir de un grupo distinto se hibrida al primer producto de extensión en un sitio predeterminado que no se solapa con un sitio predeterminado de cualquier otro cebador del segundo conjunto anidado y en donde cada cebador de cada grupo tiene un extremo 5' no complementario que contiene un segundo sitio de unión de cebador, y en donde los cebadores del primer conjunto anidado y/o los cebadores del segundo conjunto anidado comprende un marcador de secuencia dispuesto entre su porción específica a receptor y su primer o segundo sitio de unión de cebador, respectivamente; (e) extender los cebadores del segundo conjunto anidado para formar un segundo producto de extensión, de modo que cada segundo producto de extensión comprende un primer sitio de unión de cebador, un segundo sitio de unión de cebador, al menos un marcador de secuencia y o (i) una porción de un segmento V y una porción de un segmento J de una cadena de receptores de linfocitos B o (ii) una porción de un segmento V y una porción de un segmento J de una cadena ligera kappa de receptores de linfocitos B; (f) llevar a cabo una reacción en cadena de la polimerasa en la mezcla de reacción para formar un amplicón, usando la reacción en cadena de la polimerasa cebadores directos específicos para el primer sitio de unión de cebador y cebadores indirectos específicos para el segundo sitió de unión de cebador; y (g) secuenciar los ácidos nucleicos del amplicón para formar un perfil de clonotipo de múltiples cadenas de receptores de linfocitos B.

En algunas realizaciones, puede implementarse más de un ciclo de hibridación y extensión de cebadores (después de fundir el producto de extensión) en las etapas (b) y/o (e), en cuyo caso las copias de los ácidos nucleicos recombinados originales en la muestra pueden marcarse con uno o más marcadores de secuencia. En estas realizaciones, la etapa de secuenciación (g) puede incluir etapas adicionales de alineación y fusión tal como se describe a continuación para determinar clonotipos y perfiles de clonotipo. En algunas realizaciones, por ejemplo, en donde solo se llevan a cabo extensiones únicas en las etapas (b) y (e), la etapa de secuenciación incluye proporcionar una pluralidad de lecturas de secuencia teniendo cada una, una tasa de error y comprendiendo cada una, una secuencia de nucleótidos y una secuencia de marcador y alinear lecturas secuencia que tengan secuencias de marcador similares para determinar lecturas de secuencia que se corresponden con el mismo clonotipo. Como antes, en algunas realizaciones, en la PCR las posiciones y las longitudes de las porciones específicas a receptor de los cebadores del primer y segundo conjunto se seleccionan de modo que niveles relativos de distintos ácidos nucleicos recombinados en el amplicón son sustancialmente los mismos que los de los diferentes ácidos nucleicos recombinados en la muestra.

En algunas de las realizaciones, los marcadores de secuencia están unidos a un polinucleótido diana o una copia del mismo en una etapa de extensión de cebador, en donde sustancialmente cada polinucleótido diana distinto y copia del mismo se marca con el mismo marcador de secuencia. En otras realizaciones, los polinucleótidos diana de una muestra o copias de los mismos pueden marcarse con más de un marcador de secuencia distinto. Tal como se explica adicionalmente a continuación, en algunas realizaciones, pueden llevarse a cabo múltiples extensiones o múltiples ciclos de una PCR en presencia de cebadores que contienen marcadores de secuencia (bien un primer conjunto de cebadores o bien un segundo conjunto de cebadores), que puede dar como resultado distintos marcadores de secuencia que están unidos al mismo polinucleótido diana y/o sus copias.

Marcadores de secuencia en análisis de clonotipo

En el presente documento se describe un método para obtener y analizar datos de secuencias a partir de un repertorio de moléculas inmunitarias, tales como receptores de linfocitos T (RLT) o receptores de linfocitos B (RLB) o fragmentos definidos de los mismos, para determinar rápida y eficazmente un perfil de clonotipo. Los datos de secuencias normalmente incluyen una gran colección de lecturas de secuencias, es decir, secuencias de identificaciones de nucleótidos y puntuaciones de calidad asociadas, a partir de un secuenciador de ADN usado para analizar las moléculas inmunitarias. Un reto clave en la construcción de perfiles de clonotipo es distinguir rápidamente y con precisión lecturas de secuencias que contengan diferencias genuinas de aquellas que contienen errores de fuentes no biológicas, tal como las etapas de extracción, química de secuenciación, química de amplificación o similares. Un aspecto de la invención incluye unir un único marcador de secuencia a cada polinucleótido diana, por ejemplo, ácidos nucleico recombinado, en una muestra para ayudar a determinar si las lecturas de secuencia de tales conjugados derivan del mismo polinucleótido diana original. De acuerdo con un aspecto de la invención, los marcadores de secuencia están unido a las moléculas de ácidos nucleicos somáticamente recombinadas para formar conjugados de marcador-molécula en donde cada ácido nucleico recombinado de tal conjugado tiene un único marcador de secuencia. Normalmente, tal unión se realiza después de extraer las moléculas de ácido nucleico de una muestra que contiene linfocitos T y/o linfocitos B y/o ADN sin células. Preferentemente, tales marcadores de secuencia únicos difieren lo máximo posible entre sí tal como se ha determinado mediante mediciones de distancia convencionales para secuencias, tales como una distancia de Hamming o similares. Maximizando la distancia entre los marcadores de secuencia en conjugados de marcador-molécula, incluso una alta tasa de errores de secuenciación y amplificación, un marcador de secuencia de un

conjugado permanece bastante más cerca de su secuencia de marcador ancestral que el de cualquier otra secuencia de marcador de un conjugado distinto. Por ejemplo, si se emplean marcadores de secuencia de 16 meros y cada tal marcador sobre un conjunto de clonotipos tiene una distancia Hamming de al menos el cincuenta por ciento, u ocho nucleótidos, a partir de cada otro marcador de secuencia sobre los clonotipos, entonces, serían necesario al menos ocho errores de secuenciación o amplificación para transformar tal marcador en otro para una lectura errónea de un marcador de secuencia (y la agrupación incorrecta de un lectura de secuencia de un clonotipo con el marcador de secuencia erróneo). En una realización, los marcadores de secuencia se seleccionan de modo que después de la unión a moléculas de ácidos nucleicos recombinados para formar conjugados de marcador-molécula, la distancia Hamming entre los marcadores y los conjugados marcador-molécula en un número al menos el veinticinco por ciento de la longitud total de tales marcadores de secuencia (es decir, cada marcador de secuencia difiere en secuencia desde cada otro marcado en al menos un 25 por ciento de sus nucleótidos); en otra realización, la distancia Hamming entre tales marcadores de secuencia es un número al menos del 50 por ciento de la longitud total de tales marcadores de secuencia.

En un caso, un método se implementa mediante las siguientes etapas: (a) obtener una muestra de un individuo que comprende linfocitos T y/o linfocitos B y/o ADN sin células; (b) unir marcadores de secuencias a moléculas de ácidos nucleicos recombinados de genes de receptores de linfocitos T o genes de inmunoglobulina en la muestra para formar conjugados marcador-molécula, en donde sustancialmente cada molécula de los conjugados marcador-molécula tiene un marcador de secuencia único; (c) amplificar los conjugados de marcador-molécula; (d) secuenciar los conjugados de marcador-molécula; y (e) alinear las lecturas de secuencia de marcadores de secuencia similares para determinar las lecturas de secuencia que se corresponden con el mismo ácido nucleico recombinado en la muestra. Las muestras que contienen linfocitos B o linfocitos T se obtienen usando técnicas convencionales. En la etapa de unión de marcadores de secuencia, preferentemente, los marcadores de secuencia no son solo únicos sino que también son suficientemente distintos entre sí de modo que la probabilidad de incluido un número grande de errores de secuenciación o amplificación en la transformación de un marcador de secuencia en otro sería cerca de cero. Después de unir los marcadores de secuencia, la amplificación del conjugado de marcador-molécula es necesario para la mayoría de tecnologías de secuenciación; sin embargo, cuando se emplean tecnologías de secuenciación de molécula única una etapa de amplificación es opcional. Las tecnologías de secuenciación de molécula única incluyen, pero sin limitación, secuenciación en tiempo real de molécula única (SMRT), secuenciación de nanoporos o similares, por ejemplo, patentes de los EE.UU. 7.313.308; 8.153.375; 7.907.800; 7.960.116; 8.137.569; Manrao y col., Nature Biotechnology , 4(8): 2685-2693 (2012); y similares.

En el presente documento se describe un método para la determinación del número de linfocitos en una muestra mediante el recuento de marcadores de secuencia únicos. Incluso sin marcadores de secuencia, los clonotipos de genes de RLTβ o IgH, en particular, los que incluyen las regiones V(D)J, proporcionan para un linfocito y sus clones un marcador único. Siempre que se obtengan ácidos nucleicos recombinados a partir de un ADN genómico, entonces, un recuento de linfocitos en una muestra puede estimarse por el número de clonotipos únicos que se recuenten después de su secuenciación. Este enfoque deja de funcionar siempre que hay significantes poblaciones clonales de linfocitos idénticos asociados con el mismo clonotipo (o cuando se obtienen ácidos nucleicos recombinados a partir del ARNm de una muestra, cuya cantidad de secuencias individuales puede reflejar, o depender de, la tasa de expresión así como el número de identificación). El uso de marcadores de secuencia supera este punto débil y es especialmente útil para proporcionar recuentos de linfocitos en pacientes que padezcan cualquier trastorno linfoides, tal como linfomas o leucemias. Los marcadores de secuencia pueden usarse para obtener un recuento absoluto de linfocitos en una muestra independientemente de si hay un gran clon dominante presente, tal como con la leucemia. Tal método puede implementarse con las etapas de: (a) obtener una muestra de un individuo que comprende linfocitos; (b) unir marcadores de secuencias a moléculas de ácidos nucleicos recombinados de genes de receptores de linfocitos T o de genes de inmunoglobulina de los linfocitos para formar conjugados marcador-molécula, en donde sustancialmente cada molécula de los conjugados marcador-molécula tiene un marcador de secuencia único; (c) amplificar los conjugados de marcador-molécula; (d) secuenciar los conjugados de marcador-molécula; y (e) recontar el número de marcadores de secuencia distintos para determinar el número de linfocitos en la muestra. En algunas realizaciones, las moléculas de ácidos nucleicos recombinados son de ADN genómico.

En una realización de la invención, los marcadores de secuencia se unen a moléculas de ácido nucleico recombinado de una muestra mediante marcación y muestreo, por ejemplo, tal como se desvela por Brenner y col., patente de los EE.UU. 5.846.719; Brenner y col., patente de los EE.UU. 7.537.897; Macevicz, publicación de patente internacional WO 2005/111242; y similares. En la marcación mediante muestreo, los polinucleótidos de una población a macar (o únicamente marcados) se usan para muestrear (mediante unión, enlace o similares) marcadores de secuencia de una población mucho más grande. Es decir, si la población de polinucleótidos tiene miembros K (incluidas réplicas del mismo polinucleótido) y la población de marcadores de secuencia tiene miembros N, entonces N>>K. En una realización, el tamaño de una población de marcadores de secuencia usado con la invención es de al menos 10 veces el tamaño de la población de clonotipos en una muestra; en otra realización, el tamaño de una población de marcadores de secuencia usado con la invención es de al menos 100 veces el tamaño de la población de clonotipos en una muestra; y, en otra realización, el tamaño de una población de marcadores de secuencia usado con la invención es de al menos 1000 veces el tamaño de la población de clonotipos en una muestra. En otras realizaciones, se selecciona un tamaño de una población de marcador de secuencia de modo que

sustancialmente cada clonotipo en una muestra tendrá un marcador de secuencia único siempre que tales clonotipos se combinen con tal población de marcador de secuencia, por ejemplo, en una reacción de unión, tal como una reacción de ligación, reacción de amplificación o similares. En algunas realizaciones, sustancialmente cada clonotipo significa al menos el 90 por ciento de tales clonotipos tendrá un marcador de secuencia único; en otras realizaciones, sustancialmente cada clonotipo significa al menos el 99 por ciento de tales clonotipos tendrá un marcador de secuencia único; en otras realizaciones, sustancialmente cada clonotipo significa al menos el 99,9 por ciento de tales clonotipos tendrá un marcador de secuencia único. En muchas muestras de tejido o biopsias el número de linfocitos T o linfocitos B puede ser de hasta o aproximadamente 1 millón de células; por tanto, en algunas realizaciones de la invención que emplean tales muestras, el número de marcadores de secuencia único empleados en la marcación mediante muestreo es de al menos 10^8 o en otras realizaciones de al menos 10^9.

En dichas realizaciones, en las que hasta 1 millón de clonotipos se marcan mediante muestreo, grandes conjuntos de marcadores de secuencia pueden producirse eficazmente mediante síntesis combinatoria haciendo reaccionar una mezcla de todos los cuatros precursores de nucleótidos en cada etapa de adición de una reacción de síntesis, por ejemplo, tal como se desvela en Church, patente de los EE.UU. 5.149.625. El resultado es un conjunto de marcadores de secuencia que tienen una estructura de "N_1N_2 ... N_k" donde cada N_1=A, C, G o T y k es el número de nucleótidos en los marcadores. El número de marcadores de secuencia en un conjunto de marcadores de secuencia realizado mediante tal síntesis combinatoria es 4^k. Por lo tanto, un conjunto de tales marcadores de secuencia con k al menos 14, o k en el intervalo de 14 a 18, es adecuado para unir marcadores de secuencia a una población de miembros de 10^6 de moléculas mediante marcación por muestreo. Los conjuntos de marcadores de secuencia con la anterior estructura incluyen muchas secuencias que pueden introducir complicaciones o errores mientras que se implementan los métodos de la invención. Por ejemplo, el conjunto sintetizado combinatoriamente anterior de marcadores de secuencia incluye muchos marcadores de miembros con segmentos de homopolímeros que algunos enfoques de secuenciación, tales como enfoques de secuenciación por síntesis, tiene complicaciones en determinar con precisión por encima de una determinada longitud. Por lo tanto, la invención incluye marcadores de secuencia combinatoriamente sintetizados que tienen estructuras que son eficaces para etapas de métodos particulares, tales como secuenciación. Por ejemplo, varias estructuras de marcador de secuencia para químicas de secuenciación por síntesis pueden realizarse dividiendo los cuatro nucleótidos naturales en subconjuntos disociados que se usan alternativamente en síntesis combinatoria, evitando, de este modo, segmentos de homopolímeros por encima de una longitud dada. Por ejemplo, si z es o A o C y x es o G o T, para proporcionar una estructura de marcador de secuencia de

$$[(z)_1(z)_2 \ldots (z)_i][(x)_1(x)_2 \ldots (x)_j] \ldots$$

en donde i y j, que pueden ser iguales o diferentes, se seleccionan para limitar el tamaño de cualquier segmento de homopolímero. En una realización, i y j se encuentran en el intervalo de 1 a 6. En dichas realizaciones, los marcadores de secuencia pueden tener longitudes en el intervalo de 12 a 36 nucleótidos; y, en otras realizaciones, tales marcadores de secuencia pueden tener longitudes en el intervalo de 12 a 24 nucleótidos. En otras realizaciones, se pueden usar otros apareamientos de nucleótidos, por ejemplo, z es A o T y x es G o C; o z es A o G y x es T o C. Alternativamente, si z' es cualquier combinación de tres de los cuatro nucleótidos naturales y x' es cualquier nucleótido que no es una z' (por ejemplo, z' es A, C o G y x' es T). Esto proporciona una estructura de marcador de secuencia que sigue:

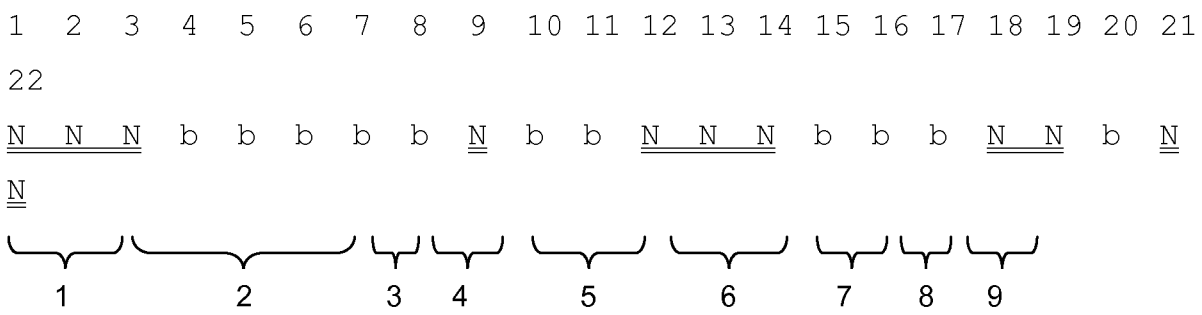
$$[(z')_1(z')_2 \ldots (z')_i]x'[(z')_1(z')_2 \ldots (z')_i]x' \ldots$$

en donde i se selecciona como antes y la aparición de x' sirve como una puntuación para finalizar cualquier homopolímero indeseado.

Marcadores de secuencia adicionales

La invención usa métodos de marcación de ácidos nucleicos, tales como fragmentos de ADN genómico, con marcadores de secuencia únicos, que pueden incluir "marcadores en mosaico", antes de la amplificación y secuenciación. Tales marcadores de secuencia son útiles para identificar errores de amplificación y secuenciación. Los marcadores en mosaico minimizan los artefactos de secuenciación y amplificación debido a una hibridación inadecuada, sensibilización, formación de horquillados o similares, que pueden producirse con marcadores de secuencia completamente aleatorios de la técnica anterior. En un aspecto, los marcadores en mosaico son marcadores de secuencia que comprenden alternar regiones constantes y regiones variables, en donde cada región constante tiene una posición en el marcador en mosaico y comprende una secuencia predeterminada de nucleótidos y cada región variable tiene una posición en el marcador en mosaico y comprende un número predeterminado de nucleótidos seleccionados aleatoriamente. A modo de ilustración, un marcador en mosaico de 22 meros (SEQ ID NO: 1) puede tener la siguiente forma:

Posición de nucleótido:



Posición de región

Hay nueve regiones constantes y variables, con las regiones 1 (nucleótidos 1-3), 3 (nucleótido 9), 5 (nucleótidos 12-14), 7 (nucleótidos 18-19) y 9 (nucleótidos 21-22) siendo variables (nucleótidos con doble subrayado) y las regiones 2 (nucleótidos 4-8), 4 (nucleótidos 10-11), 6 (nucleótidos 15-17) y 8 (nucleótido 20) siendo constantes. N representa un nucleótido seleccionado aleatoriamente del conjunto de A, C, G o T; por tanto, el número de marcadores en mosaico de este ejemplo es de 4^{11} = 4.194.304 marcadores. b representa un nucleótido predeterminado en la posición indicada. En algunas realizaciones, la secuencia de b, "***bbbbb*bb***bbb**b**", se selecciona para minimizar la probabilidad de tener una coincidencia perfecta en un genoma del organismo que conforma la muestra.

En un aspecto, para los marcadores en mosaico de una realización particular del método de la invención, todas las regiones constantes con la misma posición tienen la misma longitud y todas las regiones variables con la misma posición tienen la misma longitud. Esto permite que los marcadores en mosaico se sinteticen usando síntesis combinatoria parcial con químicas e instrumentos convencionales.

En un aspecto, los marcadores en mosaico comprenden de 10 a 100 nucleótidos, o de 12 a 80 nucleótidos, o de 15 a 60 nucleótidos. En algunas realizaciones, los marcadores en mosaico comprende al menos ocho posiciones de nucleótidos con nucleótidos aleatoriamente seleccionados; en otras realizaciones, siempre que los marcadores en mosaico tengan una longitud de al menos 15 nucleótidos, comprenden al menos 12 posiciones de nucleótidos con nucleótidos aleatoriamente seleccionados. En otro aspecto, ninguna región variable dentro de un marcador de mosaico puede tener una longitud que sea superior a siete nucleótidos.

En otro aspecto, los marcadores en mosaico se pueden usar en las siguientes etapas: (i) preparación de modelos de ADN a partir de ácidos nucleicos en una muestra; (ii) marcación mediante muestreo de modelos de ADN para formar una multiplicidad de conjugados de marcador-modelo, en donde sustancialmente cada modelo de ADN de un conjugado de marcado-modelo tiene un marcador en mosaico único que comprende regiones constantes y regiones variables alternantes, teniendo cada región constante una posición en el marcador en mosaico y una longitud de 1 a 10 nucleótidos de una secuencia predeterminada y teniendo cada región variable una posición en el marcador en mosaico y una longitud de 1 a 10 nucleótidos aleatoriamente seleccionados, de modo que las regiones constantes que tienen las mismas posiciones tienen las mismas longitudes y la región variable que tiene las mismas posiciones tienen las mismas longitudes; (iii) amplificación de la multiplicidad de los conjugados de marcador-modelo; (iv) generación de una pluralidad de lecturas de secuencia para cada uno de los conjugados de marcador-modelo amplificados; y (v) determinación de una secuencia de nucleótidos de cada uno de los ácidos nucleicos determinando un nucleótido de consenso en cada posición de nucleótido de cada pluralidad de lecturas de secuencia que tienen marcadores en mosaico idénticos. En otro aspecto, los marcadores en mosaico se pueden usar en las siguientes etapas: (a) preparación de modelos de ADN unicatenarios a partir de ácidos nucleicos en una muestra; (b) marcación mediante muestreo de los modelos de ADN unicatenarios para formar conjugados de marcador-modelo, en donde sustancialmente cada modelos de ADN unicatenario de un conjugado de marcador-modelo tiene un marcador de secuencia único (es decir, un marcador en mosaico) que tiene una longitud de al menos 15 nucleótidos y que tiene la siguiente forma:

$$[(N_1N_2 \ldots N_{Kj})(b_1b_2 \ldots b_{Lj})]_M$$

en donde cada N_i, para i= 1, 2, ... K_j, es un nucleótidos aleatoriamente seleccionado entre el grupo que consiste en A, C, G y T; K_j es un número entero en el intervalo de 1 a 10 para cada j inferior a o igual a M (es decir, las regiones N_1N_2 ... N_{Kj} son regiones variables); cada b_i, para i= 1, 2, ... L_j, es un nucleótido; L_j es un número entero en el intervalo de 1 a 10 para cada j inferior a o igual a M; de modo que cada marcador de secuencia (i) tiene el mismo Kj para cada jy (ii) tiene las mismas secuencia b_1b_2 ... b_{Lj} para cada j (es decir, las regiones b_1b_2 ... b_{Lj} son regiones constantes); y M es un número entero superior a o igual a 2; (c) amplificación de los conjugados de marcador-modelo; (d) generación de una pluralidad de lecturas de secuencia para cada uno de los conjugados de marcador-modelo amplificados; y (e) determinación de una secuencia de nucleótidos de cada uno de los ácidos nucleicos

determinando un nucleótido de consenso en cada posición de nucleótido de cada pluralidad de lecturas de secuencia que tienen marcadores de secuencia idénticos. En algunas realizaciones, la pluralidad de lecturas de secuencia es de al menos 10^4; en otras realizaciones, la pluralidad de lecturas de secuencia es de al menos 10^5; en todavía otras realizaciones, la pluralidad de lecturas de secuencia es de al menos 10^6. En algunas realizaciones, la longitud total del anterior marcador de secuencia se encuentra en el intervalo de 15 a 80 nucleótidos.

Unión de marcadores de secuencia

Se puede usar una variedad de distintas reacciones de unión para unir marcadores única a sustancialmente cada clonotipo en una muestra además de los ilustrados anteriormente. Muchas técnicas para capturar subconjuntos de ácidos nucleicos de muestras, por ejemplo, para deducir la complejidad de muestra en tecnología de secuenciación de micromatriz o genómica, puede usarse con la modificación habitual en la presente invención a marcadores de secuencia unidos a ácidos nucleicos recombinados. Técnicas ejemplares para capturar diversos conjuntos de ácidos nucleicos diana para su posterior manipulación, incluidos marcadores de secuencia de unión, secuenciación y similares, incluyen las siguientes: Willis y col., patente de los EE.UU. 7.700.323; Jones y col., publicación de patente de los EE.UU. 2005/0142577; Gullberg y col., publicación de patente de los EE.UU. 2005/0037356; Porreca y col., Nature Methods, 4(11): 931-936 (2007); Turner y col., Nature Methods, 6(5): 315-316 (2009); Church, patente de los EE.UU. 5.149.625; Macevicz, patente de los EE.UU. 8.137.936; y similares.

En una realización, tal unión se consigue mediante la combinación de una muestra que contiene moléculas de ácidos nucleicos recombinados (que, a su vez, comprende secuencias de clonotipo) con una población o biblioteca de marcadores de secuencia de modo que los miembros de las dos poblaciones de moléculas pueden combinarse aleatoriamente y asociarse o enlazarse, por ejemplo, covalentemente. Por ejemplo, tal combinación aleatoria puede producirse en una reacción biomolecular en donde un cebador que contiene marcador se hibrida a un ácido nucleico diana y se extiende o en donde un adaptador que contiene marcador se liga al extremo de un ácido nucleico diana. En algunas realizaciones, el método de unión, de marcadores depende en parte del enfoque de secuenciación de ADN. Por ejemplo, en métodos de secuenciación que producen lecturas de secuencia relativamente largas precisas, tales como secuenciación 454, una biblioteca de ADNc puede realizarse a partir de ARnm que comprende ácidos nucleicos recombinados usando técnicas convencionales, por ejemplo, 5'-RACE, tal como se desvela en Freeman y col., Genome Research, 19: 1817-1824 (2009), después la cual los marcadores de secuencia puede unirse ligando adaptadores que contienen marcadores de secuencia a uno o ambos extremos. En otras realizaciones, cuando se usan métodos de secuenciación, tales como secuenciación "Illumina" o secuenciación "Ion Torrent", que producen lecturas de secuenciación relativamente cortas y con tendencia a error, pueden requerirse etapas adicionales de modo que los amplicones para la secuenciación tienen longitudes que se cubren por lecturas de secuencia generadas a partir de las técnicas. En tales reacciones de unión de marcador, las secuencias de clonotipo comprenden polinucleótidos unicatenario o bicatenarios lineales y marcadores de secuencia se llevan a cabo mediante reactivos tales como cebadores de amplificación, tal como cebadores de PCR, adaptadores de ligación, sondas circularizables, plásmidos o similares. Tales varios reactivos capaces portar poblaciones de marcador de secuencia se desvelan en Macevicz, patente de los EE.UU. 8.137.936; Faham y col., patente de los EE.UU. 7.862.999; Landegren y col., patente de los EE.UU. 8.053.188; Unrau y Deugau, Gene, 145: 163-169 (1994); Church, patente de los EE.UU. 5.149.625; y similares.

Las Fig. 2A y 2B ilustran una reacción de unión que comprende una PCR en la que una población de marcadores de secuencia (T_1, T_2, T_3 ... T_j, T_{j+1} ... T_k, T_{k+1} ... T_{n-1}, T_n) se incorpora en los cebadores (2100). La población de marcadores de secuencia tiene un tamaño mucho superior al de las moléculas de ácidos nucleicos recombinados (2102). Los marcadores de secuencia se unen a las moléculas de ácidos nucleicos recombinados mediante hibridación de los cebadores a las moléculas de ácidos nucleicos y extensión de los cebadores con ADN polimerasa en el primer ciclo de una PCR. La figura ilustra el modo en el que las moléculas de ácidos nucleicos recombinados seleccionan, o muestrean, una pequeña fracción de la población total de marcadores de secuencia hibridándose aleatoriamente a los cebadores a modo de sus regiones de unión de cebador comunes (2104), por ejemplo, en la región V (2108). Puesto que los cebadores (y, por lo tanto, marcadores de secuencia) se combinan con las moléculas de secuencia de ácidos nucleicos recombinados aleatoriamente, existe una pequeña posibilidad de que el mismo marcador de secuencia se haya podido unir a distintas moléculas de ácidos nucleicos; sin embargo, si la población de marcadores de secuencia es grande tal como se enseña en el presente documento, entonces la posibilidad será insignificantemente pequeña de modo que sustancialmente cada molécula de ácido nucleico recombinado tendrá un marcador de secuencia unido. El otro cebador (2106) de par de cebador directo e indirecto se hibrida a la región C (2110) de modo que después de múltiples ciclos de hibridación, extensión y fusión, se forma el amplicón (2112), uniendo, de este modo, marcadores de secuencia únicos a las regiones V(D)J que comprenden los clonotipos de población (2102). Es decir, el amplicón (2112) comprende los conjugados de marcador-molécula de la reacción de unión.

Las Fig. 2C y 2D ilustran un método de unión de un par de marcadores de secuencia a cada, o sustancialmente cada, ácido nucleico recombinado en una muestra. Como en el método de las Fig. 2A y 2B, los cebadores (2200) que portan marcadores de secuencia (T_1, T_2, T_3 ... T_j, T_{j+1} ... T_k, T_{k+1} ... T_{n-1}, T_n) se usan como cebadores corriente abajo y, adicionalmente, reemplazan el cebador común (2106), los cebadores (2206) que portan los marcadores de secuencia (Tm, Tm+1, Tm+2 ... Tq, Tq+1, Tq+2, ... Tr, Tr+1, Tr+2, ... Ts, Ts+1, Ts+2, ...) se usan cebadores

corriente arriba. Como con el conjunto de cebadores corriente abajo, el número de distintos marcadores de secuencia distintos portados por cebadores corriente arriba (2206) puede ser grande en comparación con el número de moléculas de ácidos nucleicos recombinados (2202) de modo que sustancialmente cada ácido nucleico recombinado (2202) tendrá un marcador único después de la amplificación. En algunas realizaciones, cada conjunto de marcadores de secuencia en cebadores (2206) y (2200) no necesitan ser tan grandes como los marcadores de secuencia en la realización de las Fig. 2A y 2B. Puesto que cada ácido nucleico recombinado está marcado de forma única por un par de marcadores de secuencia, compartir un marcador de secuencia del par con una diferencia de ácido nucleico recombinado no restará valor a la sustancial unicidad con respecto a un par de marcadores de secuencia que marcan un ácido nucleico recombinado único. Por lo tanto, en la realización de las Fig. 2C y 2D, los marcadores de secuencia de cada conjunto de cebador (2200) y (2206) pueden ser menos diversos que los marcadores de secuencia del conjunto de cebador (2100). Por ejemplo, si se emplean marcadores de secuencia aleatorios y cebadores (2100) contienen marcadores de secuencia de 16 meros, entonces, los cebadores (2200) y (2206) pueden cada uno contener marcadores de secuencia de 8 meros para proporcionar la misma diversidad de marcador de secuencia total. Por otra parte, la realización de las Fig. 2C y 2D funciona de forma similar a la de las Fig. 2A y 2B. Los marcadores de secuencia se unen a las moléculas de ácidos nucleicos recombinados mediante hibridación de los cebadores a las moléculas de ácidos nucleicos y extensión de los cebadores con ADN polimerasa en el primer ciclo de una PCR. Como antes, La Fig. 2C ilustra el modo en el que las moléculas de ácidos nucleicos recombinados seleccionan, o muestrean, una pequeña fracción de la población total de pares de marcadores de secuencia hibridándose aleatoriamente a los cebadores a modo de sus regiones de unión de cebador (2204) y (2205), por ejemplo, en la región V (2208) y la región C (2210), respectivamente. Puesto que los cebadores (y, por lo tanto, marcadores de secuencia) se combinan con las moléculas de secuencia de ácidos nucleicos recombinados aleatoriamente, existe una pequeña posibilidad de que el mismo par de marcadores de secuencia se haya podido unir a distintas moléculas de ácidos nucleicos; sin embargo, si la población de marcadores de secuencia es grande tal como se enseña en el presente documento, entonces la posibilidad será insignificantemente pequeña de modo que sustancialmente cada molécula de ácido nucleico recombinado tendrá un par único de marcadores de secuencia unido. Después de múltiples ciclos de hibridación, extensión y fusión, se forma el amplicón (2212), uniendo, de este modo, pares únicos de marcadores de secuencia únicos a las regiones V(D)J que comprenden los clonotipos de población (2202). Es decir, el amplicón (2212) comprende los conjugados de marcador-molécula de la reacción de unión.

En algunas realizaciones, se pueden usar sondas circularizables para capturar y unir marcadores de secuencia a ácidos nucleicos recombinados deseados, por ejemplo, con modificación habitual de técnicas desveladas por Porreca y col. (citado anteriormente); Willis y co.l (citado anteriormente); o referencias similares. Tal como se ilustra en las Fig. 2E y 2F, se proporciona una sondas circularizable (2302) que comprende los siguientes elementos: un segmento de unión diana corriente arriba (2304), un segmento de unión diana corriente abajo (2306) que tiene un extremo 5' fosforilado (2305); un marcador de secuencia (2310); un segundo sitio de unión de cebador común (2314); un sitio de escisión opcional (2308); y un primer sitio de unión de cebador común (2312). La sonda circularizable (2302) se combina en una mezcla de reacción en condiciones de hibridación con una muestra que contiene polinucleótidos diana (2300), que pueden ser, por ejemplo, primeras o segundas cadenas de un ADNc preparado a partir de ARNm usando técnicas convencionales. Como se muestra, los polinucleótidos diana comprenden regiones V, NDN, J y C de ácidos nucleicos recombinados que codifican cadenas IgH o RLTβ. En algunas realizaciones, las secuencias de segmentos de unión diana corriente arriba y corriente abajo (2304) y (2306), respectivamente, se seleccionan de modo que expanden una porción de la región VDJ de los polinucleótidos diana. La sonda circularizable (2302) y polinucleótidos diana (2300) forman complejo (2330) en la mezcla de reacción cuando se hibridan segmentos de unión diana corriente arriba y corriente abajo (2304 y 2306). En la presencia de un ADN polimerasa y dNTP, el segmento de unión diana corriente arriba (2304) se extiende (2340) hasta el segmento de unión diana corriente abajo (2306) que copia (y, por lo tanto, captura) una porción de la región VDH del polinucleótido diana. En presencia de una actividad ligasa, el segmento de unión diana corriente arriba extendido se liga a el segmento de unión diana corriente abajo (2306) formando, de este modo, un círculo de ADN unicatenario cerrado (2342). La mezcla de reacción puede, opcionalmente, a continuación tratarse (2344) con una exonucleasa para retirar sonda no reaccionada y polinucleótidos diana. En algunas realizaciones, los círculos unicatenarios (2342) se linealizan escindiendo el sitio de escisión (2308), que puede ser, por ejemplo, un sitio de reconocimiento de endonucleasa de corte infrecuente o insertando un monómero de ARN en la sonda y escindiendo con RNasa H o similares, después de lo cual insertos de marcador de VDJ de las sondas linealizadas (2348) pueden amplificarse mediante los cebadores (2350) y (2352). Los cebadores (2350) y (2352) pueden incluir regiones no complementarias para añadir elementos para permitir una secuenciación de ADN posterior (2354). Como alternativa, se puede usar un círculo unicatenario para generar modelos de nanobola para secuenciación directa, por ejemplo, Drmanac y col., Science, 327(5961): 78-81 (2010); patente de los EE.UU. 8.445.196; y similares.

La Fig. 2G ilustra otra realización para unir un marcador de secuencia a un ácido nucleico recombinado que codifica una molécula de receptor inmunitario. Directrices para implementar esta realización se pueden encontrar en Faham y Zheng, patente de los EE.UU. 7.208.295. Se combinan ácidos nucleicos recombinados (2450) en una mezcla de reacción en condiciones de hibridación para las sondas (2454) y adaptadores (2456). Las sondas (2454) comprenden una porción específica a receptor (2455) y una porción específica a adaptador (2457). Por ejemplo, las sondas (2454) pueden comprender una mezcla de sondas en sonda distintas sondas tienen porciones específicas a receptores específicas para distintas regiones J o, en otras realizaciones, específicas para distintas regiones V. Los

adaptadores (2456), que están fosforilados en el extremo 5', comprenden una porción específica a sonda (2458) en su extremo 5', un marcador de secuencia (2460) y un primer sitio de unión de cebador (2462). Los emplazamientos, secuencias y longitudes de la porción específica a receptor (2455) y la porción específica a adaptador (2457) de la sonda (2454) y porción específica a sonda (2458) se seleccionan de modo que se hibridan entre sí para formar estructuras (2452). Después de que se forme la estructura (2452), se escinde la porción unicatenaria (2461) del ácido nucleico recombinado (2450) y el extremo 3' libre del ácido nucleico recombinado (2450) se liga al extremo 5' fosforilado del adaptador (2456) para formar un primer producto de extensión (2459), después de lo cual la sonda (2454) se retira (2474). La escisión de (2461) puede efectuarse mediante una nucleasa unicatenaria, tal como se describe en Faham y Zheng. En una realización, la sonda (2454) se sintetiza con timidinas reemplazadas con uracilos, por ejemplo, en una PCR con dUTP en lugar de dTTp y se retira tratándola con glicosilasa de ADN de uracilo (UDG), por ejemplo, tal como se enseña por Faham y col., patente de los EE.UU. 7.208.295. El tratamiento UDG escinde la sonda (2454) en los uracilos para proporcionar fragmentos (2455). Después de la sonda libre, se retiran los adaptadores y solapas (2476), cebadores directos (2466) y cebadores indirectos (2468) se añaden al producto de extensión (2464) y se lleva a cabo la PCR (2470), después de la cual se secuencia (2472) una muestra del amplicón resultante.

En una realización similar a la de la Fig. 2G, se pueden usar sondas y adaptadores similares a marcadores de secuencia unidos en sitios predeterminados de un polinucleótido diana, en donde una endonucleasa solapa, tal como FEN-1, se usa para escindir una porción unicatenaria que se corresponde con (2461). En esta realización además de una nucleasa distinta, la polaridad de la sonda y secuencias de adaptador se invierten; concretamente, un sustrato para una endonucleasa solapa requiere que el extremo 3' del adaptador que se corresponde con (2454) se hibride a una secuencia diana (2450) y que la porción unicatenaria que se corresponde con (2452) sea un extremo 5' de la secuencia diana. Después de la escisión y retiración de la secuencia de sonda, las etapas restantes son sustancialmente las mismas. Las directrices para usar endonucleasas solapa en ensayos de detección puede encontrarse en las siguientes referencias: Lyamichev y col., Nature Biotechnology , 17: 292-296 (1999); Eis y col., Nature Biotechnology , 19: 673-676 (2001); y referencias similares.

En algunas realizaciones, los ácidos nucleicos recombinados codifican cadenas de molécula de receptor inmunitario que forma normalmente un repertorio inmunitario que puede comprender un muy gran conjunto de polinucleótidos muy similares (por ejemplo, >1000, pero más de 10.000 y aún más normalmente de 100.000 a 1.000.000 o más) que pueden tener una longitud de menos de 500 nucleótidos o, en otras realizaciones, menos de 400 nucleótidos o, en aún otras realizaciones, menos de 300 nucleótidos. En un aspecto de la invención, los inventores reconocieron y apreciaron que estas características permitían el uso de marcadores de secuencia altamente diferentes para comparar eficazmente lecturas de secuencia de clonotipos altamente similares para determinar si derivan de la misma secuencia original o no.

Muestras

El término "muestra" se refiere a una cantidad de material biológico, que en algunas realizaciones se obtiene de un paciente y que contiene células y/o ADN libre de células; es decir, el término de usa de forma indistinta con el término "especimen", o "muestra de tejido". El término "muestra" también se usa a veces en un sentido estadístico de obtener un subconjunto o porción, de un conjunto o cantidad más grande, respectivamente, de, por ejemplo, ácidos nucleicos recombinados; en particular, el uso estadístico del término "muestra" también se puede entender que significa "muestra representativa", de modo que tal muestra se entiende que refleja, o es aproximada, a las frecuencias relativas de distintos ácidos nucleicos en un tejido (por ejemplo). Un experto en la técnica es capaz de distinguir el uso adecuado a partir del contexto de los términos.

Los perfiles de clonotipo se pueden obtener a partir de muestras de células o fluidos inmunitarios, tales como sangre, que contienen ácidos nucleicos sin células que codifican cadenas de receptor inmunitario. Por ejemplo, las células inmunitarias pueden incluir linfocitos T y/o linfocitos B. Los linfocitos T incluyen, por ejemplo, células que expresan receptores de linfocitos T. Las células T incluyen linfocitos T auxiliares (linfocitos T o linfocitos Th efectores), linfocitos T citotóxicos (LTC), linfocitos T de memoria y linfocitos T reguladores. En un aspecto, una muestra de linfocitos T incluye al menos 1.000 linfocitos T; pero más normalmente, una muestra incluye al menos 10.000 linfocitos T y, más normalmente, al menos 100.000 linfocitos T. En otro aspecto, una muestra incluye un número de linfocitos T en el intervalo de 1.000 a 1.000.000 de células. Una muestra de células inmunitarias también comprende linfocitos B. Los linfocitos B incluyen, por ejemplo, linfocitos B plasmáticos, linfocitos B de memoria, células B1, células B2, linfocitos B de zona marginal y linfocitos B foliculares. Los linfocitos B pueden expresar inmunoglobulinas (anticuerpos, receptor de linfocitos B). Como antes, en un aspecto una muestra de linfocitos B incluye al menos 1.000 linfocitos B; pero más normalmente, una muestra incluye al menos 10.000 linfocitos B y, más normalmente, al menos 100.000 linfocitos B. En otro aspecto, una muestra incluye un número de linfocitos B en el intervalo de 1.000 a 1.000.000 de linfocitos B.

Las muestras usadas en los método de la invención pueden provenir de una variedad de tejidos, que incluyen, por ejemplo, tejido tumoral, sangre y plasma sanguíneo, fluido linfático, líquido cefalorraquídeo que rodea el cerebro y la médula espinal, fluido sinovial que rodea las articulaciones ósea y similares. En una realización, la muestra es una muestra de sangre. La muestra de sangre puede ser de aproximadamente 0,1, 0,2, 0,3, 0,4, 0,5, 0,6, 0,7, 0,8, 0,9,

1,0, 1,5, 2,0, 2,5, 3,0, 3,5, 4,0, 4,5, o 5,0 m. La muestra puede ser una biopsia tumoral. La biopsia puede ser, de, por ejemplo, un tumor del cerebro, hígado, pulmón, corazón, colon, riñón o médula ósea. Cualquier técnica de biopsia usada por un experto en la técnica puede usarse para aislar una muestra del sujeto. Por ejemplo, una biopsia puede ser una biopsia abierta, en la que se usa anestesia general. La biopsia puede ser una biopsia cerrada, en la que se realiza un corte más pequeño que en una biopsia abierta. La biopsia puede ser una biopsia central o por incisión, en la que se retira parte del tejido. La biopsia puede ser una biopsia por escisión, en la que se intenta retirar una lesión completa. La biopsia puede ser una biopsia de aspiración por aguja fina, en la que se retira una muestra de tejido o fluido con una aguja.

En algunas realizaciones, los perfiles de clonotipo para los métodos de la invención se generan a partir de un tumor o sangre periférica en el caso de muestras de diagnóstico o de sangre periférica en el caso de muestras para controlar enfermedad residual. Uno o más clonotipos correlacionados con una enfermedad, tal como trastorno proliferativo linfoide o mieloide, se determinan a partir de una muestra de diagnóstico. Normalmente, el uno o más clonotipos correlacionados con un trastorno proliferativo linfoide o mieloide son los presentes en un perfil de clonotipo con las frecuencias más altas. En algunos casos, puede haber un único clonotipo correlacionado y, en otros casos, puede haber múltiples clonotipos correlacionados con un trastorno proliferativo linfoide o proliferativo. Se pueden tomar muestras tumorales a partir de cualquier tejido afectado por tal trastorno, que incluye ganglios linfáticos u otros tejidos fuera del sistema linfático. Como se ha mencionado anteriormente, los perfiles de clonotipo para controlar enfermedad residual pueden generarse a partir de una muestra de ácidos nucleicos extraídos a partir de sangre periférica. Los ácidos nucleicos de la muestra pueden ser de linfocitos B de una fracción que contiene células de la sangre periférica o a partir de un fracción sin células de la sangre periférica, tal como plasma o suero. En una realización, una muestra de sangre periférica incluye al menos 1.000 linfocitos B; pero más normalmente, tal muestra incluye al menos 10.000 linfocitos B y, más normalmente, al menos 100.000 linfocitos B. En otro aspecto, una muestra incluye un número de linfocitos B en el intervalo de 1.000 a 1.000.000 de linfocitos B. En algunas realizaciones, el número de células en una muestra establece un límite sobre la sensibilidad de una medición. Es decir, una mayor sensibilidad de detección de una enfermedad residual se logra usando una muestra más grande de sangre periférica. Por ejemplo, en una muestra que contiene 1.000 linfocitos B, la frecuencia más baja de clonotipo detectable es de 1/1.000 o ,001, independientemente de cuántas lecturas de secuenciación se obtienen cuando el ADN de tales células se analiza mediante secuenciación. Los ácidos nucleicos de la muestra pueden ser de linfocitos T de una fracción que contiene células de la sangre periférica o a partir de una fracción sin células de la sangre periférica, tal como plasma o suero. En una realización, una muestra de sangre periférica incluye al menos 1.000 linfocitos T; pero más normalmente, tal muestra incluye al menos 10.000 linfocitos T y, más normalmente, al menos 100.000 linfocitos T. En otro aspecto, una muestra incluye un número de linfocitos T en el intervalo de 1.000 a 1.000.000 de linfocitos T. En algunas realizaciones, el número de células en una muestra establece un límite sobre la sensibilidad de una medición. Es decir, una mayor sensibilidad de detección de una enfermedad residual se logra usando una muestra más grande de sangre periférica. Por ejemplo, en una muestra que contiene 1.000 linfocitos T, la frecuencia más baja de clonotipo detectable es de 1/1.000 o ,001, independientemente de cuántas lecturas de secuenciación se obtienen cuando el ADN de tales células se analiza mediante secuenciación.

Una muestra para su uso con la invención puede incluir ADN (por ejemplo, ADN genómico) o ARN (por ejemplo, ARN mensajero). El ácido nucleico puede ser ADN o ARN sin células, por ejemplo, extraido del sistema circulatorio, Vlassov y col., Curr. Mol. Med., 10: 142-165 (2010); Swarup y col., FEBS Lett., 581: 795-799 (2007). En los métodos de la invención que se proporciona, la cantidad de ARN o ADN de un sujeto que puede analizarse incluye, por ejemplo, tan bajo como una única célula en algunas aplicaciones (por ejemplo, ensayo de calibración con otros criterios de selección celular, por ejemplo, criterios morfológicos) y tantas como 10 millones de células o más, que se traduce en una cantidad de ADN en el intervalo de 6 pg-60 ug y una cantidad de ARN en el intervalo de 1 pg-10 ug. En algunas realizaciones, la muestra de ácido nucleico es una muestra de ADN de 6 pg a 60 ug. En otras realizaciones, la muestra de ácido nucleico es una muestra de ADN de 100 µl a 10 ml de sangre periférica; en otras realizaciones, una muestra de ácido nucleico es una muestra de ADN de una fracción sin células de 100 µl a 10 ml de sangre periférica.

En algunas realizaciones, la muestra de linfocitos o ácido nucleico libre es suficientemente grande para que sustancialmente cada linfocito B o linfocito T con un clonotipo distinto se represente en este, formando, de este modo, un "repertorio" de clonotipos. En una realización, para conseguir una representación sustancial de cada clonotipo distinto, se toma una muestra que contiene una probabilidad del noventa y nueve por ciento cada clonotipo de una población presente a una frecuencia de ,001 por ciento o superior. En otra realización, se toma una muestra que contiene una probabilidad del noventa y nueve por ciento cada clonotipo de una población presente a una frecuencia de ,0001 por ciento o superior. Y, en otra realización, se toma una muestra que contiene una probabilidad del noventa y nueve por ciento cada clonotipo de una población presente a una frecuencia de ,00001 por ciento o superior. En una realización, una muestra de linfocitos B o linfocitos T incluye al menos medio millón de células y, en otra realización, tal muestra incluye al menos un millón de células.

Las muestras de ácido nucleico pueden obtenerse de sangre periférica usando técnicas convencionales, por ejemplo, Innis y col., editors, PCR Protocols (Academic Press, 1990); o similar. Por ejemplo, se pueden separar glóbulos blancos de muestras de sangre usando técnicas convencionales, por ejemplo, RosetteSep kit (Stem Cell Technologies, Vancouver, Canadá). Las muestras de sangre pueden variar en un volumen de 100 µl a 10 ml; en un

aspecto, los volúmenes de muestra de sangre se encuentran en el intervalo de 100 µl a 2 ml. A continuación, puede extraerse el ADN y/o ARN de tal muestra de sangre usando técnicas convencionales para su uso en métodos de la invención, por ejemplo, DNeasy Blood & Tissue Kit (Qiagen, Valencia, CA). Opcionalmente, los subconjuntos de glóbulos blancos, por ejemplo, linfocitos, pueden aislarse adicionalmente usando técnicas convencionales, por ejemplo, clasificación de células activadas por fluorescencia (FACS)(Becton Dickinson, San Jose, CA), clasificación de células magnéticamente activadas (MACS)(Miltenyi Biotec, Auburn, CA) o similares. Por ejemplo, los linfocitos B de memoria pueden aislarse a modo de marcadores de superficie CD19 y CD27.

También puede extraerse ADN sin células de muestras de sangre periférica usando técnicas convencionales, por ejemplo, Lo y col., patente de los EE.UU. 6.258.540; Huang y col., Methods Mol. Biol., 444: 203-208 (2008); y similares. A modo de ejemplo no limitante, puede recogerse sangre periférica en tubos EDTA, después de lo cual puede fraccionarse en plasma, componentes de glóbulos blancos y glóbulos rojos mediante centrifugación. ADN de la fracción plasmática libre de células (por ejemplo, de 0,5 a 2,0 ml) puede extraerse usando un Kit QIAamp DNA Blood Mini Kit (Qiagen, Valencia, CA) o un kit similar, de acuerdo con el protocolo del fabricante.

En un aspecto, una muestra de linfocitos para generar un perfil de clonotipo es suficientemente grande para que sustancialmente cada linfocito T o linfocito B con un clonotipo distinto se represente en este. En una realización, se toma una muestra que contiene una probabilidad del noventa y nueve por ciento cada clonotipo de una población presente a una frecuencia de ,001 por ciento o superior. En otra realización, se toma una muestra que contiene una probabilidad del noventa y nueve por ciento cada clonotipo de una población presente a una frecuencia de ,0001 por ciento o superior. En otra realización, se toma una muestra que contiene una probabilidad del noventa y nueve por ciento cada clonotipo de una población presente a una frecuencia de ,00001 por ciento o superior. En otras realizaciones, se toma una muestra que contiene una probabilidad del noventa y cinco por ciento cada clonotipo de una población presente a una frecuencia de ,001 por ciento o superior. En otra realización, se toma una muestra que contiene una probabilidad del noventa y cinco por ciento cada clonotipo de una población presente a una frecuencia de ,0001 por ciento o superior. En otra realización, se toma una muestra que contiene una probabilidad del noventa y cinco por ciento cada clonotipo de una población presente a una frecuencia de ,00001 por ciento o superior. En otra realización adicional más, una muestra de linfocitos B o linfocitos T incluye al menos medio millón de células y, en otra realización, tal muestra incluye al menos un millón de células.

Cuando la fuente de un material de cual se toma una muestra sea pobre, tal como, muestras de estudio clínico o similares, el ADN del material puede amplificarse mediante una técnica de no sesgado, tal como amplificación de genoma completo (WGA), amplificación de desplazamiento múltiple (MDA); u otra técnica similar, por ejemplo, Hawkins y col., Curr. Opin. Biotech., 13: 65-67 (2002); Dean y col., Genome Research, 11: 1095-1099 (2001); Wang y col., Nucleic Acids Research, 32: e76 (2004); Hosono y col., Genome Research, 13: 954-964 (2003); y similares.

Las muestras de sangre son de particular interés y pueden obtenerse usando técnicas convencionales, por ejemplo, Innis y col., editors, PCR Protocols (Academic Press, 1990); o similar. Por ejemplo, se pueden separar glóbulos blancos de muestras de sangre usando técnicas convencionales, por ejemplo, RosetteSep kit (Stem Cell Technologies, Vancouver, Canadá). Las muestras de sangre pueden variar en un volumen de 100 µl a 10 ml; en un aspecto, los volúmenes de muestra de sangre se encuentran en el intervalo de 100 µl a 2 ml. A continuación, puede extraerse el ADN y/o ARN de tal muestra de sangre usando técnicas convencionales para su uso en métodos de la invención, por ejemplo, DNeasy Blood & Tissue Kit (Qiagen, Valencia, CA). Opcionalmente, los subconjuntos de glóbulos blancos, por ejemplo, linfocitos, pueden aislarse adicionalmente usando técnicas convencionales, por ejemplo, clasificación de células activadas por fluorescencia (FACS)(Becton Dickinson, San Jose, CA), clasificación de células magnéticamente activadas (MACS)(Miltenyi Biotec, Auburn, CA) o similares.

Puesto que la identificación de recombinaciones está presente en el ADN de cada célula de inmunidad adaptativo del individuo así como sus transcripciones de ARN asociadas, cualquier ARN o ADN puede secuenciarse en los métodos de la invención que se proporciona. Una secuencia recombinada de un linfocito T o linfocito B que codifica una cadena de un receptor de linfocitos T o molécula de inmunoglobulina o una porción de la misma, se hace referencia como un clonotipo. El ADN o ARN puede corresponder a secuencias de genes de receptor de linfocitos T (RLT) o genes de inmunoglobulina (Ig) que codifican anticuerpos. Por ejemplo, el ADN o ARN puede corresponderse con secuencias que codifican cadenas α, β, γ, o δ de un RLT. En la mayoría de linfocitos T, el RLT es un heterodímero que consiste en una cadena α y una cadena β. La cadena de RLTα se genera mediante recombinación VJ y el receptor de cadena β se genera mediante recombinación V(D)J. Para la cadena RLTβ, en los humanos hay 48 segmentos V, 2 segmentos D y 13 segmentos J. Se pueden suprimir varias bases y otras añadir (denominados nucleótidos N y P) en cada de las dos uniones. En una minoría de linfocitos T, los RLT consisten en cadenas delta γ y δ. La cadena de RLT γ se genera mediante recombinación VJ y la cadena de RLT δ se genera mediante recombinación V(D)J (Kenneth Murphy, Paul Travers, and Mark Walport, Janeway's Immunology 7ª edición, Garland Science, 2007).

El ADN y ARN analizado en los métodos de la invención puede corresponderse con secuencias que codifican inmunoglobulinas de cadena pesada (IgH) con regiones constantes (α, δ, ε, γ, o µ) o inmunoglobulinas de cadena ligera (IgK o IgL) con regiones constantes λ o k. Cada anticuerpo tiene dos cadenas ligeras idénticas y dos cadenas pesadas idénticas. Cada cadena está compuesta de una región constante (C) y una variable. Para cada cadena

pesada, la región variable está compuesta de una variable (V), diversidad (D) y segmentos de unión (J). Varias secuencias distintas que codifican cada tipo de estos segmentos están presentes en el genoma. Un evento de recombinación VDJ específico se produce durante el desarrollo de un linfocito B, haciendo que esa célula genere una cadena pesada específica. La diversidad en la cadena ligera se genera de un modo similar excepto en que no hay región D de modo que solo es una recombinación VJ. La mutación somática a menudo se produce cerca del sitio de recombinación, provocando la adición o supresión de varios nucleótidos, aumentando adicionalmente la diversidad de cadenas pesadas y ligeras generadas por los linfocitos B. La posible diversidad de los anticuerpos generados por un linfocito B es entonces el producto de las distintas cadenas pesadas y ligeras. Las regiones variables de las cadenas pesadas y ligeras contribuyen a formar el sitio o región de reconocimiento de antígeno (o unión). Añadida a esta diversidad hay un proceso de hipermutación somática que puede producirse después de que se haya montado una respuesta específica frente a algún epítopo.

Como se ha mencionado anteriormente, de acuerdo con la invención, se pueden seleccionar cebadores para generar amplicones que contienen porciones de ácidos nucleicos recombinados de linfocitos o de ácidos nucleicos sin células a partir de un tejido, tal como sangre. Tales porciones se pueden hacer referencia en el presente documento como "regiones somáticamente reorganizadas". Las regiones somáticamente reorganizadas pueden comprender ácidos nucleicos de linfocitos en desarrollo o completamente desarrollados, donde los linfocitos en desarrollo son células en la que la reorganización de genes inmunitarios no se ha completado para formar moléculas que tengan (por ejemplo) regiones V(D)J completas. Regiones somáticamente reorganizadas incompletas ejemplares incluyen moléculas de IgH incompletas (tales como, moléculas que contienen solo regiones D_J), moléculas de RLTδ incompletas como, moléculas que contienen solo regiones D-JJ) y IgK inactivo (por ejemplo que comprende regiones Kde-V).

Amplificación de poblaciones de ácidos nucleicos

En algunas realizaciones, las secuencias de cebador del primer y segundo conjunto de cebadores pueden seleccionarse de acuerdo con reacciones en cadena de la polimerasa (PCR) de multiplexación. Por ejemplo, directrices para seleccionar cebadores y para llevar a cabo PCR de multiplexación de ácidos nucleicos que codifican diversas cadenas de receptores inmunitarios se encuentran en las siguientes referencias: Faham y Willis, patentes de Estados Unidos 8.236.503 y 8.628.927; Morley, patente de los EE.UU. 5.296.351; Gorski, patente de los EE.UU. 5.837.447; Dau, patente de los EE.UU. 6.087.096; Van Dongen y col., publicación de patente de los EE.UU. 2006/0234234; publicación de patente europea EP 1544308B1; Van Dongen y col., Leukemia, 17: 2257-2317 (2003); y similares, Directrices para PCR de multiplexación pueden encontrarse en Henegariu y col., BioTechniques, 23: 504-511 (1997) y referencias similares. En algunas realizaciones, los cebadores se seleccionan de modo que las frecuencias de secuencias amplificadas en un producto final son sustancialmente las mismas que las frecuencias de las secuencias en la mezcla de reacción de partida. Tal selección de cebadores puede incluir la sección de las longitudes del cebador, los sitios de unión del cebador y las concentraciones dle cebador. Tal como se ha señalado anteriormente, dependiendo de los métodos seleccionados para generar lecturas de secuencias y marcadores de secuencia unidos, el nivel de multiplexación puede variar en gran medida.

En algunas realizaciones, una etapa de amplificación de ácidos nucleicos diana incluye la amplificación lineal de ácidos nucleicos diana, tal como, por ejemplo, mediante ciclos repetidos de hibridación de un conjunto de cebadores (por ejemplo, un primer conjunto de cebadores "corriente arriba" o "directos"), extendiendo los cebadores, fundiendo la cadena extendida del modelo, de modo que la cantidad de cadenas extendidas se amplifica como una función lineal del número de ciclos. En otras palabras, una etapa de amplificación incluye copiar un polinucleótidos diana (es decir, al menos una cadena de un polinucleótido diana) mediante extensiones repetidas de un conjunto de cebadores. En algunas realizaciones, tal extensión única o repetida en una dirección puede seguirse por etapas de retiración de cebadores no extendidos y una extensión única o repetida de otro conjunto de cebadores en la otra dirección (por ejemplo, como segundo conjunto de cebadores "corriente abajo" o "indirectos").

El número de cebador en el primer conjunto de cebadores y un segundo conjunto de cebadores puede variar ampliamente dependiendo del número y tipo de ácidos nucleicos de cadena de receptores inmunitarios que se amplifiquen en un ensayo. En algunas realizaciones, pueden usarse cebadores de consenso para diversas cadenas. En otras realizaciones, pueden designarse cebadores específicos para cada polinucleótido diana distinto amplificado. Normalmente, tanto el primer como el segundo conjunto de cebadores comprende cada uno una pluralidad de cebadores. En algunas realizaciones, la pluralidad de cebadores en el primer conjunto o el segundo conjunto de cebadores es de al menos 50 cebadores; en otras realizaciones, la pluralidad de cebadores en el primer conjunto o el segundo conjunto de cebadores es de al menos 100 cebadores; en otras realizaciones, la pluralidad de cebadores en el primer conjunto o el segundo conjunto de cebadores es de al menos 150 cebadores; en otras realizaciones, la pluralidad de cebadores en el primer conjunto o el segundo conjunto de cebadores es de al menos 200 cebadores; en otras realizaciones, la pluralidad de cebadores en el primer conjunto o el segundo conjunto de cebadores es de al menos 250 cebadores. El número de cebadores en el primer conjunto puede ser el mismo o distinto del número de cebadores en el segundo conjunto.

En algunas realizaciones, los cebadores del primer conjunto y el segundo conjunto se seleccionan de modo que la longitud de los clonotipos es de al menos 30 nucleótidos; en otras realizaciones, los cebadores del primer conjunto y

el segundo conjunto se seleccionan de modo que la longitud de clonotipos se encuentra en el intervalo de 30 a 500 nucleótidos; en otras realizaciones, los cebadores del primer conjunto y el segundo conjunto se seleccionan de modo que la longitud de los clonotipos se encuentra en l intervalo de 30 a 400 nucleótidos. En otras realizaciones, los cebadores del primer conjunto y el segundo conjunto se seleccionan de modo que la longitud de clonotipos se encuentra en el intervalo de 30 a 300 nucleótidos; en otras realizaciones, los cebadores del primer conjunto y el segundo conjunto se seleccionan de modo que la longitud de los clonotipos se encuentra en el intervalo de 30 a 200 nucleótidos.

Se pueden encontrar protocolos de amplificación por PCR ejemplares en Dongen y col., Leukemia, 17: 2257-2317 (2003) o van Dongen y col., publicación de patente de los EE.UU. 2006/0234234. En resumen, un protocolo ejemplar es el siguiente: Tampón de reacción: Tampón ABI II o tampón dorado ABI (Life Technologies, San Diego, CA); 50 µl de volumen de reacción final; 100 ng de ADN de muestra; 10 pmol de cada cebador (sujeto a ajustes para equilibrar la amplificación tal como se describe a continuación); dNTP a una concentración final de 200 µM; $MgCl_2$ a una concentración final de 1,5 mM (sujeto a optimización dependiendo de las secuencias diana y polimerasa); polimerasa Taq (1-2 U/tubo); condiciones de ciclo: preactivación 7 min a 95 ºC; hibridación a 60 ºC; tiempos del ciclo: 30 s desnaturalización; 30 s hibridación; 30 s extensión. Las polimerasas que se pueden usar para la amplificación en los métodos de la invención están comercialmente disponibles e incluyen, por ejemplo, polimerasa Taq, polimerasa AccuPrime o Pfu. La elección de la polimerasa a usar puede basarse en si se prefiere fidelidad o eficacia.

PCR en tiempo real, tinción de picogreen, electroforesis nanofluídica (por ejemplo, LabChip) o mediciones de absorción UV pueden usarse como etapa inicial para evaluar la cantidad funcional de material amplificable en una muestra.
En un aspecto, las amplificaciones de multiplexación de la invención se llevan a cabo de modo que las cantidades relativas de secuencias en una población de partida son sustancialmente las mismas que las de en la población amplificada o amplicón. Es decir, las amplificaciones de multiplexación se llevan a cabo con un sesgado de amplificación mínimo entre secuencias miembro de una población de muestra. En una realización, tales cantidades relativas son sustancialmente las mismas si cada cantidad relativa de un amplicón se encuentra dentro de cinco veces su valor en la muestra de partida. En otra realización, tales cantidades relativas son sustancialmente las mismas si cada cantidad relativa de un amplicón se encuentra dentro de dos veces su valor en la muestra de partida. Como se describe más completamente a continuación, el sesgo de amplificación en PCR puede detectarse y corregirse usando técnicas convencionales de modo que un conjunto de cebadores de PCR puede seleccionarse para un repertorio predeterminado que proporciona amplificación no sesgada de cualquier muestra.

En algunas realizaciones, el sesgo de amplificación puede evitarse llevando a cabo una amplificación de dos etapas (por ejemplo, tal como se describe en Fahan y Willis, citados anteriormente), en donde una pequeña cantidad de ciclos de amplificación (por ejemplo, 2-5 o 2-10, o 2-15 ciclos) se implementan en una primera o primaria, etapa usando cebadores que tienen colas no complementarias con las secuencias diana. Las colas incluyen sitios de unión de cebador que se añaden a los extremos de las secuencias del amplicón primario de modo que tales sitios se usan en una segunda etapa de amplificación que usa solo un cebador único directo y un cebador único indirecto, eliminando, de este modo, una causa primaria del sesgo de amplificación. Antes del inicio de la segunda etapa de amplificación, los cebadores no extendidos de la primera etapa se retiran de la mezcla de reacción o se inactivan de otro modo. En algunas realizaciones, la PCR primaria tendrá un número suficientemente pequeño de ciclos (por ejemplo, 2-10) para minimizar la amplificación diferencial mediante los distintos cebadores. Entonces, se lleva a cabo la amplificación secundaria con un par de cebadores, que elimina una fuente de amplificación diferencial. En algunas realizaciones, un pequeño porcentaje o porción, por ejemplo, uno por ciento del volumen de reacción, de la PCR primaria se doma directamente a la mezcla de reacción de PCR secundaria. En algunas realizaciones, un total de al menos treinta y cinco ciclos emplazados entre una primera etapa de amplificación y una segunda etapa de amplificación

En algunas realizaciones se pueden combinar estándares internos con amplificados en la misma reacción que los ácidos nucleicos recombinados de una muestra. El estándar interno son ácidos nucleicos con secuencias conocidos y concentraciones conocidas. Por ejemplo, pueden ser copias clonadas de un ácido nucleico natural que codifica porciones de una cadena de receptor inmunitario o pueden ser ácidos nucleicos sintéticos. En algunas realizaciones, las longitudes y composiciones de base de los estándares internos se seleccionan para que sen representativas de las cadenas de receptor inmunitario particular que se está amplificando. Al controlar los cambios en las concentraciones relativas de los estándares internos tras su amplificación, se puede detectar el sesgo de amplificación y se pueden determinar las condiciones para una amplificación no sesgada. Por ejemplo, las longitudes del cebador, posiciones y concentraciones pueden variarse para minimizar el sesgo en el producto de amplificación. En algunas realizaciones, se usa una pluralidad de estándares internos en una reacción; en algunas realizaciones, se usan de 2 a 50 estándares internos distintos en una reacción; en otras realizaciones, se usan de 2 a 25 estándares internos distintos en una reacción; y en algunas realizaciones, se usan de 2 a 10 estándares internos distintos en una reacción. En algunas realizaciones, el sesgo de amplificación se determina midiendo las frecuencias relativas de las secuencias de distintos nucleótidos diana (por ejemplo, todos los clonotipos o los seleccionados o estándares internos) en un producto de amplificación. En otras realizaciones, la presencia, ausencia o nivel de sesgo de amplificación puede determinarse mediante PCR cuantitativa en tiempo real de ácidos nucleicos seleccionados, tales como dos o más de los estándares internos. Los estándares internos también se pueden usar para cuantificar

los números de distintos clonotipos en la muestra original. Técnicas para tal recuento molecular son bien conocidas, por ejemplo Brenner y col., patente de los EE.UU. 7.537.897.

Generación de lecturas de secuencia

Cualquier técnica de alto rendimiento para secuenciar ácidos nucleicos puede usarse en el método de la invención. Preferentemente, tal técnica tiene la capacidad de generar de un modo económico un volumen de datos de secuencia a partir de los cuales al menos 1.000 clonotipos pueden determinarse y, preferentemente, a partir de los cuales al menos 10.000 o 1.000.000 clonotipos pueden determinarse. Las técnicas de secuenciación de ADN incluyen reacciones de secuenciación didesoxi clásica (método Sanger) usando terminadores o cebadores marcados y separación por gel en plancha o capilares, secuenciación mediante síntesis usando nucleótidos marcados inversamente terminados, pirosecuenciación, secuenciación 454, hibridación específica de alelos con respecto a una librería de sondas de oligonucleótidos marcadas, secuenciación mediante síntesis usando hibridación específica de alelos con respecto a una librería de clones marcados que está seguida de ligación, control en tiempo real o la incorporación de nucleótidos marcados durante una etapa de polimerización, secuenciación de polonia y secuenciación SOLiD. La secuenciación de las moléculas separadas ha demostrado más recientemente mediante reacciones de extensión secuencial o única usando polimerasas o ligasas así como mediante hibridaciones diferenciales únicas o secuenciales con bibliotecas de sondas. Estas reacciones se han llevado a cabo sobre muchas secuencias clonales en paralelo incluyendo demostraciones en aplicaciones comerciales actuales de sobre 100 millones de secuencias en paralelo. Estos enfoques de secuenciación pueden, de este modo, usarse para estudiar el repertorio de receptor de linfocitos T (RLT) y/o receptor de linfocitos B (RLB). En un aspecto de la invención, los métodos de alto rendimiento de secuenciación se emplean que comprenden una etapa de aislamiento espacial de moléculas individuales sobre una superficie sólida donde se secuencian en paralelo. Tales superficies sólidas pueden incluir superficies no porosas (tales como en secuenciación Solexa, por ejemplo, Bentley y col., Nature, 456: 53-59 (2008) o secuenciación genómica completa, por ejemplo, Drmanac y col., Science, 327: 78-81 (2010)), matrices de pocillos, que pueden incluir modelos unidos a perlas o partículas (tales como con 454, por ejemplo, Margulies y col., Nature, 437: 376-380 (2005) o secuenciación de Ion Torrent, publicación de patente de los EE.UU. 2010/0137143 o 2010/0304982), membranas micromecanizadas (tales como secuenciación SMRT, Eid y col., Science, 323: 133-138 (2009)) o matrices de perlas (como con secuenciación SOLiD o secuenciación de polonia, por ejemplo, Kim y col., Science, 316: 1481-1414 (2007)). En otro aspecto, tales métodos comprenden la amplificación de las moléculas aisladas o bien antes o bien después de que se aíslen espacialmente sobre una superficie sólida. Antes de la amplificación puede comprender la amplificación a base de emulsión, tal como PCR de emulsión o amplificación de círculo rodante.

Son de particular interés enfoques que usan secuenciación mediante síntesis con terminadores reversibles, tales como secuenciación a base de SOlexa donde las moléculas de modelo individuales se aíslan espacialmente sobre una superficie sólida, después de lo cual se amplifican en paralelo mediante PCR en puente para formar poblaciones clonales separadas, o clústeres y, a continuación, se secuencian, tal como describe Bentley y col., (citado anteriormente) y en las instrucciones del fabricante (por ejemplo, Hoja de cálculo y Kit de preparación de TruSeq™, Illumina, San Diego, CA, 2010); y adicionalmente en las siguientes referencias: patentes de los EE.UU. 6.090.592; 6.300.070; 7.115.400; y documento EP0972081B1. En una realización, moléculas individuales dispuestas y amplificadas sobre una superficie sólida forma clústeres en una densidad de al menos 10^5 clústeres por cm^2; o en una densidad de al menos $5x10^5$ por cm^2; o en una densidad de al menos 10^6 clústeres por cm^2. La secuenciación a base de Solexa también proporciona la capacidad de generar dos lecturas de secuencia a partir de la misma secuencia diana (o modelo) en un clúster, una lectura de secuencia a partir de extremos opuestos de una secuencia diana. En algunas realizaciones, tales pares de lecturas de secuencia pueden combinarse y tratarse como una única lectura de secuencia en análisis posteriores, o tales pares pueden tratarse por separado pero teniendo en cuenta que se originan a partir del mismo clúster. Algunas veces el par de lecturas de secuencia a partir del mismo modelo se denominan "pares mate", y el proceso de secuenciación de ambos extremos de un modelo se denomina secuenciación "bidireccional". En algunas realizaciones, una etapa de secuenciación mediante síntesis usando nucleótidos marcados inversamente terminados incluye la generación de una única lectura de secuencia para cada clúster o población clonal de modelos y la generación de una pluralidad de lecturas de secuencia (incluidas pero no limitadas a pares mate) para cada clúster o población clonal de modelos. En todavía otras realizaciones, cuando se genera una pluralidad de lecturas de secuencia para cada clúster o población clonal de modelos, tal pluralidad de lecturas de secuencia puede combinarse para formar una única lectura de secuencia eficaz que se usa en análisis posteriores, tal como una etapa de fusión.

En un caso, un perfil de clonotipo a base de secuencia de una muestra de un individuo se obtiene usando las siguientes etapas: (a) obtener una muestra de ácido nucleico de linfocitos T y/o linfocitos B del individuo; (b) aislar espacialmente moléculas individuales derivadas de tal muestra de ácido nucleico, comprendiendo las moléculas individuales al menos un modelo generado a partir de un ácido nucleico en la muestra, cuyo modelo comprende una región somáticamente reorganizada o una porción de la misma, siendo cada molécula individual capaz de producir al menos una lectura de secuencia; (c) secuenciar dichas moléculas individuales espacialmente aisladas; y (d) determinar las abundancias de distintas secuencias de las moléculas de ácido nucleico a partir de la muestra de ácido nucleico para generar un perfil de clonotipo. En una realización, cada una de las regiones somáticamente reorganizadas comprende una región V y una región J. En otra realización, la etapa de secuenciación incluye la

generación de una pluralidad de lecturas de secuencia para cada clonotipo determinado. En todavía otras realizaciones, la etapa de secuenciación incluye la combinación de información o datos a partir de una pluralidad de lecturas de secuencia para formar cada clonotipo. En algunas realizaciones, tal etapa de combinación puede llevarse a cabo fusionando lecturas de secuencia tal como se describe en Fahan y WIllis, patente de los EE.UU. 8.628.927 o usando marcadores de secuencia como se describe en Faham y col., publicación de patente de los EE.UU. 2013/0236895A1. En otra realización, la etapa de secuenciación comprende secuenciar bidireccionalmente cada una de las moléculas individuales espacialmente aisladas para producir al menos una lectura de secuencia directa y al menos una lectura de secuencia indirecta.

Además de la última realización, al menos una de las lecturas de secuencia directas y al menos una de las lecturas de secuencia indirectas tienen una región de solapamiento de modo que las bases de tal región de solapamiento se determinan mediante una relación complementaria inversa entre tales lecturas de secuencia. En otra realización adicional más, cada una de las regiones somáticamente reorganizadas comprende una región V y una región J y la etapa de secuenciación incluye adicionalmente la determinación de una secuencia de cada una de las moléculas de ácido nucleico individuales a partir de una o más de sus lecturas de secuencia directas y al menos una lectura de secuencia indirecta que parte de una posición en una región J y se extiende en la dirección de su región V asociada. En otra realización, las moléculas individuales comprenden ácidos nucleicos seleccionados entre el grupo que consiste en moléculas de IgH completas, moléculas de IgH incompletas, IgK completo, moléculas de IgK inactivas, moléculas de RLTβ, moléculas de RLTγ, moléculas de RLTδ completas y moléculas de RLTδ incompletas. En otro caso, la etapa de secuenciación comprende generar las lecturas de secuencia que tienen puntuaciones de calidad monotónicamente en disminución. En otra realización, el anterior método comprende las siguientes etapas: (a) obtener una muestra de ácido nucleico de linfocitos T y/o linfocitos B del individuo; (b) aislar espacialmente moléculas individuales derivadas de tal muestra de ácido nucleico, comprendiendo las moléculas individuales conjuntos anidados de modelos generados a partir de un ácido nucleico en la muestra y cada una conteniendo una región somáticamente reorganizada o una porción de la misma, cada conjunto anidado siendo capaz de producir una pluralidad de lecturas de secuencia cada una extendiéndose en la misma dirección y cada una empezando desde una posición distinta sobre el ácido nucleico desde la cual el conjunto anidado se generó; (c) secuenciar dichas moléculas individuales espacialmente aisladas; y (d) determinar las abundancias de distintas secuencias de las moléculas de ácido nucleico a partir de la muestra de ácido nucleico para generar un perfil de clonotipo. En una realización, la etapa de secuenciación incluye la producción de una pluralidad de lecturas de secuencia para cada uno de los conjuntos anidados. En otra realización, cada una de las regiones somáticamente reorganizadas comprende una región V y una región J, y cada una de la pluralidad de lecturas de secuencia se inicia desde una posición distinta en la región V y se extiende en la dirección de su región asociada J.

En un aspecto, para cada muestra de un individuo, la técnica de secuenciación usada en los métodos de la invención genera secuencias de al menos 1.000 clonotipos por tirada; en otro aspecto, tal técnica genera secuencias de al menos 10.000 clonotipos por tirada; en otro aspecto, tal técnica genera secuencias de al menos 100.000 clonotipos por tirada; en otro aspecto, tal técnica genera secuencias de al menos 500.000 clonotipos por tirada; y en otro aspecto, tal técnica genera secuencias de al menos 1.000.000 clonotipos por tirada. En otro aspecto adicional, tal técnica genera secuencias de entre 100.000 a 1.000.000 clonotipos por tirada por muestra individual. En cada una de las anteriores, cada clonotipo por tirada se determina a partir de al menos 10 lecturas de secuencia.

La técnica de secuenciación usada en los métodos de la invención que se proporciona pueden generar aproximadamente 30 pb, aproximadamente 40 pb, aproximadamente 50 pb, aproximadamente 60 pb, aproximadamente 70 pb, aproximadamente 80 pb, aproximadamente 90 pb, aproximadamente 100 pb, aproximadamente 110, aproximadamente 120 pb por lectura, aproximadamente 150 pb, aproximadamente 200 pb, aproximadamente 250 pb, aproximadamente 300 pb, aproximadamente 350 pb, aproximadamente 400 pb, aproximadamente 450 pb, aproximadamente 500 pb, aproximadamente 550 pb o aproximadamente 600 pb por lectura.

Determinación de clonotipo a partir de datos de secuencia

En la invención, los marcadores de secuencia en combinación con una etapa de fusión de lectura de secuencia se usan para determinar clonotipos. En realizaciones en las que un único de marcador único se una a sustancialmente cada polinucleótido diana distinto, la determinación de clonotipo usando marcadores de secuencia no es directa. En dichas realizaciones, los clonotipos de una muestra se determinan, en primer lugar, agrupando lecturas de secuencia basadas en sus marcadores de secuencia. Tal agrupación puede conseguir mediante métodos de alineación de secuencias convencionales. Directrices para seleccionar los métodos de alineación están disponibles en Batzoglou, Briefings in Bioinformatics, 6: 6-22 (2005). Después de ensamblar las lecturas de secuencia en grupos que se corresponden con marcadores de secuencia únicos, entonces, las secuencias de los clonotipos asociados pueden analizarse para determinar la secuencia del clonotipo a partir de la muestra. La Fig. 4A ilustran una alineación ejemplar y método a partir de la determinación de la secuencia (SEQ ID NO: 2) de un clonotipo asociado con un marcador de secuencia único. En este ejemplo, se alinean once lecturas de secuencia a modo de sus respectivos marcadores de secuencia (4302) después de los cuales se comparan los nucleótidos en cada posición de las porciones del clonotipo (4304) de las lecturas de secuencia, indicadas como 1, 2, 3, 4, ... n. Por ejemplo, los nucleótidos en la posición 6 (4306) son t, t, g, t, t, t, t, t, t, c, t; es decir, nueve identificaciones de nucleótidos son t,

una es "g" (4308) y una es "c" (4310) (SEQ ID NO: 3 y SEQ ID NO: 4). En una realización, la correcta identificación de nucleótidos de la secuencia de clonotipo en la posición es cualquiera sea la identidad de la mayoría de base. En el ejemplo de la posición 6 (4306), la identificación de nucleótidos es "t", puesto que es el nucleótido en la mayoría de las lecturas de secuencia en esa posición. En otras realizaciones, otros factores se pueden tener en cuenta para determinar una correcta identificación de nucleótidos para una secuencia de clonotipo, tal como puntuaciones de calidad de las identificaciones de nucleótidos de las lecturas de secuencia, identidades de bases adyacentes o similares. Una vez se han determinado los clonotipos tal como se ha descrito anteriormente, puede ensamblarse un perfil de clonotipo que comprende las abundancias o frecuencias de cada clonotipo distinto de una muestra.

En algunas realizaciones, se puede llevar a cabo más de una etapa de extensión usando cebadores que contienen marcador de secuencia para aumentar la fracción de polinucleótidos diana en una muestra que están marcados con marcadores de secuencia antes de su amplificación. En dichas realizaciones, la más de una etapa de extensión en presencia de los cebadores que contienen marcador de secuencia da como resultado un polinucleótido diana y/o sus copias que están marcadas con una pluralidad de distintos marcadores de secuencia. El tamaño de la pluralidad depende del número de etapas de extensión llevadas a cabo en la presencia de los cebadores que contienen marcador de secuencia, la eficacia de la reacción de amplificación, aunque solo uno o ambos de los cebadores directos e indirectos tengas marcadores de secuencia y similares. En algunas de tales realizaciones, la pluralidad se encuentra en el intervalo de 2 a 15, o en el intervalo de 2 a 10, o en el intervalo de 2 a 5. En algunas de tales realizaciones, después de la amplificación, las copias de cada polinucleótido diana de una muestra pueden dividirse en una pluralidad de grupos o subconjuntos en donde los miembros de cada grupo o subconjunto se marca con el mismo marcador de secuencia y los miembros de cada miembro distinto o subconjunto se marca con un marcador de secuencia distinto; es decir, los miembros del mismo grupo tienen el mismo marcador de secuencia y los miembros de distintos grupos tienen distintos marcadores de secuencia. En otras palabras, después de la amplificación, en algunas realizaciones, cada copia de un polinucleótido diana a partir de una muestra se marcará con uno de los dos marcadores de secuencia distintos; o, en otras realizaciones, cada copia de un polinucleótido diana a partir de un amuestra se marcará con uno de los tres marcadores de secuencia distintos; o, en otras realizaciones, las copias de un polinucleótido diana a partir de una muestra se marcará con uno de los cuatro marcadores de secuencia distintos; etcétera. En estas realizaciones, los clonotipos pueden determinarse mediante una combinación de alineación de marcador de secuencia seguido por etapas de fusión para tratar lecturas de secuencia dentro de un grupo como que se origina de la misma secuencia pariente basándose en una probabilidad que es de que el origen común es verdadero como fusión de tasas de error, frecuencias relativas y similares. La Fig. 4B ilustra lecturas de secuencia a partir de tal realización. En un enfoque, las lecturas de secuencia se agrupan, en primer lugar, mediante marcadores de secuencia comunes (4402) que en la ilustración resultan en tres grupos (4420), (4422) y (4424). En algunas realizaciones, dentro de cada grupo, las secuencias (4404) se analizan para determinar una secuencia de consenso del grupo; por ejemplo, como antes, en cada posición de nucleótido puede identificarse una base según la mayoría de base o la base de frecuencia más alta, o similares. Las secuencias de consenso del grupo pueden, entonces, fusionarse entre sí para determinar los clonotipos.

En algunas realizaciones, el anterior aspecto de la invención puede implementarse en un método para elaborar perfiles de prácticamente cualquier población de ácidos nucleicos en una muestra. Tal método puede comprender las etapas de: (a) obtener una muestra que comprende una población de ácidos nucleicos; (b) unir marcadores de secuencia a ácido nucleicos de la población para formar conjugados de marcador-ácido nucleico, en donde al menos un ácido nucleico de la población o copias del mismo tienen distintos marcadores de secuencia unidos; (c) amplificar los conjugados de marcador-ácido nucleico; (d) secuenciar los conjugados de marcador-ácido nucleico para generar lecturas de secuencia que tienen tasas de error y que comprenden una secuencia de ácido nucleico y una secuencia de marcador; (e) alinear las lecturas de secuencias como secuencias de marcadores para formar grupos de lecturas de secuencia que tienen los mismos marcadores de secuencia; (f) fusionar lecturas de secuencia de grupos para determinar secuencias de ácidos nucleicos, en donde los grupos de las lecturas de secuencias de fusionan en distintas secuencias siempre que dichos grupos de lecturas de secuencias sean distintos con una probabilidad de al menos el noventa y cinco por ciento; y (g) determinar el perfil de secuencia de la población determinando los niveles de las secuencias. Como se ha implementado para la elaboración de perfiles de una población de ácidos nucleicos recombinados, tal método puede implementarse mediante las etapas de: (a) obtener una muestra de un individuo que comprende linfocitos T y/o linfocitos B y/o ADN sin células; (b) unir marcadores de secuencias a moléculas de ácidos nucleicos recombinados de genes de receptores de linfocitos T o genes de inmunoglobulina a partir de la muestra para formar conjugados marcador-ácido nucleico, en donde al menos un ácido nucleico recombinado a partir de la muestra o copias del mismo tienen distintos marcadores de secuencia unidos; (c) amplificar los conjugados de marcador-ácido nucleico; (d) secuenciar los conjugados de marcador-ácido nucleico para proporcionar lecturas de secuencia que tienen tasas de error y que comprenden una secuencia de marcador y una secuencia de ácido nucleico recombinado; (e) alinear las lecturas de secuencias como secuencias de marcadores para formar grupos de lecturas de secuencia que tienen los mismos marcadores de secuencia; (f) fusionar lecturas de secuencia de grupos para determinar clonotipos, en donde los grupos de las lecturas de secuencias de fusionan en distintas secuencias siempre que dichos grupos de lecturas de secuencias sean distintos con una probabilidad de al menos el noventa y cinco por ciento; y (g) determinar el perfil de clonotipo de la muestra determinando los niveles de los clonotipos.

En las realizaciones anteriores, y otras realizaciones que se desvelan en el presente documento, la etapa de

secuenciación de conjugados de marcador-ácido nucleico comprende la secuenciación de una muestra de conjuntados de marcador-ácido nucleico a partir de un amplicón. Normalmente, tal muestra en una muestra representativa en que las frecuencias relativas de los polinucleótidos diana en la muestra original (es decir, la muestra de tejido, muestra de sangre, o similar) se mantienen en la muestra de conjugados de marcador-ácido nucleico a partir del producto de una reacción de amplificación. En algunas realizaciones en las que se analiza la población de ácidos nucleicos recombinados que codifican moléculas de receptores inmunitarios, una muestra de conjugados de marcador-ácido nucleico comprende al menos 10^4 conjugados de marcador-ácido nucleico; en otras realizaciones, tal muestra comprende al menos 10^5 conjugados de marcador-ácido nucleico; en otras realizaciones, tal muestra comprende al menos 10^6 conjugados de marcador-ácido nucleico; en otras realizaciones, tal muestra comprende al menos 10^7 conjugados de marcador-ácido nucleico.

Fusión de lecturas de secuencia

Cuando se unen múltiples marcadores de secuencia a un ácido nucleico recombinado o copias de los mismos, se puede llevar a cabo una etapa de fusión de lecturas de secuencia (o lecturas de secuencia de consenso de grupos) para determinar los clonotipos. Reducir un conjunto de lecturas de secuencia para una muestra dada a un conjunto de clonotipos distintos y registrar el número de lecturas para cada clonotipo sería un trivial si la tecnología de secuenciación no tuviera errores. Sin embargo, en presencia de errores de secuenciación, cada clonotipo genuino está rodeado por una 'nube' de lecturas de secuencia con números variantes de errores con respecto los de su secuencia. La "nube" de errores de secuenciación disminuye en densidad según aumenta la distancia desde el clonotipo en el espacia de la secuencia. Hay disponible una variedad de algoritmos para convertir lecturas de secuencia en clonotipos. En un aspecto, la fusión de lecturas de secuencia (es decir, mezclar clonotipos candidatos determinados por que tienen uno o más errores de secuenciación) depende de al menos tres factores: el número de secuencias obtenidas para cada uno de los clonotipos que se está comparando; el número de bases en las que difieren; y la puntuación de calidad de secuenciación en las posiciones en las que son discordantes. En algunas realizaciones, una relación de probabilidad puede interpretarse y evaluarse que se basa sobre las tasas de error esperadas y una distribución binomial de errores. Por ejemplo, dos clonotipos, uno con 150 lecturas y otro con 2 lecturas con una diferencia entre ellos en un área de pobre calidad de secuenciación se fusionará probablemente ya que son probables de generarse mediante error de secuenciación. Por otro lado, dos clonotipos, uno con 100 lecturas y el otro con 50 lecturas con dos diferencias entre ellos no se fusionan ya que se considera que es poco probable que se generen mediante error de secuenciación. En algunas realizaciones, el algoritmo descrito a continuación puede usarse para determinar clonotipos a partir de lecturas de secuencia. Algunos de estos conceptos se ilustran en la FIg. 5A. En algunas realizaciones de la etapa de fusión, las lecturas de secuencia se convierten, en primer lugar en clonotipos cantidad. Tal conversión depende de la plataforma de secuenciación empleada. Para plataformas que generan lecturas de secuencia largas de alta puntuación Q, la lectura de secuencia o una porción de la misma puede tomarse directamente como clonotipo candidato. Para plataformas que generar lecturas de secuencia más cortas de puntuación Q, se puede requerir algunas etapas de alineación y ensamblaje para convertir un conjunto de lecturas de secuencia relacionadas en un clonotipo candidato. Por ejemplo, para plataformas basadas en Slexa, en algunas realizaciones, los clonotipos candidato se generar a partir de colecciones de lecturas emparejadas a partir de múltiples clústeres, por ejemplo, 10 o más, tal como se ha mencionado anteriormente.

Las frecuencias de clonotipos candidato pueden trazarse en especio de secuencia, tal como se ilustra en la Fig. 5A, donde tal espacio se reduce en una dimensión (el eje horizontal) para fines ilustrativos. El eje vertical proporciona la magnitud de frecuencia de cada clonotipo candidato, log (recuento de lectura) o alguna medición similar. En la figura, los clonotipos candidato se representan mediante los diversos símbolos (530). De acuerdo con una realización de la invención, si se fusionan o no do clonotipos candidato depende de sus respectivas frecuencias o recuentos de lectura (tal como se ha indicado anteriormente), el número de diferencias de base entre ellos (cuantas más diferencias, es menos probable la fusión) y las puntuaciones de calidad de las bases en los emplazamientos en los que las respectivas secuencias difieren (mayores puntuaciones de calidad hace que la fusión sea menos probable). Los clonotipos candidato pueden considerarse en el orden de sus respectivas frecuencias. La Fig. 5A muestra el clonotipo candidato 1 (532), el clonotipo candidato 7 (534) y el clonotipo candidato 11 (536) como los tres candidatos con las tres frecuencias más altas. Relacionado con cada tal clonotipo candidato hay otros clonotipos candidato que se encuentran cerca en la secuencia, pero con frecuencias inferiores, tales como (i) para el clonotipo candidato 1 (532) está el clonotipo candidato 2 (538) y los clonotipos candidato 3, 4, 5 y 6 encerrados por el cono (540); para el clonotipo candidato 7 (534) hay clonotipos candidato 8, 9 y 10 y encerrado por el cono (542); y (iii) para el clonotipo candidato 11, está el clonotipo candidato 12 encerrado por el cono (544). Los conos representan los límites de probabilidad dentro de cual clonotipo candidato de menor frecuencia se fusionaría con uno de los clonotipos candidatos 1,7 u 11 de mayor frecuencia. Tales límites de probabilidad son funciones de la frecuencia de los clonotipos candidatos cercanos (3, 4, 5 y 6 para 1; 8, 9 y 10 para 7; y 12 para 11) y sus distancias en el espacio de secuencia desde los respectivos clonotipos candidato de frecuencia superior. El clonotipo candidato 2 (538) se encuentra fuera del cono (540); por tanto, no se fusionaría con el clonotipo candidato 1 (532). De nuevo, los límites probabilidad (de fusión) se muestran como conos puesto que los clones candidato con frecuencias superiores son más probables que sean clonotipos genuinamente distintos de aquellos de frecuencias inferiores y múltiples diferencias en frecuencias inferiores son más probable que sean errores que múltiples diferencias en frecuencias superiores.

La nube de lecturas de secuencia que rodea cada clonotipo candidato puede modelarse usando la distribución binomial y un modelo simple para la probabilidad de un error de base única. Este último modelo de error puede inferirse a partir del mapeo de segmentos V y J o a partir del algoritmo de hallazgo de clonotipo mismo, a través de autoconsistencia y convergencia. Se construye un modelo para la probabilidad a una secuencia 'nube' dada Y con recuento de lectura C2 y errores E (con respecto a la secuencia C) que es parte de una verdadera secuencia de clonotipo X con un recuento de lectura perfecto C1 con el modelo nulo que X es el único clonotipo verdadero en esta región de espacio de secuencia. Se realiza una decisión de si fusionar o no la secuencia Y en el clonotipo X de acuerdo con los parámetros C1, C2 y E. Para cualquier C1 y E dado se precalcula un valor máx. C2 para decidir fusionar la secuencia Y. Los valores máx. para C2 se escogen de modo que la probabilidad de no realizar la fusión Y con la hipótesis nula de que Y es parte del clonotipo X es menos que algún valor P después de integrar sobre todas las secuencias posibles Y con error E en los alrededores de la secuencia X. El valor P controla el comportamiento del algoritmo y hace que la fusión sea más o menos permisiva.
Si no se fusiona una secuencia Y en el clonotipo X debido a que su recuento de lectura se encuentra por encima del umbral C2 para fusionarse en el clonotipo X, entonces se convierte en una candidato para sembrar clonotipos separados (tales como con el clonotipo candidato 2 (538) en la Fig. 5A). Un algoritmo que implementa tales principios también se aseguraría que cualquier otra secuencia Y2, Y3, etc. que son las "más cercanas" a esta secuencia Y (que se han considerado independientes de X) no se agreguen en X. Este concepto de "cercanía" incluye tanto recuentos de errores con respecto a Y y X como el recuento de lectura absoluto de X e Y, es decir, se modela del mismo modo que el anterior modelo para la nube de secuencias de error alrededor del clonotipo X. De este modo, las secuencias "nube" pueden atribuirse adecuadamente a su clonotipo correcto si parecen estar "cerca" más de un clonotipo. Por lo tanto, volviendo a la Fig. 5A, si el clonotipo candidato 2 se considera que es genuinamente distinto del clonotipo candidato 1 (532), entonces, una rutina especial o subalgoritmo, proporcionaría una norma para determinar cuál de los clonotipos candidato 1 (532) y 2 (538), candidatos 4 y 5, entre 1 y 2, debe fusionarse con (si alguno).

En una realización, un algoritmo procede de un modo de arriba a abajo empezando con la secuencia X con el recuento de lectura más alto. Esta secuencia siembre el primer clonotipo. Las secuencias cercanas se fusionan o bien en este clonotipo si sus recuentos se encuentran por debajo de los umbrales precalculados (véase, anteriormente) o se dejan solas si se encuentran por encima del umbral o "cerca" de otra secuencia que no se fusionó. Después de buscar todas las secuencias cercanas dentro de un máximo recuento de error, es proceso de fusionar lectura en el clonotipo X se finaliza. Sus lecturas y todas las lecturas que se han fusionado en este se recuentan y retiran de la lista de lecturas disponibles para realizar otros clonotipos. La siguiente secuencia se mueve con el recuento de lectura más alto. Las lecturas cercanas se fusionan en este clonotipo tal como se ha indicado anteriormente y este proceso se continúa hasta que ya no hay más secuencias con recuentos de lectura por encima de un umbral dado, por ejemplo, hasta que todas las secuencias con más de 1 recuento se hayan usado como siembras para lo clonotipos.

Como se ha mencionado anteriormente, en otra realización del anterior algoritmo, puede añadirse un ensayo adicional para determinar si fusionar una secuencia candidato Y en un clonotipo X existente, que tiene en cuenta la puntuación de calidad de las lecturas de secuencia relevantes. La(s) puntuación(es) de calidad promedio se determinan para secuencia(s) Y (ponderadas por todas las lecturas con la secuencia Y) que son secuencias Y y X que difieren. Si la puntuación promedio se encuentra por encima de un valor predeterminado entonces es más probable que la diferencia indique un verdadero clonotipo distinto que no debe fusionarse y si la puntuación promedio se encuentra por debajo de tal valor predeterminado entonces es más probable que la secuencia Y se provoque mediante errores de secuenciación, y por lo tanto, debe fusionarse en X.

La implementación exitosa del anterior algoritmo para fusionar clonotipos candidato depende de tener un modo eficaz para encontrar todas las secuencias con menos de erros E (es decir, menos de alguna medición de distancia de secuencia) desde una secuencia X de entrada. Este problema puede resolverse usando un árbol de secuencia. La implementación de tales árboles tiene algunas características inusuales en que los nodos del árbol no están restringidos a que sean letras únicas de las secuencias de ADN de los clonotipos candidato, tal como se ilustra en la Fig. 5D. Los nodos pueden tener secuencias arbitrariamente largas, que permiten un uso más eficaz de la memoria del ordenador.

Todas las lecturas de una muestra dada se colocan en el árbol de secuencia. Cada nodo de hoja mantiene indicadores a sus lecturas asociadas. Se recupera una única secuencia de un clonotipo candidato recorriendo hacia atrás en el árbol desde la hoja hasta el nodo raíz. La primera secuencia se coloca en un simple árbol con un nodo raíz y un nodo de hoja que contiene la secuencia completa de la lectura. Las secuencias se añaden a continuación una por una. Para cada secuencia añadida o bien se forma una nueva rama en el último punto de secuencia común entre la lectura y el árbol existente o se añade la lectura a un nodo de hoja existente si el árbol ya contiene la secuencia. Habiendo colocado todas las lecturas en el árbol es fácil de usar el árbol para los siguientes fines: 1) Encontrar el recuento de lectura más alto: clasificar los nodos de hoja mediante recuento de lectura permite encontrar el nodo de hoja (es decir, secuencia) con la mayoría de lecturas y sucesivamente números inferiores de lecturas; 2) Encontrar hojas cercanas: para cualquier secuencia todas las trayectorias a través del árbol que tiene menos de X errores con respecto a esta secuencia se pueden buscar. Una trayectoria se inicia en la raíz y ramifica esta trayectoria para separar trayectoria que proceden a lo largo del árbol. Se anota el recuento de error actual de

cada trayectoria según procede a lo largo del árbol. Cuando el recuento de error supera el máx. permitido de errores se finaliza la trayectoria dada. De este modo, las partes grandes del árbol se cortan lo más temprano posible. Es un modo eficaz de encontrar todas las trayectorias (es decir, todas las hojas), dentro de X errores a partir de cualquier secuencia dada.

Las características de los anteriores conceptos se ilustran en más detalle en el diagrama de flujo de la FIg. 5B. Un conjunto de clonotipos candidato se obtiene a partir de datos de secuencia obtenidos mediante la secuenciación de ácidos nucleicos recombinados a partir de una muestra de linfocitos T o linfocitos B. En un aspecto, los clonotipos candidato incluyen cada uno una región NDN y porciones de regiones V y J. Estas secuencias se organizan en una estructura de datos (550) que puede ser un árbol de secuencia. No se muestra en la Fig. 5B, como parte de generación de un conjunto de clonotipos candidato, en una realización, los árboles de secuencia también se pueden construir para regiones V conocidos y regiones J conocidas. Las lecturas de secuencia que conforman un clonotipo candidato pueden, posteriormente, mapearse o alinearse, con respecto a estas secuencias conocidas mediante árboles de secuencia para determinar eficazmente las secuencias V y J más probablemente conocidas de los clonotipos candidato. Volviendo a la Fig. 5B, una vez se han generados los clonotipos candidato, una estructura de datos, tal como un árbol de secuencia, se construye para su uso en un método para distinguir clonotipos genuinos a partir de clonotipos candidato que contienen errores experimentales o de medición, tales como errores de secuenciación. El clonotipo candidato que tiene la frecuencia más alta de aparición entre los clonotipos candidato actuales ($HFCC_k$) se selecciona (552) entre la estructura de datos, por ejemplo, un árbol de secuencia; en otras palabras, $HFCC_k$ es el clonotipo candidato con el número más alto de copias o recuentos de lectura en el ciclo k. A continuación, se identifican los clonotipos candidato con menor frecuencia cercanos (LFCC) (554); es decir, se identifican los clonotipos candidato con una distancia de D_k. En un aspecto de la invención, esta identificación se lleva a cabo usando un árbol de secuencia, que permite comparaciones de secuencia eficaces de secuencias relativamente cortas (<300 pb).

En una realización, las comparaciones o alineamientos de secuencia, se llevan a cabo usando programación dinámica, tal como se desvela por Gusfield (citado anteriormente). En una realización adicional, tal programación dinámica es una programación dinámica con bandas donde no se consideran las secuencias que difieren del HFCC seleccionado por más de una distancia predeterminada, lo que acelera la informatización. Los candidatos de $HFCC_k$ y $LFCC_j$ pueden compararse basándose en muchos criterios o propiedades distintas. En un aspecto, como se ha mencionado anteriormente, los clonotipos candidato se comparan basándose en al menos dos propiedades: (i) frecuencia o recuentos de lectura y (ii) diferencias de secuencia. En otro aspecto, como se ha mencionado anteriormente, los clonotipos candidato se comparan basándose en al menos tres propiedades: (i) frecuencia o recuentos de lectura, (ii) diferencias de secuencia y (iii) puntuaciones de calidad o mediciones de las bases en las que se producen las diferencias. En una realización, las diferencias de secuencia incluyen sustituciones de base; en otra realización, les diferencias de secuencia incluyen sustituciones, deleciones e inserciones de bases. La última realización es especialmente aplicable siempre que se generen datos de secuencia mediante químicas de secuenciación por síntesis que no emplean terminadores, tales como secuenciadores 454 o secuenciadores Ion Torrent. Tales enfoques de secuenciación diferencian distintos tramos de homopolímeros medidos mediante amplitud de señal; por tanto, las rutinas de identificación de nucleótidos en tales enfoques tienden a errores de inserción y deleción, debido a que la diferencia del nivel de señal de los homopolímeros que difieren en un nucleótido desciende de forma precipitada con un aumento del tamaño del homopolímero (es decir, uno de 2 meros se distinguen más fácilmente que de uno de 3 meros, pero de 8 meros es casi indistinguible de uno de 9 meros). En un aspecto, la comparaciones de HFCCs y LFCC puede implementarse usando una función (denominada en el presente documento como una "función de probabilidad de fusión"), tal como $P(HFCC_k, LFCC_j, D, Q)$ que se muestra en el recuadro de decisión (558) que depende de las cantidades de (i) a (iii) descritas anteriormente. Tal función puede tomar muchas formas distintas, pero, en general, el valor de P cambia con cambios en (i), (ii) y (iii) del siguiente modo: El valor de P aumenta monotónicamente con la frecuencia del HFCC y la relación de la frecuencia de HFCC con la de LFCC, de modo que cuanto mayor sea la relación de la frecuencia de HFCC con la de LFCC, mayor será la probabilidad de que LFCC se fusione en HFCC. Del mismo modo, el valor de P preferentemente disminuye monotónicamente con el grado en el que las secuencias de HFCC y LFCC difieren, de modo que cuanto mayor sea la diferencia entre HFCC and LFCC (por ejemplo, según se ha medido mediante el número mínimo de sustituciones, inserciones o deleciones para cambiar entre sí) menor será la probabilidad de que LFCC se fusione con HFCC. Finalmente, el valor de P preferentemente disminuye monotónicamente con puntuaciones de calidad en aumento de los emplazamientos en los que las secuencias de HFCC y LFCC difieren, de modo que cuando mayor sean las puntuaciones de calidad, menor será la probabilidad de que LFCC se fusione con HFCC.

Cuando las secuencias HFCC y LFCC difieren en más de un emplazamiento, las puntuaciones de calidad en distintos emplazamientos pueden combinarse en una variedad de modos de diferencia. En una realización, siempre que haya una pluralidad de tales diferencias, la pluralidad de las puntuaciones de calidad se expresa como un valor promedio, que puede ser o un promedio no ponderado o un promedio ponderado. La Fig. 5C muestra una función ejemplar, P, informatizada para distintos valores de calidad (curvas de a a e) para una diferencia de secuencia dada. Tal como se ilustra en la Fig. 5C. siempre que HFCC se encuentra en un nivel de aproximadamente 200 recuentos de lectura (570) entonces si las puntuaciones de calidad se determinan mediante la curva (a), cualquiera de LFCC inferior a aproximadamente 50 recuentos de lectura (572) se fusiona en HFCC. El argumento, D, de función P es una medición de la distancia entre las secuencias $HFCC_k$ y $LFCC_j$ y su valor puede variar de ciclo a ciclo según progresa

el análisis. (Los índices "k" indican que los valores de constantes con un subíndice "k" pueden depender del ciclo computacional, k.) En una realización, $D=D_k$, de modo que su valor es una función del número de ciclos. En otra realización, D=D(frecuencia de HFCC), de modo que su valor es una función de la frecuencia de HFCC, independientemente del número de ciclos. Por ejemplo, según disminuye la frecuencia de HFCC, entonces, la distancia, D, de candidatos a comparar disminuye. En una realización, D es una distancia de Hamming entre $HFCC_k$ y $LFCC_j$; sin embargo, se pueden usar otras mediciones de distancia. En una realización, D_k es una función no creciente de k; y, en otra realización, D_k es una función decreciente de k. Disminuir la magnitud de D con un número de cliclos en aumento o con frecuencia decreciente de HFCC, resulta ventajoso en algunas realizaciones puesto que según progresa la informatización a menos, y candidatos clonotipo de menor frecuencia, la mayoría de los cuales son componentes, de modo que la distancia de secuencia (en lugar de la diferencia de frecuencia) se convierte en una comparación predominante. Disminuyendo la D según progresa la informatización, las comparaciones improductivas con respecto a los clonotipos candidato de baja frecuencia distante se reducen, acelerando, de este modo, la informatización. La función P pueden ser una expresión complicada dependiendo del número de factores a tener en cuenta. La Fig. 5C ilustra los valores informatizados para una realización de P que se refiere a umbrales de recuento de lectura para fusionar un LFCC dado un recuento de lectura de un HFCC para distintas puntuaciones de calidad, tal como se describió anteriormente. Las curvas de "a" a "e" representan las relaciones para las distintas puntuaciones de calidad (correspondiéndose la curva "a" a la puntuación de calidad más alta).

Volviendo a la Fig. 5B, si $P<P_k$, entonces LFCCj no se fusiona con $HFCC_k$ y se selecciona (560) otro LFCC. Si $P>P_k$, entonces LFCCj se fusiona con $HFCC_k$ (562), en cuyo caso se selecciona (566) otro LFCC, a menos que ya no haya más LFCC para evaluar (564). Si ya no hay más LFCC para evaluar (564), entonces el $HFCC_k$ actual (incluidos todos los LFCC fusionados en este) se retira (568) de la estructura de datos, tal como el árbol de secuencia. Tal retiración se ilustra en el árbol de secuencia (590) sencillo de las Fig. 5D-5E. Aquí, la trayectoria (592) (indicada por una línea discontinua) en el árbol de secuencia (590) se corresponde con HFCC (596) que se fusiona con LFCC (598). Después de la fusión, el segmento de la trayectoria (592) en el área discontinua (599) se retira del árbol de secuencia (590) para proporcionar un árbol de secuencia reducido (597) que se muestra en la Fig. 5E, que se usa en computaciones posteriores para encontrar LFCC (554) cercano. Después de tal retiración, se finaliza la determinación de clonotipo si se cumple un criterio de detención (570). En una realización, el criterio de detención (570) es si el último clonotipo candidato no componente se ha procesado o no (552). En otra realización, el criterio de detención (570) es si la frecuencia o los recuentos de lectura del HFCC seleccionado se encuentran por debajo o no de los que se corresponden con un único linfocito. En un aspecto del método de la invención, una etapa de amplificación puede resultar en cada linfocito en una muestra que es representada por múltiples copias del mismo clonotipo; por tanto, en una realización, siempre que HFCC tenga un número de recuentos de lectura por debajo del número que se corresponde con un único linfocito, entonces, la computación se detiene. En algunas realizaciones, tal número de recuentos de lectura (o copias de clonotipo candidato) es de al menos 10; en otra realización, tal número es de al menos 20; en otra realización, tal número es de al menos 30; en otra realización, tal número es de al menos 40. Si no se cumple el criterio de detención, entonces se selecciona (572) el siguiente HFCC. Las etapas analíticas que se resumen en el diagrama de flujo de la Fig. 5B pueden implementarse en cualquier idioma de programación adecuado, tal como C, C++, Java, C#, Fortran, Pascal o similares.

El anterior método para determinar clonotipos y/o perfiles de clonotipo puede comprender las etapas de (a) formar una estructura de datos de moléculas inmunitarias recombinadas a partir de lecturas de secuencia obtenidas mediante secuenciación de ácidos nucleicos de alto rendimiento, (b) fusionar con un clonotipo candidato de frecuencia más alta cualquier clonotipo candidato de frecuencia más baja siempre que tal frecuencia inferior se encuentre por debajo de un valor de frecuencia predeterminado y una diferencia de secuencia entre estos se encuentra por debajo de un valor de diferencia predeterminado para formar un clonotipo, (c) retirar el clonotipo candidato fusionado de la estructura de datos y (d) repetir las etapas (b) y (c) hasta que se forme un perfil de clonotipo. En una realización, la estructura de datos en un árbol de secuencia.

El anterior método de determinación de clonotipos se puede llevar a cabo mediante etapas que comprenden: (a) proporcionar un conjunto de lecturas de secuencia a partir de un repertorio de moléculas inmunitarias recombinadas que tiene cada una, una región V, una región NDN y una región J en donde para cada molécula al menos una lectura de secuencia abarca al menos una porción de la región NDN de tal molécula; (b) formar a partir de lecturas de secuencias que abarcan al menos una porción de una región NDN un árbol de secuencia que tiene hojas que representan clonotipos candidato, cada hoja y su clonotipo candidato correspondiente teniendo una frecuencia; (c) fusionar con un clonotipo candidato de frecuencia más alta cualquier clonotipo candidato de frecuencia más baja siempre que tal frecuencia inferior se encuentre por debajo de un valor de frecuencia predeterminado y una diferencia de secuencia entre estos se encuentra por debajo de un valor de diferencia predeterminado para formar un clonotipo que tenga una secuencia del clonotipo candidato de frecuencia más alta; (d) retirar las hojas que se corresponden con los clonotipos candidato fusionados a partir del árbol de secuencia; y (e) repetir las etapas (c) y (d) hasta que una frecuencia más alta de un clonotipo candidato de menor frecuencia se encuentre por debajo de un valor de detención predeterminado. En una realización, la etapa de formación de más incluye seleccionar un clonotipo candidato de frecuencia más alta e identificar dichos clonotipos candidato de menor frecuencia que tienen una diferencia de secuencia entre ellos inferior a un valor de diferencia predeterminado para formar un subconjunto de fusión. Por lo tanto, en tal realización, uno puede limitar el número total de LFCC que debe compararse para la operación de fusión (solo se consideran los que se encuentran dentro del valor de diferencia predeterminado). Tal

valor es una entrada de proceso dependiendo de la aplicación, por ejemplo, el tamaño del repertorio, cuánto tiempo de ordenador se usa, etcétera. Como se ha mencionado anteriormente, la función usada para decidir si fusionar un HFCC con un LFCC puede tener una variedad de formas. En un aspecto general, la etapa de fusión, tal función puede tener las siguientes propiedades: Depende de las frecuencias de HFCC, LFCC, la diferencia de secuencia entre estos (que puede expresarse como una medición de diferencia de cadena convencional, tal como una distancia Hamming) y puntuaciones de calidad del uno o más emplazamientos de nucleótidos en donde el HFCC y el LFCC difieren; de modo que la función (i) monotónicamente aumenta con una relación creciente de HFCC y frecuencia de LFCC, (ii) monotónicamente disminuye con una diferencia de secuencia en aumento entre HFCC y LFCC y (iii) monotónicamente disminuye con puntuaciones de calidad en aumento del uno o más emplazamientos de nucleótidos. Es decir, con respecto a la propiedad (iii), lo seguro es que HFCC y LFCC son distintos (por ejemplo, porque hay un alto nivel de confianza en las identificaciones de nucleótidos), entonces, es menos probable que se fusionen.

En algunas realizaciones, se selecciona una función de probabilidad de fusión de modo que las lecturas de secuencia se fusionan en distintos clonotipos (o polinucleótidos diana, tal como, ácidos nucleicos recombinados) siempre que tales lecturas de secuencia sean distintas con una probabilidad de al menos el 95 por ciento; en otras realizaciones, se selecciona una función de probabilidad de fusión de modo que las lecturas de secuencia se fusionan en distintos clonotipos siempre que tales lecturas de secuencia son distintas con una probabilidad de al menos el 99 por ciento; en otras realizaciones, se selecciona una probabilidad de fusión de modo que las lecturas de secuencia se fusionan en distintos clonotipos siempre que tales lecturas de secuencia sean distintas con una probabilidad de al menos el 99,9 por ciento. Como se ha mencionado anteriormente, en algunas realizaciones, una función de probabilidad de coalescencia depende de una tasa de error de una química de secuenciación usada, el número de nucleótidos discrepantes en las lecturas de secuencia que se están comparando y las frecuencias relativas de las lecturas de secuencia que se están comparando; en otra realización, una función de probabilidad de coalescencia depende de una tasa de error de una química de secuenciación usada, el número de nucleótidos discrepantes en lecturas de secuencia que se están comparando, las frecuencias relativas de las lecturas de secuencia que se están comparando y las puntuaciones de calidad de los nucleótidos discrepantes. En lo anterior, la selección de un valor de secuencia predeterminado y un valor de diferencia predeterminado es una elección de diseño que depende de las aplicaciones particulares. Los factores que afectan tales elecciones pueden incluir detalles de la biología, velocidad de implementación y similares.

Aplicaciones de control

En el presente documento se describen métodos para controlar enfermedad mínima residual determinando la presencia, ausencia y/o nivel de ácidos nucleicos en una muestra que es característica o está correlacionada con una enfermedad. En algunos casos, tales ácidos nucleicos son ácidos nucleicos recombinados o clonotipos, que están correlacionados con una afección pre-cancerosa o cancerosa, tal como trastorno proliferativo linfoide o mieloide y que puede usarse para controlar el estado del trastorno o afección. Tales ácidos nucleicos y, en particular, clonotipos, son útiles para controlar enfermedad mínima residual de un cáncer después de su tratamiento, en el que el resultado de tal control es un factor clave en la determinación de si continuar, no continuar o, de otro modo, modificar el tratamiento. En muchos neoplasmas linfoides y mieloides malignos, una muestra de tejido diagnóstico, tal como una muestra de sangre periférica o una muestra de médula ósea, se obtiene antes del tratamiento a partir de la cual se genera del perfil de clonotipo (un "perfil de clonotipo diagnóstico"). Para trastornos proliferativos linfoide o mieloide, no se conoce normalmente antes de una muestra diagnóstica qué cadena(s) de receptores inmunitarios está(n) correlacionada(s) con el clon linfoide o mieloide del trastorno o afección. Por consiguiente, en la práctica actual se deben llevar a cabo muchas amplificaciones y secuenciaciones por separado sobre distintos ácidos nucleicos recombinados que codificas distintas cadenas de receptores inmunitarios para identificar los clonotipos correlacionados con una enfermedad o afección del paciente. Uno o más clonotipos correlacionados con la enfermedad (es decir, "clonotipos de correlación") se identifican en perfiles de clonotipo que resultan de tales esfuerzos de amplificación y secuenciación. Típicamente, los clonotipos que tienen las frecuencias más altas en los perfiles de clonotipo se toman como los clonotipos de correlación. En un caso, el número de tiradas de amplificaciones y secuenciaciones por separado necesarias para identificar clonotipos de correlación se reduce en gran medida proporcionando amplificaciones de multiplexación a una escala más grande en una única reacción de porciones de ácidos nucleicos recombinados que codifican un pluralidad de distintas cadenas de receptores inmunitarios. En algunos casos, tal pluralidad se encuentra en el intervalo de 2 a 4 cadenas de receptores inmunitarios por separado; y, en otras realizaciones, tal pluralidad se encuentra en el intervalo de 2 a 3 cadenas de receptores inmunitarios por separado. Más particularmente, en algunas realizaciones, entre las cadenas RLB los siguientes se amplifican en una única reacción de multiplexación: ácidos nucleicos recombinados que codifican IgH que incluye al menos una porción de la región VDJ, IgH que incluye al menos una porción de la región DH y IgK; y, en otras realizaciones, entre las cadenas RLT los siguientes se amplifican en una única reacción de multiplexación: RLTβ, RLTδ y RLTγ.

Después del tratamiento y, preferentemente después de conseguir una remisión completa del cáncer, la presencia, ausencia o frecuencia de tales clonotipos de correlación para ácidos nucleicos se evalúa periódicamente para determinar si la remisión permanece o si el neoplasma vuelve o reincide, basándose en la presencia de, o un aumento en la frecuencia de, los ácidos nucleicos de correlación o clonotipo (o clonotipos relacionados) en un perfil

de clonotipo post-tratamiento o perfil de ácidos nucleicos. Es decir, después del tratamiento, se evalúa la enfermedad mínima residual basándose en la presencia, ausencia o frecuencia de los clonotipos de correlación o ácidos nucleicos característicos. Como se ha mencionado anteriormente, cuando tales clonotipos de correlación son comunes o se corresponden con un segmento de receptor reorganizado que no tiene suficiente diversidad (de modo que las células no cancerosas pueden compartir el clonotipo), la aparición de tales clonotipos en un perfil de clonotipo post-tratamiento puede provocar una indicación de falso positivo de recaída.

Los métodos de la invención son aplicables para controlar cualquier enfermedad proliferativa en la que un ácido nucleico reorganizado que codifica un receptor inmunitario o porción del mismo puede usarse como marcador de células implicadas en la enfermedad. En un aspecto, los métodos de la invención son aplicables a trastornos proliferativos linfoides y mieloides. En otro aspecto, los métodos de la invención son aplicables a linfomas y leucemias. En otro aspecto, los métodos de la invención son aplicables para controlar EMR en linfoma folicular, leucemia linfocítica crónica (CLL), leucemia linfocítica aguda (ALL), leucemia mielógena crónica (CML), leucemia mielógena aguda (AML), linfomas de Hodgkin y de no Hodgkin, mieloma múltiple (MM), gamopatía monoclonal de significancia indeterminada (MGUS), linfoma de células del manto (MCL), linfoma difuso de linfocitos B grandes (DLBCL), síndromes mielodisplásicos (MDS), linfoma de linfocitos T o similares. En una realización particular, un método de la invención está particularmente bien adecuado para controlar MRD en ALL, MM o DLBCL.

En algunas realizaciones, una muestra de paciente, tal como sangre o médula ósea, se somete a un ensayo diagnóstico para identificar cuál de una pluralidad de cadenas de receptores inmunitarios puede incluir el clonotipo producido por un clon de un trastorno (es decir, un clonotipo de correlación). Una vez se ha determinado la cadena de receptores inmunitarios de un clonotipo de correlación, entonces los ensayos de control posteriores pueden ser específicos para esa cadena de receptores inmunitarios particular. Por ejemplo, en algunas realizaciones, un ensayo diagnóstico puede generar en la misma reacción perfiles de clonotipo basados en secuencias de una pluralidad de cadenas de RLB, tal como, IgH(VDJ), IgH(DJ) y IgK. Si un clonotipo de correlación es una cadena IgH(VDJ), entonces los ensayos de control posteriores pueden solo generar perfiles de clonotipo IgH(VDJ). En algunas realizaciones, la profundidad de secuenciación en la muestra diagnóstica puede ser distinta de la de la muestra de control. "Profundidad de secuenciación" significa el número total de lecturas de secuencia analizadas para construir perfiles de clonotipo. Para cánceres, tales como leucemias o linfomas, puesto que los ensayos diagnósticos se llevan a cabo sobre muestras del paciente antes de su tratamiento, la frecuencia o nivel de un clonotipo de correlación en la muestra es normalmente elevado y fácilmente identificado. Por ejemplo, cualquier clonotipo con una frecuencia sobre un nivel predeterminado puede definirse como un clonotipo de correlación. Tal nivel predeterminado puede variar con otros indicadores del paciente; sin embargo, a menudo un nivel predeterminado puede encontrarse en el intervalo del 2 al 5 por ciento; o, en algunas realizaciones, el cinco por ciento. Por lo tanto, en algunas realizaciones, la profundidad de secuenciación que se lleva a cabo es la que es necesaria para detectar de forma fiable clonotipos presentes en una frecuencia del uno o dos por ciento o superior. En algunas realizaciones, la profundidad de secuenciación de una muestra diagnóstica produce al menos 10.000 lecturas de secuencia; o, en otras realizaciones, es de al menos 100.000 lecturas de secuencia; en todavía otras realizaciones, la profundidad de secuenciación de una muestra diagnóstica produce al menos 10^6 lecturas de secuencia. En algunas realizaciones, la profundidad de secuenciación de una muestra de control es de al menos 100.000 lecturas de secuencia; en otras realizaciones, la profundidad de secuenciación de una muestra de control es de al menos 10^6 lecturas de secuencia.

En algunas realizaciones, un trastorno proliferativo linfoide, tal como leucemia o linfoma, en un paciente puede controlarse generando perfiles de clonotipo a partir de muestras exitosamente obtenidas (o muestras de tejido) del paciente. Tales perfiles de clonotipo pueden generarse tal como se ha descrito anteriormente. En algunos casos, tal control puede implementarse mediante las siguientes etapas: (a) obtener una muestra de un individuo que comprende linfocitos T y/o linfocitos B y/o ADN sin células; (b) unir marcadores de secuencias a moléculas de ácidos nucleicos recombinados de genes de receptores de linfocitos T o genes de inmunoglobulina a partir de la muestra para formar conjugados marcador-ácido nucleico, en donde al menos un ácido nucleico recombinado o copias del mismo tienen distintos marcadores de secuencia unidos; (c) amplificar los conjugados de marcador-ácido nucleico; (d) secuenciar una muestra de los conjugados de marcador-ácido nucleico para proporcionar lecturas de secuencias teniendo cada una tasa de error y comprendiendo cada una secuencia de marcador y una secuencia de ácido nucleico recombinado; (e) alinear las lecturas de secuencias como secuencias de marcadores para formar grupos de lecturas de secuencia que tienen los mismos marcadores de secuencia; (f) fusionar lecturas de secuencia de grupos para determinar clonotipos, en donde los grupos de las lecturas de secuencias de fusionan en distintas secuencias de ácidos nucleicos recombinados siempre que dichos grupos de lecturas de secuencias sean distintos con una probabilidad de al menos el noventa y cinco por ciento; (g) determinar el perfil de clonotipo de la muestra determinando los niveles de los clonotipos; y (h) determinar el nivel de clonotipos de correlación en el perfil de clonotipo. En algunos casos, las etapas de (a) a (h) pueden repetirse en el proceso de control del paciente para determinar si el nivel de clonotipos de correlación es prueba de reincidencia de la enfermedad. En algunos casos, las etapas de unir y amplificar pueden comprender las siguientes etapas: (a) combinar en una mezcla de reacción con condiciones de extensión de cebador un primer conjunto de cebadores con una muestra de ácidos nucleicos recombinados de células inmunitarias que expresan un receptor inmunitario y/o ADN sin células, en donde cada cebador del primer conjunto tiene una porción específica a receptor de modo que la porción específica a receptor se hibrida a un ácido nucleico recombinado distinta en un emplazamiento predeterminado y se extiende para formar un primer producto de extensión, y en donde cada cebador del primer conjunto tiene un extremo 5' no complementario

que contiene un primer sitio de unión del cebador; (b) retirar de la mezcla de reacción cebadores no extendidos del primer conjunto; y (c) añadir a la mezcla de reacción en condiciones de extensión de cebador un segundo conjunto de cebadores, en donde cada cebador del segundo conjunto tiene una porción específica a receptor de modo que la porción específica a receptor se hibrida al primer producto de extensión en un emplazamiento predeterminado y tiene un extremo 5' no complementario que contiene un segundo sitio de unión de cebador, los cebadores del primer conjunto y/o los cebadores del segundo conjunto que comprenden un marcador de secuencia dispuestos entre la porción específica a receptor y el primer o segundo sitio de unión de cebador, respectivamente, y en donde cada cebador del segundo conjunto se extiende para formar un segundo producto de extensión, de modo que cada segundo producto de extensión comprende un primer sitio de unión de cebador, un segundo sitio de unión de cebador, al menos un marcador de secuencia y ácido nucleico recombinante que codifica una porción de una cadena de receptor de células inmunitarias. En algunas realizaciones, una etapa de fusión de ácidos nucleicos recombinados comprende fusionar lecturas de secuencia de distintos ácidos nucleicos recombinados siempre que tales lecturas de secuencia sean distintas con una probabilidad de al menos el noventa y nueva por ciento; y, en otras realizaciones, con una probabilidad de al menos el 99,9 por ciento.

Los métodos de la invención también son aplicables para controlar la enfermedad mínima residual de cáncer en un paciente, incluido un cáncer no linfoide o no mieloide, que tiene un patrón de identificación de mutaciones, por ejemplo, en un conjunto seleccionado de genes cancerosos. Tal patrón de mutaciones, es decir, la presencia, ausencia y/o nivel de genes que contienen tales mutaciones, puede indicar una probabilidad de reincidencia de la enfermedad. En algunas realizaciones, los polinucleótidos diana para tal control pueden ser exones, porciones de exones, intrones seleccionados y/o regiones de control de expresión génica, por ejemplo, promotores, de una pluralidad de genes (denominados en el presente documento como "moléculas de genes cancerosos"). Las moléculas de genes cancerosos se pueden aislar de una muestra de tejido usando técnicas convencionales, tales como técnicas de captura de exones, por ejemplo, kit de enriquecimiento de exomas de TruSeq™ (Illumina, San Diego, CA); Frampton y col., Nature Biotechnology, 31(11): 1023-1031 (2013); y similares. Después de obtener tales moléculas de genes cancerosos, los marcadores de secuencia se unen para formar un conjugado de marcador-ácido nucleico, el conjugado de marcador-ácido nucleico se amplifica y secuencia de acuerdo con la invención.

Estudios de secuenciación de genoma del cáncer recientes han mostrado que hay una heterogeneidad significante en patrones de mutación entre distintos cánceres, entre distintos pacientes con el mismo cáncer, entre células del mismo tumor y entre células de distintos sitios metastáticos en el mismo paciente; sin embargo, en el mismo paciente, las células cancerosas heterogéneas normalmente evolucionan de un ancestro común, de modo que comparten mutaciones y una relación evolutiva entre las células cancerosas puede discernirse en una sucesión de mediciones durante el tiempo, por ejemplo, Vogelstein y col., Science, 339: 1546-1558 (2013); Ding y col., Nature, 481(7382): 506-510 (2012); y similares; por lo tanto, un patrón de mutaciones correlacionadas con un cáncer medidas en una muestra diagnóstica proporciona medios para detectar una reincidencia del mismo cáncer o una versión clonalmente evolucionada de este.

Las moléculas del gen del cáncer pueden seleccionarse entre una amplia variedad de genes, que incluyen, pero sin limitación, los genes en la Tabla I.

Tabla I

Genes cancerosos ejemplares				
ABL1	AKT1	ALK	APC	ATM
BRAF	CDH1	CSF1R	CTNNB1	EGFR
ERBB2	ERBB4	FBXW7	FGFR1	FGFR2
FGFR3	FLT3	GNA11	GNAQ	GNAS
HNF1A	HRAS	IDH1	JAK2	JAK3
KDR	KIT	KRAS	MET	MLH1
MPL	NOTCH1	NPM1	NRAS	PGGFRA
PIK3CA	PTEN	PTPN11	RB1	RET
SMAD4	SMO	SRC	STK	TP53
VHL				

En algunos casos, el anterior método de control de una enfermedad mínima residual de un cáncer puede comprender las siguientes etapas: (a) obtener de un individuo una muestra de tejido; (b) unir los marcadores de secuencia a cada una de una pluralidad de moléculas de genes cancerosos en la muestra para formar conjugados de marcador-ácido nucleico, en donde al menos un ácido nucleico o copias del mismo tiene distintos marcadores de secuencia unidos y en donde las moléculas de gen cancerosas son características de un cáncer del individuo; (c) amplificar los conjugados de marcador-ácido nucleico; (d) secuenciar una muestra de los conjugados de marcador-ácido nucleico para proporcionar lecturas de secuencia que tienen tasas de error y que comprenden una secuencia de marcador y una secuencia de gen; (e) alinear las lecturas de secuencias como secuencias de marcadores para formar grupos de lecturas de secuencia que tienen los mismos marcadores de secuencia; (f) fusionar las secuencias

de genes cancerosos de grupos para determinar las secuencias de moléculas de genes cancerosos, en donde los grupos de las lecturas de secuencias de fusionan en distintas moléculas de genes cancerosos siempre que dichos grupos de secuencias de genes cancerosos sean distintos con una probabilidad de al menos el noventa y cinco por ciento; y (g) detectar en un perfil de las moléculas de genes cancerosos la presencia, ausencia y/o nivel de moléculas de genes cancerosos característico del cáncer del individuo. En algunos casos, una etapa de fusión de secuencias de genes cancerosos comprende fusionar lecturas de secuencia de distintas moléculas de genes cancerosos siempre que tales lecturas de secuencia sean distintas con una probabilidad de al menos el noventa y nueva por ciento, y, en otros casos, con una probabilidad de al menos el 99,9 por ciento.

Uso de marcadores de secuencia para detectar contaminación por arrastre

La contaminación por arrastre resulta un problema significante con técnicas que incluyen la amplificación de ácidos nucleicos, por ejemplo, Borst y col., Eur. J. Clin. Microbiol. Infect. Dis., 23(4): 289-299 (2004); Aslanzadeh, Ann. Clin. Lab. Sci., 34(4): 389-396 (2004); y similares. Tal contaminación surge cuando trazas de ácido nucleico extraño a una muestra se amplifica de forma no intencionada en un ensayo de la muestra y afecta o tiene un impacto sobre un resultado medido. En un peor caso, la contaminación por arrastre en una muestra médica de un paciente puede dar como resultado una interpretación de falso positivo en un resultado de ensayo. El ácido nucleico extraño puede provenir de una fuente no relacionada con un paciente particular; por ejemplo, puede provenir de una muestra de otro paciente. O, el ácido nucleico extraño puede provenir de una fuente relacionada con un paciente; por ejemplo, puede provenir de una muestra distinta del mismo paciente manipulada en el mismo laboratorio en el pasado o de una reacción de ensayo sobre una muestra distinta del mismo paciente que se procesó en el mismo laboratorio en el pasado.

La contaminación por arrastre es especialmente complicada en un escenario clínico cuando se miden poblaciones altamente complejas de ácidos nucleicos relacionados, tales como poblaciones de ácidos nucleicos recombinados que codifican moléculas inmunitarias, tales como receptores de linfocitos T o inmunoglobulinas. El reto aparece porque resulta complicado determinar si una lectura de secuencia o clonotipo es parte de la genuina diversidad de una muestra prevista o si originan de una fuente extraña de ácido nucleico, tal como la muestra de otro paciente o una muestra anterior del mismo paciente, que están siendo procesadas en el mismo tipo de ensayo en el mismo laboratorio. En un aspecto de la invención, tal contaminación por arrastre puede detectarse usando marcadores de secuencia no solo para determinar clonotiops a partir de lecturas de secuencia sino también determina run marcador de secuencia originado en la muestra actual o a partir de otra muestra. Esto se consigue manteniendo un registro de marcadores de secuencia determinados a partir de cada muestra de paciente, entonces, siempre que se realice una medición posterior los marcadores de secuencia de la medición actual se comparan con los de las anteriores mediciones. Tales registros de marcadores de secuencia asociados con clonotipos se mantienen de forma conveniente como registros electrónicos sobre dispositivo de almacenamiento en masa debido al gran número de marcadores de cada medición y la facilidad de búsqueda y comparación de registros electrónicos que usan algoritmos convencionales. Si se encuentra una coincidencia entonces la causa más probable es la contaminación por arrastre, siempre que las poblaciones de marcadores de secuencia empleados en las mediciones sean suficientemente grandes. Las mismas relaciones ejemplares del tamaño de la población de marcadores de secuencia con respecto a una población de clonotipos para marcar mediante muestreo descrito anteriormente son aplicables para detectar contaminación por arrastre. En una realización, tal relación es de 100:1 o superior.

Una amplia variedad de métodos de búsqueda o algoritmos pueden usarse para llevar a cabo la etapa de comparación de clonotipos medidos con clonotipos de bases de datos. Muchos algoritmos de búsqueda y alineación de secuencia convencionales están disponibles públicamente y se han descrito en las siguientes referencias: Mount, Bioinformatics Sequence and Genome Analysis, Segunda edición (Cold Spring Harbor Press, 2004); Batzoglou, Briefings in Bioinformatics, 6: 6-22 (2005); Altschul y col., J. Mol. Biol., 215(3): 403-410 (1990); Needlemanand Wunsch, J. Mol. Biol., 48: 443-453 (1970); Smith y Waterman, Advances in Applied Mathematics, 2: 482-489 (1981); y similares.

En algunos casos, los anteriores métodos para detectar y medir la contaminación, tal como contaminación de arrastre, en una muestra a partir de un material que proviene de una muestra distinta pueden comprender las siguientes etapas: (a) obtener de un individuo una muestra de tejido; (b) unir marcadores de secuencia a moléculas de gen cancerosas o ácidos nucleicos recombinados para formar conjugados de marcador-ácido nucleico, en donde al menos un ácido nucleico o copias del mismo tiene distintos marcadores de secuencia unidos y en donde las moléculas de gen cancerosas son características de un cáncer del individuo; (c) amplificar los conjugados de marcador-ácido nucleico; (d) secuenciar una muestra de los conjugados de marcador-ácido nucleico para proporcionar lecturas de secuenciación teniendo cada una, una tasa de error y comprendiendo cada una, una secuencia de marcador y una secuencia de gen canceroso o secuencia de ácido nucleico recombinado; (e) comparar secuencias de marcador con respecto a secuencias de marcador determinadas por separado de otras muestras de tejido; y (f) determinar la presencia, ausencia y/o nivel de contaminación mediante la identidad de una o más secuencias de marcador con cualquier secuencia de marcador determinada por separado de otras muestras de tejido. Una vez se han determinados las secuencias de marcador en un ensayo, pueden compararse con secuencias de marcador en una base de datos de secuencias de marcador registradas a partir de ensayos sobre otros pacientes. Tales etapas pueden implementarse en el momento de un ensayo, o tales etapas pueden implementarse

retrospectivamente, por ejemplo, en el momento después del tiempo del ensayo. En un caso, los marcadores de secuencia se unen a ácidos nucleicos recombinados en una muestra de tejido, tal como sangre o médula ósea, de un individuo que sufre un trastorno proliferativo linfoide, tal como un cáncer linfoide. En otro caso, los marcadores de secuencia se unen a moléculas de genes cancerosos, tal como se ha descrito anteriormente.

En casos adicionales en los que se controlar ácidos nucleicos recombinados a partir de contaminación cruzada de muestras de tejido, las etapas de unir y amplificar pueden implementarse del siguiente modo: (a) combinar en una mezcla de reacción con condiciones de extensión de cebador un primer conjunto de cebadores con una muestra de ácidos nucleicos recombinados de linfocitos T y/o ADN sin células, en donde cada cebador del primer conjunto tiene una porción específica a receptor de modo que la porción específica a receptor se hibrida a un ácido nucleico recombinado distinta en un emplazamiento predeterminado y se extiende para formar un primer producto de extensión, y en donde cada cebador del primer conjunto tiene un extremo 5' no complementario que contiene un primer sitio de unión del cebador; (b) retirar de la mezcla de reacción cebadores no extendidos del primer conjunto; (c) añadir a la mezcla de reacción en condiciones de extensión de cebador un segundo conjunto de cebadores, en donde cada cebador del segundo conjunto tiene una porción específica a receptor de modo que la porción específica a receptor se hibrida al primer producto de extensión en un emplazamiento predeterminado y tiene un extremo 5' no complementario que contiene un segundo sitio de unión de cebador, los cebadores del primer conjunto y/o los cebadores del segundo conjunto que comprenden un marcador de secuencia dispuestos entre la porción específica a receptor y el primer o segundo sitio de unión de cebador, respectivamente, y en donde cada cebador del segundo conjunto se extiende para formar un segundo producto de extensión, de modo que cada segundo producto de extensión comprende un primer sitio de unión de cebador, un segundo sitio de unión de cebador, al menos un marcador de secuencia y ácido nucleico recombinante que codifica una porción de una cadena de receptor de células inmunitarias; y (d) llevar cabo una reacción en cadena de la polimerasa en la mezcla de reacción para formar un amplicón, usando la reacción en cadena de la polimerasa cebadores directos específicos para el primer sitio de unión de cebador y cebadores indirectos específicos para el segundo sitió de unión de cebador.

Kits

En el presente documento se describe una variedad de kits para llevar a cabo los métodos de la invención. En algunos casos, los kits comprenden (a) un conjunto de cebadores directos y un conjunto de cebadores indirector para amplificar en una PCR de multiplexación ácidos nucleicos recombinados que codifican una pluralidad de cadenas de receptores inmunitarios en donde los cebadores directos y/o cebadores indirectos tienen cada uno una porción específica diana, un marcador de secuencia y un sitio de unión de cebador común y (b) un elemento de recitación de cebador para retirar después de al menos una primera extensión de cebadores no incorporados (es decir, cebadores no extendidos) de los conjuntos. En algunos casos, los kits comprenden adicionalmente cebadores comunes específicos para los sitios de unión de cebador comunes. En algunos casos, los kits comprenden adicionalmente instrucciones por escrito para usar componentes del kit en un método de la invención. En algunos casos, los kits comprenden adicionalmente cebadores directos e indirectos específicos para amplificar ácidos nucleicos recombinados que codifican IgH(VDJ), IgH(DJ) y IgK. En algunos casos, los kits comprenden cebadores directos e indirectos específicos para amplificar ácidos nucleicos recombinados que codifican RLTβ, RLTδ y RLTγ. En algunos casos, los kits comprenden adicionalmente estándares internos que comprenden una pluralidad de ácidos nucleicos que tienen longitudes y composiciones representativas de ácidos nucleicos recombinados diana, en donde los estándares internos se proporcionan en concentraciones conocidas. En algunos casos, los kits incluyen exonucleasa unicatenaria como un elemento de retiración de cebador, tal como exonucleasa I de E. coli. En algunos casos, los kits incluyen una columna de centrifugación capaz de seleccionar por tamaño ADN bicatenario como elemento de retiración de cebador.

Definiciones

Salvo que se indique específicamente de otra forma en el presente documento, los términos y símbolos de la química de ácidos nucleicos, bioquímica, genética y biología molecular usada en el presente documento sigue los tratados estándares y textos en el ámbito, por ejemplo, Kornberg y Baker, DNA Replication, Segunda edición (W.H. Freeman, Nueva York, 1992); Lehninger, Biochemistry, Segunda edición (Worth Publishers, Nueva York, 1975); Strachan y Read, Human Molecular Genetics, Segunda edición (Wiley-Liss, Nueva York, 1999); Abbas y col., Cellular and Molecular Immunology, 6ª edición (Saunders, 2007).

"Alineamiento" se refiere a un método de comparación una secuencia de ensayo, tal como una lectura de secuencia, con respecto a una o más secuencias de referencia para determinar qué secuencia de referencia o qué porción de una secuencia de referencia es más cercana basándose en alguna medición de distancia de secuencia. Un método ejemplar de alineamiento de secuencias de nucleótidos es el algoritmo de Smith Waterman. Las mediciones de distancia pueden incluir la distancia de Hamming, la distancia de Levenshtein o similares. Las mediciones de distancia pueden incluir un componente relacionado con los valores de calidad de nucleótidos de las secuencias que se están comparando.

"Amplicón" se refiere al producto de una reacción de amplificación de polinucleótidos; es decir, una población clonal de polinucleótidos, que pueden ser unicatenarios o bicatenarios, cuyos polinucleótidos se replican a partir de una o

más secuencias de partida. La una o más secuencias de partida puede ser una o más copias de la misma secuencia o pueden ser una mezcla de distintas secuencias. Los amplicones pueden producirse mediante una variedad de reacciones de amplificación cuyos productos comprenden réplicas de uno o más ácidos nucleicos de partida o diana. En un aspecto, las reacciones de amplificación que producen amplicones son "dirigidas por modelo" en ese emparejamiento de bases de reactivos, o bien nucleótidos bien oligonucleótidos, tienen complementos en un polinucleótido modelo que se requieren para la creación de productos de reacción. En un aspecto, las reacciones dirigidas por modelo son extensiones de cebador con una polimerasa de ácido nucleico o ligaciones de oligonucleótidos con una ligasa de ácido nucleico. Tales reacciones incluyen, pero sin limitación, reacciones en cadena de la polimerasa (PCR), reacciones de la polimerasa lineales, amplificación a base de secuencias de ácidos nucleicos (NASBA), amplificaciones de círculo rodante y similares, que se desvelan en las siguientes referencias que se incorporan en el presente documento por referencia: Mullis y col., patentes de los EE.UU. 4.683.195; 4.965.188; 4.683.202; 4.800.159 (ADN); Gelfand y col., patente de los EE.UU. 5.210.015 (PCR en tiempo real con sondas "taqman"); Wittwer y col., patente de los EE.UU. 6.174.670; Kacian y col., patente de los EE.UU. 5.399.491 ("NASBA"); Lizardi, patente de los EE.UU. 5.854.033; Aono y col., publ. de patente japonesa, JP 4-262799 (amplificación de círculo rodante); y similares. En un aspecto, los amplicones de la invención se producen mediante PCR. Una reacción de amplificación puede ser una amplificación en "tiempo real" si hay disponible una química de detección que permite que un producto de reacción se mida según progresa la reacción de amplificación, por ejemplo, "PCR en tiempo real" que se describe a continuación o "NASBA en tiempo real" como se describe en Leone y col., Nucleic Acids Research, 26: 2150-2155 (1998) y referencias similares. Tal como se utiliza en el presente documento, el término "amplificar" se refiere a realizar una reacción de amplificación. Una "mezcla de reacción" se refiere a una solución que contiene todos los reactivos necesarios para llevar a cabo una reacción, que puede incluir, pero sin limitación, agentes tamponantes para mantener el pH en un nivel seleccionado durante una reacción, sales, cofactores, neutralizantes y similares.

"Clonalidad" tal como se utiliza en el presente documento se refiere al grano en el cual la distribución de las abundancias de clonotipo entre clonotipos de un repertorio se sesgan en un único o unos pocos clonotipos. Aproximadamente, la clonalidad es una medida inversa de la diversidad de clonotipo. Muchas medidas o estadísticas están disponibles a partir de la ecología que describe las relaciones de especies-abundancia que pueden usarse para mediciones clonalidad de acuerdo con la invención, por ejemplo, capítulas 17 y 18, en Pielou, An Introduction to Mathematical Ecology, (Wiley-Interscience, 1969). En un aspecto, una medición de clonalidad usada en la invención es una función de un perfil de clonotipo (es decir, el número de distintos clonotipos detectados y sus abundancias), de modo que después de medir un perfil de clonotipo, se puede informatizar la clonalidad a partir de este para proporcionar un único número. Una medida de clonalidad es la medida de Simpson, que es simplemente la probabilidad de que dos clonotipos extraídos aleatoriamente sean los mismos. Otras medidas de clonalidad incluyen medidas a base de información y el índice de diversidad de McIntosh, que se desvela en Pielou (citado anteriormente).

"Clonotipo" se refiere a un ácido nucleico recombinado de un linfocito que codifica un receptor inmunitario o una porción del mismo. Más particularmente, clonotipo se refiere a un ácido nucleico recombinado, normalmente extraído de un linfocito T o un linfocito B, pero que también puede ser de una fuente sin células, que codifica un receptor de linfocitos T (RLT) o un receptor de linfocitos B (RLB) o una porción del mismo. En diversas realizaciones, los clonotipos pueden codificas toda o una porción de una reorganización VDJ de IgH, una reorganización DJ de IgH, una reorganización VJ de IgK, una reorganización VJ de IgL, una reorganización VDJ de RLT β, una reorganización DJ de RLT β, una reorganización VJ de RLT α, una reorganización VJ de RLT γ, una reorganización VDJ de RLT δ, una reorganización VD de RLT δ, una reorganización Kde-V o similares. Los clonotipos también pueden codificar regiones de puntos de rotura de translocación que implican genes de receptores inmunitarios, tales como Bell-IgH o Bell-IgH. En un aspecto, los clonotipos tiene secuencias que son lo suficientemente largas para representar o reflejar la diversidad de moléculas inmunitarias de las que derivan; por consiguiente, los clonotipos pueden variar ampliamente en longitud. En algunas realizaciones, los clonotipos longitudes en el intervalo de 25 a 400 nucleótidos; en otras realizaciones, los clonotipos tienen longitudes en el intervalo de 25 a 200 nucleótidos.

"Perfil de clonotipo" se refiere a una enumeración de distintos clonotipos y sus abundancias relativas que derivan de una población de linfocitos, donde, por ejemplo, la abundancia relativa puede expresar como una frecuencia en una población dada (es decir, un número entre 0 y 1). Típicamente, la población de linfocitos se obtiene a partir de una muestra de tejido. El término "perfil de clonotipo" se refiere a, pero más general que, el concepto de inmunología de "repertorio" inmunitario tal como se describe en las referencias, tales como las siguientes: Arstila y col., Science, 286: 958-961 (1999); Yassai y col., Immunogenetics, 61: 493-502 (2009); Kedzierska y col., Mol. Immunol., 45(3): 607-618 (2008); y similares. La expresión "perfil de clonotipo" incluye una amplia variedad de listados y abundancias de ácidos nucleicos que codifican receptores inmunitarios reorganizados, que pueden derivarse de subconjuntos seleccionados de linfocitos (por ejemplo, linfocitos de infiltración tisular, subconjuntos inmunofenotípicos, o similares) o que pueden codificar porciones de receptores inmunitarios que tienen diversidad reducida en comparación con receptores inmunitarios completos. En algunas realizaciones, los perfiles de clonotipo puede comprender al menos 10^3 de clonotipos distintos; en otras realizaciones, los perfiles de clonotipo puede comprender al menos 10^4 de clonotipos distintos; en otras realizaciones, los perfiles de clonotipo puede comprender al menos 10^5 de clonotipos distintos; en otras realizaciones, los perfiles de clonotipo puede comprender al menos 10^6 de clonotipos distintos. En dichas realizaciones, tales perfiles de clonotipo pueden comprender adicionalmente abundancias o frecuencias

relativas de cada uno de los distintos clonotipos. En un aspecto, un perfil de clonotipo es un conjunto de distintas secuencias de nucleótidos recombinados (con sus abundancias) que codifican receptores de linfocitos T (RLT) o receptores de linfocitos B (RLB) o fragmentos de los mismos, respectivamente, en una población de linfocitos de un individuo, en donde las secuencias de nucleótidos del conjunto tienen una correspondencia uno-a-uno con distintos linfocitos o sus subpoblaciones clonales para sustancialmente todos los linfocitos de la población. En un aspecto, los segmentos de ácidos nucleicos que definen clonotipos se seleccionan de modo que su diversidad (es decir, el número de distintas secuencias de ácidos nucleicos en el conjunto) es lo suficientemente grande de modo que sustancialmente cada linfocito T o linfocito B o clon del mismo en un individuo porta una única secuencia de ácido nucleico de tal repertorio. Es decir, preferentemente cada clon distinto de una muestra tiene un clonotipo distinto. En otros aspectos de la invención, la población de linfocitos que se corresponden con un repertorio puede ser linfocitos B circulantes o pueden ser linfocitos T circulantes o pueden ser subpoblaciones de o bien las poblaciones anteriores, incluyendo, aunque no de forma limitativa, linfocitos T CD4+ o linfocitos T CD8+ o bien otras poblaciones definidas mediante marcadores de superficie celular o similares. Tales poblaciones pueden adquirirse tomando muestras de tejidos particulares, por ejemplo, médula ósea, ganglios linfáticos o similares o clasificando o enriqueciendo células de una muestra (tal como sangre periférica) basándose en uno o más marcadores de superficie celular, tamaño, morfología o similar. En aún otros aspectos, la población de linfocitos que se corresponde con un repertorio puede derivar de tejidos enfermos, tales como un tejido tumoral, un tejido infectado o similar. En una realización, un perfil de clonotipo que comprende cadenas de RLT β de ser humano o fragmentos de las mismas comprende un número de distintas secuencias de nucleótidos en el intervalo de 0,1 x 10^6 a 1,8 x 10^6, o en el intervalo de 0,5 x 10^6 a 1,5 x 10^6, o en el intervalo de 0,8 x 10^6 a 1,2 x 10^6. En otra realización, un perfil de clonotipo que comprende cadenas de lgH de ser humano o fragmentos de las mismas comprende un número de distintas secuencias de nucleótidos en el intervalo de 0,1 x 10^6 a 1,8 x 10^6, o en el intervalo de 0,5 x 10^6 a 1,5 x 10^6, o en el intervalo de 0,8 x 10^6 a 1,2 x 10^6. En una realización particular, un perfil de clonotipo de la invención comprende un conjunto de secuencias de nucleótidos que codifican sustancialmente todos los segmentos de la región V(D)J de una cadena lgH. En un aspecto, "sustancialmente todo" tal como se utiliza en el presente documento se refiere a cada segmento que tiene una abundancia relativa del ,001 por ciento o superior; o en otro aspecto, "sustancialmente todo" tal como se utiliza en el presente documento se refiere a cada segmento que tiene una abundancia del ,0001 por ciento o superior. En otra realización particular, un perfil de clonotipo de la invención comprende un conjunto de secuencias de nucleótidos que codifica sustancialmente todos los segmentos de la región V(D)J de una cadena de RLT β. En otra realización, un perfil de clonotipo de la invención comprende un conjunto de secuencias de nucleótidos que tiene longitudes en el intervalo de 25-200 nucleótidos y que incluye segmentos de las regiones V, D y J de una cadena de RLT β. En otra realización, un perfil de clonotipo de la invención comprende un conjunto de secuencias de nucleótidos que tiene longitudes en el intervalo de 25-200 nucleótidos y que incluye segmentos de las regiones V, D y J de una cadena de lgH. En otra realización, un perfil de clonotipo de la invención comprende un número de distintas secuencias de nucleótidos que es sustancialmente equivalente al número de linfocitos que expresan una cadena de lgH distinta. En otra realización, un perfil de clonotipo de la invención comprende un número de distintas secuencias de nucleótidos que es sustancialmente equivalente al número de linfocitos que expresa una cadena de RLT β distinta. En otra realización adicional más, "sustancialmente equivalente" se refiere a que con el noventa y nueve por ciento de probabilidad un perfil de clonotipo incluirá una secuencia de nucleótidos que codifique una lgH o RLT β o porción del mismo portado o expresado por cada linfocito de una población de un individuo a una frecuencia del ,001 por ciento o superior. En otra realización adicional más, "sustancialmente equivalente" se refiere a que el con el noventa y nueve por ciento de probabilidad un repertorio de secuencias de nucleótidos incluirá una secuencia de nucleótidos que codifique una lgH o RLT β o porción del mismo portado o expresado por cada linfocito presente a una frecuencia del ,0001 por ciento o superior. En algunas realizaciones, los perfiles de clonotipo derivan de muestras que comprenden de 10^5 a 10^7 linfocitos. Tales números de linfocitos pueden obtenerse a partir de muestras de sangre periférica de 1-10 ml.

"Regiones de determinación de complementariedad" (CDR) se refieren a regiones de una inmunoglobulina (es decir, anticuerpo) o receptor de linfocitos T donde la molécula complementa una conformación del antígeno, determinando, de este modo, la especificidad de la molécula y contacto con un antígeno específico. Los receptores de linfocitos T e inmunoglobulinas tienen cada una tres CDR: CDR1 y CDR2 se encuentran en el dominio variable (V) y CDR3 incluye algunos de los dominios B, todos los diversos (D) (cadenas pesadas solo) y de unión (J) y algunos dominios constantes (C).

"Base de datos de clonotipos" se refiere a una colección de clonotipos formateados y organizados para su facilidad y velocidad de búsqueda, comparación y recuperación. En algunas realizaciones, la base de datos de clonotipos comprende una colección de clonotipos que codifica la misma región o segmento de un receptor inmunitario. En algunas realizaciones, la base de datos de clonotipos comprende clonotipos de perfiles de clonotipo de una pluralidad de individuos. En algunas realizaciones, una base de datos de clonotipos comprende clonotipos de perfiles de clonotipo de al menos 10^4 clonotipos de al menos 10 individuos. En algunas realizaciones, una base de datos de clonotipos comprende al menos 10^6 clonotipos o al menos 10^8 clonotipos, al menos 10^9 clonotipos, o al menos 10^{10} clonotipos. Una base de datos de clonotipos puede ser una base de datos pública que contiene clonotipos, tal como la base de datos IMGT (www.imgt.org), por ejemplo, descrita en Nucleic Acids Research, 31: 307-310 (2003). Las bases de datos de clonotipos pueden ser en un formato FASTA y las entradas de las bases de datos de clonotipos pueden buscarse o compararse usando un algoritmo BLAST, por ejemplo, Altschul y col., J. Mol. Biol., 215(3): 403-410 (1990), o algoritmo similar.

"Fusionar" se refiere tratar dos clonotipos candidatos con diferencia de secuencia como el mismo determinando que tales diferencias se deben a un error experimental o de medición y no debido a diferencias biológicas genuinas. En un aspecto, una secuencia de un clonotipo candidato de frecuencia superior se compara con la de un clonotipo candidato de frecuencia inferior y si se cumples los criterios predeterminados entonces el número de clonotipos candidatos de frecuencia inferior se añade al del clonotipo candidato de frecuencia superior y el clonotipo candidato de frecuencia inferior se descarta a continuación. Es decir, los recuentos de lectura asociados con el clonotipo candidato de frecuencia inferior se añaden a los del clonotipo candidato de frecuencia superior y el clonotipo candidato de frecuencia superior y el clonotipo candidato de frecuencia inferior se tratan como el mismo; es decir, la diferencia observada entre ellos se determina que se debe a un error (por ejemplo, error de secuenciación, error de amplificación o similar). En algunas realizaciones, los criterios predeterminados es una función de probabilidad que depende de factores tales como frecuencias relativas de los clonotipos candidato que se están comparando, el número de posiciones en las que los candidatos difieren, las puntuaciones de calidad de las posiciones y similares.

"Regiones de determinación de complementariedad" (CDR) se refieren a regiones de una inmunoglobulina (es decir, anticuerpo) o receptor de linfocitos T donde la molécula complementa una conformación del antígeno, determinando, de este modo, la especificidad de la molécula y contacto con un antígeno específico. Los receptores de linfocitos T e inmunoglobulinas tienen cada una tres CDR: CDR1 y CDR2 se encuentran en el dominio variable (V) y CDR3 incluye algunos de los dominios B, todos los diversos (D) (cadenas pesadas solo) y de unión (J) y algunos dominios constantes (C).

"Contaminación" tal como se utiliza en el presente documento se refiere a la presencia en una muestra de tejido de un individuo de ácido nucleico de otro individuo. En un aspecto, "contaminación" se refiere a la presencia de ácido nucleico que no se origina del paciente que puede afectar en la interpretación del perfil de clonotipo del paciente.
"Identificación genética" se refiere a una correspondencia única entre un individuo y un conjunto de valores (o estados) de marcadores genéticos a partir de uno o más loci genéticos del individuo.

"Marcador genético" se refiere a un segmento polimórfico de ADN en un locus genético, que puede usarse para identificar un individuo. Un marcador genético puede identificarse por su secuencia o por sus secuencias adyacentes o flanqueantes. Típicamente, un marcador genético puede tener una pluralidad de secuencias o valores, en distintos individuos de una población. Marcadores genéticos ejemplares incluyen, pero sin limitación, repeticiones en tándem cortas (SRT), polimorfismos de nucleótidos unidos (SNP) y similares. El segmento polimórfico de ADN puede ser ADN genómico o puede ser ARN de transcripción inversa. En una realización, el segmento polimórfico es ADN genómico. En una realización, el marcador genético para su uso con la invención se identifica mediante amplificación y secuenciación usando técnicas convencionales. En otra realización, los marcadores genéticos se amplifican y secuencian junto con moléculas inmunitarias durante el proceso para generar un perfil de clonotipo.

"Estándar interno" se refiere a una secuencia de ácidos nucleicos que se procesa en la misma reacción que uno o más polinucleótidos diana para permitir la cuantificación absoluta o relativa de los polinucleótidos diana en una muestra. En un aspecto la reacción es una reacción de amplificación, tal como PCR. Un estándar interno puede ser endógeno o exógeno. Es decir, un estándar interno puede producirse naturalmente en la muestra o puede añadirse a la muestra antes de una reacción. En un aspecto, una o más secuencias estándar internas exógenas puede añadirse a una mezcla de reacción en concentraciones predeterminada para proporcionar una calibración con la que se puede comparar una secuencia amplificada para determinar la cantidad de su polinucleótido diana correspondiente en una muestra. La selección del número, secuencias, longitud y otras características de estándares internos exógenos es una elección de diseño habitual para un experto en la técnica. Los estándares internos endógenos, también denominados en el presente documento como "secuencias de referencia", son secuencias naturales con respecto a una muestra que se corresponde con genes mínimamente regulados que muestran un nivel de transcripción constante y independiente del ciclo celular, por ejemplo, Selvey y col., Mol. Cell Probes, 15: 307-311 (2001). Estándares internos ejemplares incluyen, pero sin limitación, secuencias a partir de los siguientes genes: GAPDH, β_2-microglobulina, ARN ribosómico 18S y β-actina.

"Kit" se refiere a cualquier sistema de suministro para suministrar materiales o reactivos para llevar a cabo el método de la invención. En el contexto de la invención, tales sistemas de suministro incluyen sistemas que permiten el almacenamiento, transporte o suministro de reactivos de reacción (por ejemplo, cebadores, enzimas, estándares internos, etc. en los recipientes adecuados) y/o materiales de soporte (por ejemplo, tampones, instrucciones por escrito para llevar a cabo el ensayo, etc.) desde un emplazamiento a otro. Por ejemplo, los kits incluyen una o más envolturas (por ejemplo, cajas) que contienen los reactivos de reacción relevantes y/o materiales de soporte. Tal contenido puede suministrarse al recipiente previsto de forma conjunta o separada. Por ejemplo, un primer recipiente puede contener una enzima para su uso en un ensayo, mientras que un segundo recipiente contiene cebadores.

"Enfermedad mínima residual" se refiere a células cancerosas que permanecen después del tratamiento. La expresión se usa frecuentemente junto con el tratamiento de linfomas y leucemias.

"Trastorno proliferativo linfoide o mieloide" se refiere a cualquier trastorno proliferativo anormal en el que una o más secuencias de nucleótidos que codifican uno o más receptores inmunitarios reorganizados pueden usarse como un marcador para controlar tal trastorno. "Neoplasma linfoide o mieloide" se refiere a una proliferación anormal de

células de linfocitos o mieloides que pueden ser malignas o no malignas. Un cáncer linfoide es un neoplasma linfoide maligno. Un cáncer mieloide es un neoplasma mieloide maligno. Los neoplasmas linfoides y mieloides son el resultado de, o están asociados con, trastornos linfoproliferativos o mieloproliferativos e incluyen, pero sin limitación, linfoma folicular, leucemia linfocítica crónica (CLL), leucemia linfocítica aguda (ALL), leucemia mielógena crónica (CML), leucemia mielógena aguda (AML), linfomas de Hodgkin y de no Hodgkin, mieloma múltiple (MM), gamopatía monoclonal de significancia indeterminada (MGUS), linfoma de células del manto (MCL), linfoma difuso de linfocitos B grandes (DLBCL), síndromes mielodisplásicos (MDS), linfoma de linfocitos T o similares, por ejemplo, Jaffe y col., Blood, 112: 4384-4399 (2008); Swerdlow y col., WHO Classification of Tumours of Haematopoietic and Lymphoid Tissues (4ª ed.) (IARC Press, 2008). Tal como se utiliza en el presente documento, "cáncer de linfocitos B" se refiere a un neoplasma linfoide o mieloide que implica linfocitos B o células derivadas de los mismos, tales como células plasmáticas. Del mismo modo, "cáncer de linfocitos T" se refiere a un neoplasma linfoide o mieloide que implica linfocitos T o células derivadas de los mismos.

"Porcentaje homólogo", "porcentaje idéntico", o términos similares se usan en referencia a la comparación de una secuencia de referencia y otra secuencia ("secuencia de comparación") que se refieren a que en un alineamiento óptico entre las dos secuencias, la secuencia de comparación es idéntica a la secuencia de referencia en un número de posiciones de subunidades equivalentes a las del porcentaje indicado, siendo las subunidades nucleótidos para comparaciones de polinucleótidos o aminoácidos para comparaciones de polipéptidos. Tal como se utiliza en el presente documento, un "alineamiento óptico" de secuencia que se está comparando es uno que maximiza las coincidencias entre subunidades y minimiza el número de huecos empleados en la construcción de una alineación. El porcentaje de identidades puede determinarse con implementaciones disponibles en el mercado de algoritmos, tales como los que se describe por Needleman y Wunsch, J. Mol. Biol., 48: 443-453 (1970)("GAP" program of Wisconsin Sequence Analysis Package, Genetics Computer Group, Madison, WI), o similares. Otros paquetes de software en la técnica para construir alineamiento y calcular el porcentaje de identidad u otras mediciones de similitud incluyen el programa "BestFit", que se basa en el algoritmo de Smith y Waterman, Advances in Applied Mathematics, 2: 482-489 (1981) (Wisconsin Sequence Analysis Package, Genetics Computer Group, Madison, WI). En otras palabras, por ejemplo, para obtener un polinucleótido que tenga una secuencia de nucleótidos de al menos el 95 por ciento de identidad con una secuencia de nucleótidos de referencia, hasta el cinco por ciento de nucleótidos en la secuencia de referencia pueden eliminarse o sustituirse con otro nucleótido, o un número de nucleótidos de hasta el cinco por ciento del número total de nucleótidos en la secuencia de referencia puede insertarse en la secuencia de referente.

"Reacción en cadena de la polimerasa", o "PCR", se refiere a una reacción para la amplificación *in vitro* de secuencias de ADN específicas mediante la extensión de cebador simultánea de cadenas complementarias de ADN. En otras palabras, la PCR es una reacción para realizar múltiples copias o réplicas de un ácido nucleico recombinado flanqueado por sitios de unión de cebador, comprendiendo tal reacción una o más repeticiones de las siguientes etapas: (i) desnaturalizar el ácido nucleico diana, (ii) hibridar los cebadores con los sitios de unión de cebador y (iii) extender los cebadores mediante una polimerasa de ácido nucleico en presencie de trifosfatos de nucleosida. Tal como se utiliza en el presente documento, expresiones "cebador directos" y "cebador corriente arriba" se usan de forma indistinta y las expresiones "cebador indirecto" y "cebador corriente abajo" se usan de forma indistinta. También, tal como se utiliza en el presente documento, si se muestra un polinucleótido diana bicatenario con su cadena de sentido en una orientación de izquierda a derecha 5'→3', un cebador directo se uniría a la cadena antisentido sobre el lado izquierdo y se extendería a la derecha y un cebador indirecto se uniría a la cadena de sentido sobre el lado derecho y se extendería hacia la izquierda. Normalmente, la reacción se somete a ciclos a través de distintas temperaturas optimizadas para cada etapa en un instrumento de ciclos térmico. Temperaturas particulares, duraciones en cada etapa y tasa de cambio entre etapas depende de muchos factores bien conocidos por los expertos en la técnica, por ejemplo, ejemplificados por las referencias: McPherson y col., editores, PCR: A Practical Approach and PCR2: A Practical Approach (IRL Press, Oxford, 1991 y 1995, respectivamente). Por ejemplo, en una PCR convencionales que usa ADN polimerasa Taq, puede desnaturalizarse un ácido nucleico diana bicatenario a una temperatura de >90 °C, los cebadores hibridados a una temperatura en el intervalo de 50-75 °C y los cebadores extendidos a una temperatura en el intervalo de 72-78 ºC. El término "PCR" abarca formas derivadas de la reacción, incluyendo, aunque no de forma limitativa, RT-PCR, PCR en tiempo real, PCR anidada, PCR cuantitativa, PCR multiplexada y similares. Los volúmenes de reacción varían desde unos cientos de nanolitros, por ejemplo, 200 nl a unos pocos cientos de µl, por ejemplo, 200 µl. "PCR de transcripción inversa", o "RT-PCR", se refiere a una PCR que está precedida por una reacción de transcripción inversa que convierte una ARN diana en una ADN unicatenario complementario, que, a continuación, se amplifica, por ejemplo, Tecott y col., patente de los EE.UU. 5.168.038, cuya patente se incorpora al presente documento por referencia. "PCR en tiempo real" se refiere a una PCR para cual la cantidad de producto de reacción, es decir, amplicón, se controla según procede la reacción. Existen muchas formas de PCR en tiempo real ue difieren principalmente en las químicas de detección usadas para controlar el producto de reacción, por ejemplo, Gelfand y col., patente de los EE.UU. 5.210.015 ("taqman"); Wittwer y col., patentes de Estados Unidos 6.174.670 y 6.569.627 (tintes de intercalado); Tyagi y col., patente de Estados Unidos 5.925.517 (balizas moleculares); cuyas patentes se incorporan al presente documento por referencia. Las químicas de detección para PCR en tiempo real se revisan en Mackay y col., Nucleic Acids Research, 30: 1292-1305 (2002), que también se incorpora en el presente documento por referencia. "PCR anidada" se refiere a una PCR de dos etapas en donde el amplicón de un primer PCR se convierte en la muestra para la segunda PCR que usa un nuevo conjunto de cebadores, al menos uno del cual se une en un

emplazamiento interior del primer amplicón. Tal como se utiliza en el presente documento, "cebadores iniciales" en referencia a una reacción de amplificación anidada se refiere a los cebadores usados para generar un primer amplicón y "cebadores secundarios" se refiere al uno o más cebadores usados para generar un segundo, o anidado, amplicón. "PCR multiplexada" se refiere a una PCR en donde múltiples secuencias diana (o una única secuencia diana y una o más secuencias de referencia) se llevan a cabo simultáneamente en la misma mezcla de reacción, por ejemplo, Bernard y col., Anal. Biochem., 273: 221-228 (1999)(PCR en tiempo real de dos colores). Normalmente, se emplean distintos conjuntos de cebadores para cada secuencia que se está amplificando. "PCR cuantitativa" se refiere a una PCR designada para medir la abundancia de una o más secuencias diana específicas en una muestra o espécimen. La PCR cuantitativa incluye tanto la cuantificación absoluta como la cuantificación relativa de tales secuencias diana. Las mediciones cuantitativas se realizan usando unas o más secuencias de referencia o estándares internos que pueden someterse a ensayo por separado o juntos con una secuencia diana. La secuencia de referencia puede ser endógena o exógena con respecto a una muestra o espécimen y, en el último caso, puede comprender uno o más modelos de competidor. Secuencias de referencia endógenas típicas incluyen segmentos de transcripciones de los siguientes genes: β-actina, GAPDH, β_2-microglobulina, ARN ribosómico y similares. Las técnicas para la PCR cuantitativa son bien conocidos de aquellas personas normalmente expertas en la materia, como se ejemplifican en las siguientes referencias: Freeman y col., Biotechniques, 26: 112-126 (1999); Becker-Andre y col., Nucleic Acids Research, 17: 9437-9447 (1989); Zimmerman y col., Biotechniques, 21: 268-279 (1996); Diviacco y col., Gene, 122: 3013-3020 (1992); Becker-Andre y col., Nucleic Acids Research, 17: 9437-9446 (1989); y similares.

"Cebador" se refiere a un oligonucleótido, o bien natural o bien sintético que es capaz, cuando se forma un híbrido con un modelo de polinucleótido, de actuar como punto de inicio de la síntesis de ácidos nucleico y de extenderse desde su extremo 3' a lo largo del modelo de modo que se forma un híbrido extendido. La extensión de un cebador se lleva a cabo normalmente con una polimerasa de ácido nucleico, tal como ADN o ARN polimerasa. La secuencia de nucleótidos añadida en el proceso de extensión se determina mediante la secuencia del polinucleótido modelo. Normalmente, los cebadores se extienden mediante ADN polimerasa. Los cebadores tienen normalmente una longitud en el intervalo de 14 a 40 nucleótidos o en el intervalo de 18 a 36 nucleótidos. Los cebadores se emplean en una variedad de reacciones de amplificación nucleicas, por ejemplo, las reacciones de amplificación lineal que usan un único cebador o reacciones de cadena de la polimerasa, que emplean dos o más cebadores. Directrices para seleccionar las longitudes y secuencias de cebadores para aplicaciones particulares son bien conocidos por los expertos en la técnica, tal como se muestra en las siguientes referencias: Dieffenbach, editor, PCR Primer: A Laboratory Manual, 2ª Edition (Cold Spring Harbor Press, Nueva York, 2003).

"Puntuación de calidad" se refiere a una medición de la probabilidad de que una asignación de base en un emplazamiento de secuencia particular sea correcta. Una variedad de métodos es bien conocida por el experto en la técnica para calcular puntuaciones de calidad para circunstancias particulares, tal como, las bases identificadas como resultado de distintas químicas de secuenciación, sistemas de detección, algoritmos de identificación de nucleótidos, etcétera. En general, los valores de puntuación de calidad están monotónicamente relacionados con las probabilidades de una identificación de nucleótidos correcta. Por ejemplo, una puntuación de calidad, o Q, de 10 puede significar que hay un 90 por ciento de probabilidad de que se un nucleótido se identifique correctamente, una Q de 20 puede significar que hay un 99 por ciento de probabilidad de que un nucleótido se identifique correctamente, etcétera. Para algunas plataformas de secuenciación, particularmente aquellas que usan químicas de secuenciación por síntesis, las puntuaciones de calidad promedio disminuyen como una función de la longitud de lectura de secuencia, de modo que las puntuaciones de calidad al inicio de una lectura de secuencia son superiores a las del final de una lectura de secuencia, debiéndose tal declive al fenómeno tal como extensiones incompletas, extensiones por arrastre, perdida de modelo, pérdida de polimerasa, fallo de protección con capuchón, fallos de desprotección y similares.

"Lectura de secuencia" se refiere a una secuencia o nucleótidos determinados a partir de una secuencia o corriente de datos generada mediante una técnica de secuenciación, cuta determinación se realiza, por ejemplo, por medio de un software de lectura de nucleótidos asociados con la técnica, por ejemplo, software de lectura de nucleótidos de un suministrador comercial de una plataforma de secuenciación de ADN. Una lectura de secuencia normalmente incluye puntuaciones de calidad para cada nucleótido en la secuencia. Típicamente, las lecturas de secuencia se realizan extendiendo un cebador a lo largo de un ácido nucleico modelo, por ejemplo, una ADN polimerasa o una ADN ligasa. Se generan datos registrando señales, tales como ópticas, químicas (por ejemplo, cambio en pH) o señales eléctricas, asociadas con tal extensión. Tales datos iniciales se convierten en una lectura de secuencia.

"Marcador de secuencia" (o "marcador") o "código de barras" se refiere a un oligonucleótido que está unido a un polinucleótido o molécula modelo y se usa para identificar y/o rastrear el polinucleótido o modelo en una reacción o una serie de reacciones. Cada marcador de secuencia tiene una secuencia de nucleótidos que a veces se denomina en el presente documento como una "secuencia de marcador". Un marcador de secuencia puede unirse en el extremo 3' o 5' de un polinucleótido o modelo o puede insertarse en el interior de tal polinucleótido o modelo para formar un conjugado lineal o circular, a menudo denominado en el presente documento como "polinucleótido marcado", o "modelo macado", o "conjugado de marcador-polinucleótido", "conjugado de marcador-molécula", o similar. Los marcadores de secuencia pueden variar ampliamente de tamaño y composiciones, las siguientes referencias, que se incorporan en el presente documento por referencia, proporcionan directrices para seleccionar

conjuntos de marcadores de secuencia adecuados para realizaciones particulares: Brenner, patente de los EE.UU. 5.635.400; Brenner y Macevicz, patente de los EE.UU. 7.537.897; Brenner y col., Proc. Natl. Acad. Sci., 97: 1665-1670 (2000); Church y col., publicación de patente europea 0 303 459; Shoemaker y col., Nature Genetics, 14: 450-456 (1996); Morris y col., publicación de patente europea 0799897A1; Wallace, patente de los EE.UU. 5.981.179; y similares. La selección de longitudes y/o composiciones de marcador particulares puede depender de varios factores entre los que se incluye, sin limitación, la tecnología de secuenciación usada para descodificar un marcador; el número de marcadores distinguibles requeridos para identificas de forma no ambigua un conjunto de polinucleótidos diana, cómo de diferentes deben ser los marcadores de un conjunto para asegurar una identificación fiable, por ejemplo, libertar de hibridación cruzada o identificación errónea a partir de errores de secuenciación y similares. En algunas realizaciones, los marcadores de secuencia pueden tener cada una, una longitud dentro del intervalo de 6 a 100 nucleótidos, o de 10 a 100 nucleótidos o de 12 a 50 nucleótidos o de 12 a 25 nucleótidos, respectivamente. En algunas realizaciones, se usan conjuntos de marcadores de secuencia en donde cada marcador se secuencia de un conjunto tiene una secuencia de nucleótidos única que difiere de cada uno de los otros marcadores distintos del mismo conjunto en al menos cuatro bases; en otras realizaciones, se usan conjuntos de marcadores de secuencia en donde cada uno de los marcadores de un conjunto difieren en cada un otro marcador del mismo conjunto en al menos cinco bases; en todavía otras realizaciones, se usan conjuntos de marcadores de secuencia en donde la secuencia de cada marcador difiere de cada otro marcador del mismo conjunto en al menos el diez por ciento de sus nucleótidos; o, en otras realizaciones, al menos el veinticinco por ciento de sus nucleótidos; o, en otras realizaciones, al menos el cincuenta por ciento de sus nucleótidos.

LISTADO DE SECUENCIAS

<110> Sequenta, Inc.
Asbury, Thomas
Hervold, Kieran
Kotwaliwale, Chitra
Faham, Malek
Moorhead, Martin
Weng, Li
Wittkop, Tobias
Zheng, Jianbiao
<120> ANÁLISIS BIOMOLECULAR A GRAN ESCALA CON MARCADORES DE SECUENCIA
<130> 848US00 (37623-739.201)
<150> 61/841878
<151> 01-07-2013
<150> 62/001580
<151>21-05-2014
<160> 6
<170> PatentIn versión 3.5
<210> 1
<211> 24
<212> ADN
<213> Secuencia artificial
<220>
<223> cebador
<400> 1

agttctggct aacctgtaga gcca

24

<210> 2
<211> 24
<212> ADN
<213> Secuencia artificial
<220>
<223> cebador
<400> 2

agttcgggct aacctgtcga gcca

24

<210> 3
<211> 24
<212> ADN

<213> Secuencia artificial
<220>
<223> cebador
<400> 3

agttccggct aacctgtcga gcca

24

<210> 4
<211> 22
<212> ADN
<213> Secuencia artificial
<220>
<223> cebador
<220>
<221> misc_feature
<222> (1) .. (22)
<223> n es a, c, g, ort
<400> 4

nnnnnnnnnn nnnnnnnnnn nn

22

<210> 5
<211> 12
<212> ADN
<213> Secuencia artificial
<220>
<223> cebador
<400> 5

gtatttttt ct

12

<210> 6
<211> 13
<212> ADN
<213> Secuencia artificial
<220>
<223> cebador
<400> 6

ttcaggggg gct

13

LISTADO DE SECUENCIAS

<110> Sequenta, Inc.
Asbury, Thomas
Hervold, Kieran
Kotwaliwale, Chitra
Faham, Malek
Moorhead, Martin
Weng, Li
Wittkop, Tobias
Zheng, Jianbiao

<120> ANÁLISIS BIOMOLECULAR A GRAN ESCALA CON MARCADORES DE SECUENCIA

<130> 848US00 (37623-739.601)

<150> 61/841878

<151> 01-07-2013

<150> 62/001580
<151>21-05-2014

<160> 6

<170> PatentIn versión 3.5

<210> 1
<211> 24
<212> ADN
<213> Secuencia artificial

<220>
<223> cebador

<400> 1
agttctggct aacctgtaga gcca 24

<210> 2
<211> 24
<212> ADN
<213> Secuencia artificial

<220>
<223> cebador

<400> 2
agttcgggct aacctgtcga gcca 24

<210> 3
<211> 24
<212> ADN
<213> Secuencia artificial

<220>
<223> cebador

<400> 3
agttccggct aacctgtcga gcca 24

<210> 4
<211> 22
<212> ADN
<213> Secuencia artificial

<220>
<223> cebador

<220>

<221> misc_feature
<222> (1)..(22)
<223> n es a, c, g, ort

<400> 4
nnnnnnnnnn nnnnnnnnnn nn 22

<210> 5
<211> 12
<212> ADN
<213> Secuencia artificial

<220>

<223> cebador

<400> 5
gtatttttt ct 12

<210> 6
<211> 13
<212> ADN
<213> Secuencia artificial

<220>
<223> cebador

<400> 6
ttcaggggg gct 13

REIVINDICACIONES

1. Un método de determinación de un perfil de clonotipo de ácidos nucleicos recombinados que codifican una pluralidad de cadenas de receptor inmunitario en una muestra, comprendiendo el método las etapas de:

(a) unir marcadores de secuencia a moléculas de ácido nucleico recombinadas de genes de receptores de linfocitos T o genes de inmunoglobulina a partir de una muestra de un individuo que comprende linfocitos T y/o linfocitos B y/o ADN sin células para formar conjugados de marcador-ácido nucleico, en donde al menos un ácido nucleico recombinado o copias del mismo tienen distintos marcadores de secuencia unidos;
(b) amplificar los conjugados de marcador-ácido nucleico;
(c) secuenciar una muestra de los conjugados de marcador-ácido nucleico para proporcionar lecturas de secuencias teniendo cada una, una tasa de error y comprendiendo cada una, una secuencia de marcados y una secuencia de ácido nucleico recombinado;
(g) alinear las lecturas de secuencias como secuencias de marcadores para formar grupos de lecturas de secuencia que tienen los mismos marcadores de secuencia;
(e) fusionar lecturas de secuencia de grupos para determinar clonotipos, en donde los grupos de las lecturas de secuencias de fusionan en distintas secuencias de ácidos nucleicos recombinados siempre que dichos grupos de lecturas de secuencias sean distintos con una probabilidad de al menos el noventa y cinco por ciento; y
(f) determinar el perfil de clonotipo de la muestra determinando los niveles de los clonotipos;

y en donde las etapas de unión y amplificación comprenden:

(i) combinar en una mezcla de reacción con condiciones de extensión de cebador un primer conjunto de cebadores con la muestra, en donde cada cebador del primer conjunto tiene una porción específica a receptor de modo que la porción específica a receptor se hibrida a un ácido nucleico recombinado distinta en un emplazamiento predeterminado y se extiende para formar un primer producto de extensión, y en donde cada cebador del primer conjunto tiene un extremo 5' no complementario que contiene un primer sitio de unión y un marcador de secuencia dispuestos entre la porción específica a receptor y el primer sitio de unión del cebador;
(ii) retirar de la mezcla de reacción cebadores no extendidos del primer conjunto;
(iii) añadir a la mezcla de reacción en condiciones de extensión de cebador un segundo conjunto de cebadores, en donde cada cebador del segundo conjunto tiene una porción específica a receptor de modo que la porción específica a receptor se hibrida al primer producto de extensión en un emplazamiento predeterminado y tiene un extremo 5' no complementario que contiene un segundo sitio de unión de cebador, y en donde cada cebador del segundo conjunto se extiende para formar un segundo producto de extensión, de modo que cada segundo producto de extensión comprende un primer sitio de unión de cebador, un segundo sitio de unión de cebador, al menos un marcador de secuencia y ácido nucleico recombinante que codifica una porción de una cadena de receptor de linfocitos T o una cadena de receptor de linfocitos B; y
(iv) llevar a cabo una reacción en cadena de la polimerasa en la mezcla de reacción para formar un amplicón, usando la reacción en cadena de la polimerasa cebadores directos específicos para el primer sitio de unión de cebador y cebadores indirectos específicos para el segundo sitió de unión de cebador.

2. El método de la reivindicación 1, que incluye adicionalmente una etapa de retirar de una mezcla de reacción cebadores no extendidos del segundo conjunto después de que dicho segundo producto de extensión se haya formado.

3. El método de la reivindicación 1, en donde la hibridación y la extensión de los cebadores del primer conjunto se repite después de fusionar el primer producto de extensión.

4. El método de la reivindicación 1, en donde la hibridación y la extensión de los cebadores del segundo conjunto se repite después de fusionar el segundo producto de extensión.

5. El método de la reivindicación 1 o la reivindicación 2, en donde la etapa de retiración incluye la digestión de exonucleasa, opcionalmente en donde la exonucleasa es exonucleasa I.

6. El método de la reivindicación 1, en donde las etapas de unión y amplificación generan conjugados de marcador-ácido nucleico en donde al menos un ácido nucleico recombinado y sus respectivas copias tienen una pluralidad de distintos marcadores de secuencia unidos.

7. El método de la reivindicación 6, en donde los ácidos nucleicos recombinados comprenden ácidos nucleicos recombinados que codifican cadenas de RLTβ, RLTδ y RLTγ y en donde los cebadores del primer conjunto y los cebadores del segundo conjunto comprenden cebadores que flanquean regiones de los ácidos nucleicos recombinados que codifican regiones VDJ de RLTβ y RLTδ y cebadores que flanquean regiones VJ de RLTγ.

8. El método de la reivindicación 6, en donde los ácidos nucleicos recombinados comprenden ácidos nucleicos recombinados que codifican lgH y lgK y en donde los cebadores del primer conjunto y los cebadores del segundo conjunto comprenden cebadores que flanquean regiones de los ácidos nucleicos recombinados que codifican

regiones VDJ de IgH, regiones DJ de IgH y regiones VJ de IgK.

9. El método de la reivindicación 1, en donde la etapa de fusión incluye fusionar los grupos de lecturas de secuencia en distintas secuencias de ácidos nucleicos recombinados siempre que dichos grupos de lecturas de secuencias sean distintos con una probabilidad de al menos el 99,9 por ciento.

10. El método de la reivindicación 8, en donde los cebadores del primer conjunto incluyen al menos un conjunto anidado de cebadores específico para una pluralidad de distintos sitios de unión de cebador en las regiones V de las cadenas de IgH.

11. El método de la reivindicación 1, en donde los cebadores indirectos específicos para el segundo sitio de unión de cebador comprenden un extremo 5' no complementario que comprende un marcador de muestra.

12. El método de la reivindicación 1, en donde los marcadores de secuencia son marcadores en mosaico, en donde los marcadores en mosaico comprenden alternar regiones constantes y regiones variables.

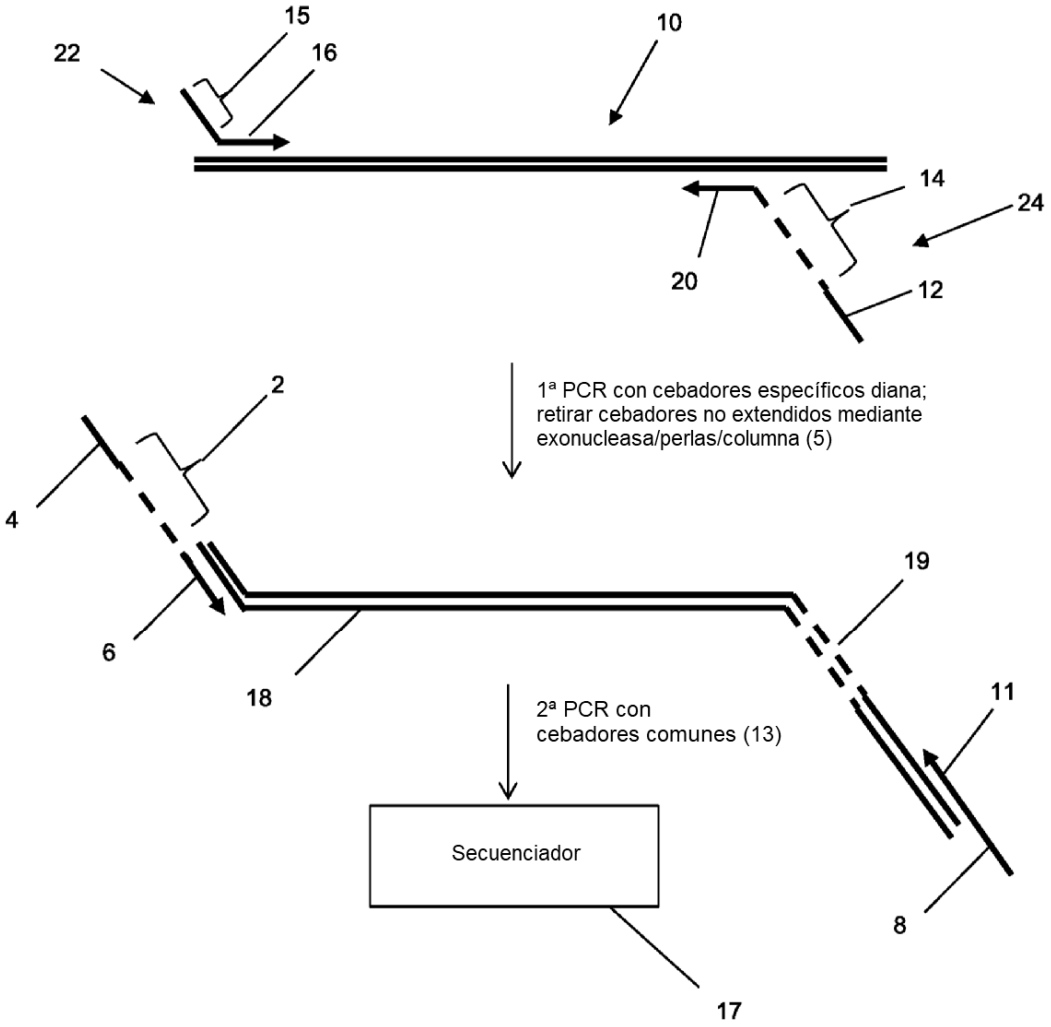

15
22
16
10
14
24
20
12
1ª PCR con cebadores específicos diana;
retirar cebadores no extendidos mediante
exonucleasa/perlas/columna (5)
2
4
19
6
18
11
2ª PCR con
cebadores comunes (13)
Secuenciador
8
17

Fig. 1A

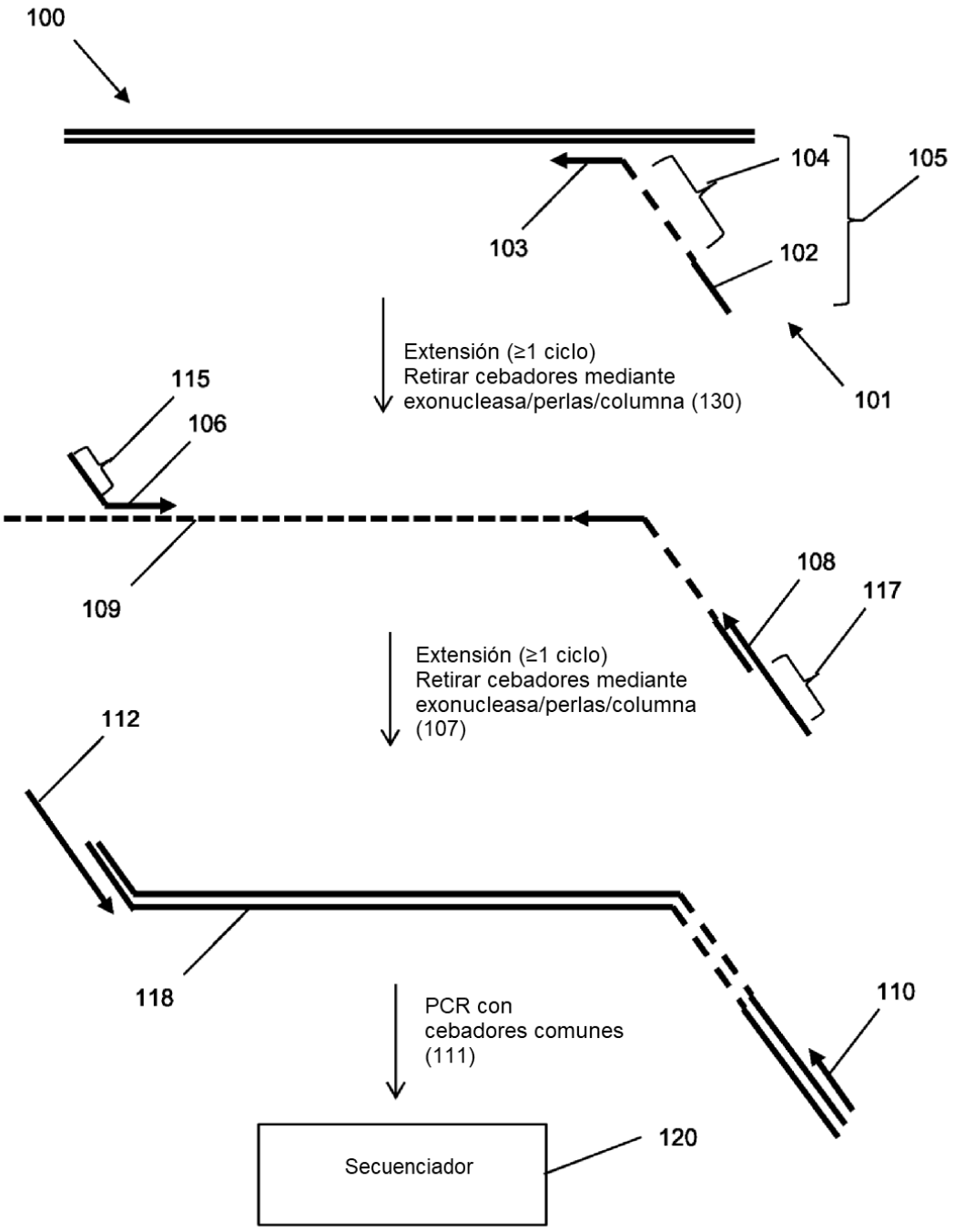

100
104
105
103
102
Extensión (≥1 ciclo)
Retirar cebadores mediante
exonucleasa/perlas/columna (130)
101
115
106
109
108
117
Extensión (≥1 ciclo)
Retirar cebadores mediante
exonucleasa/perlas/columna
(107)
112
118
110
PCR con
cebadores comunes
(111)
Secuenciador
120

Fig. 1B

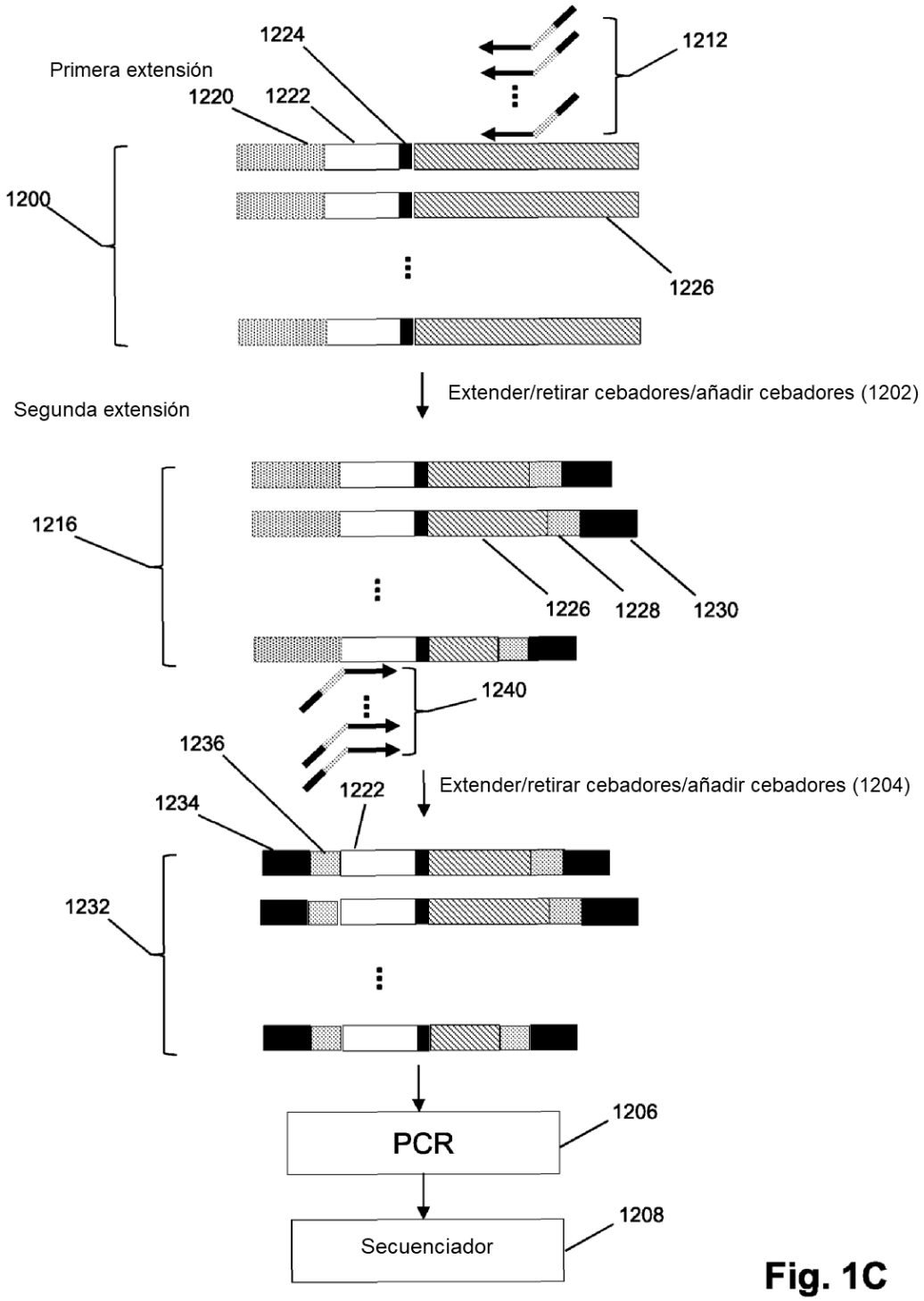

1224
1212
Primera extensión
1220
1222
1200
1226
Extender/retirar cebadores/añadir cebadores (1202)
Segunda extensión
1216
1226
1228
1230
1240
1236
1222
Extender/retirar cebadores/añadir cebadores (1204)
1234
1232
1206
PCR
1208
Secuenciador

Fig. 1C

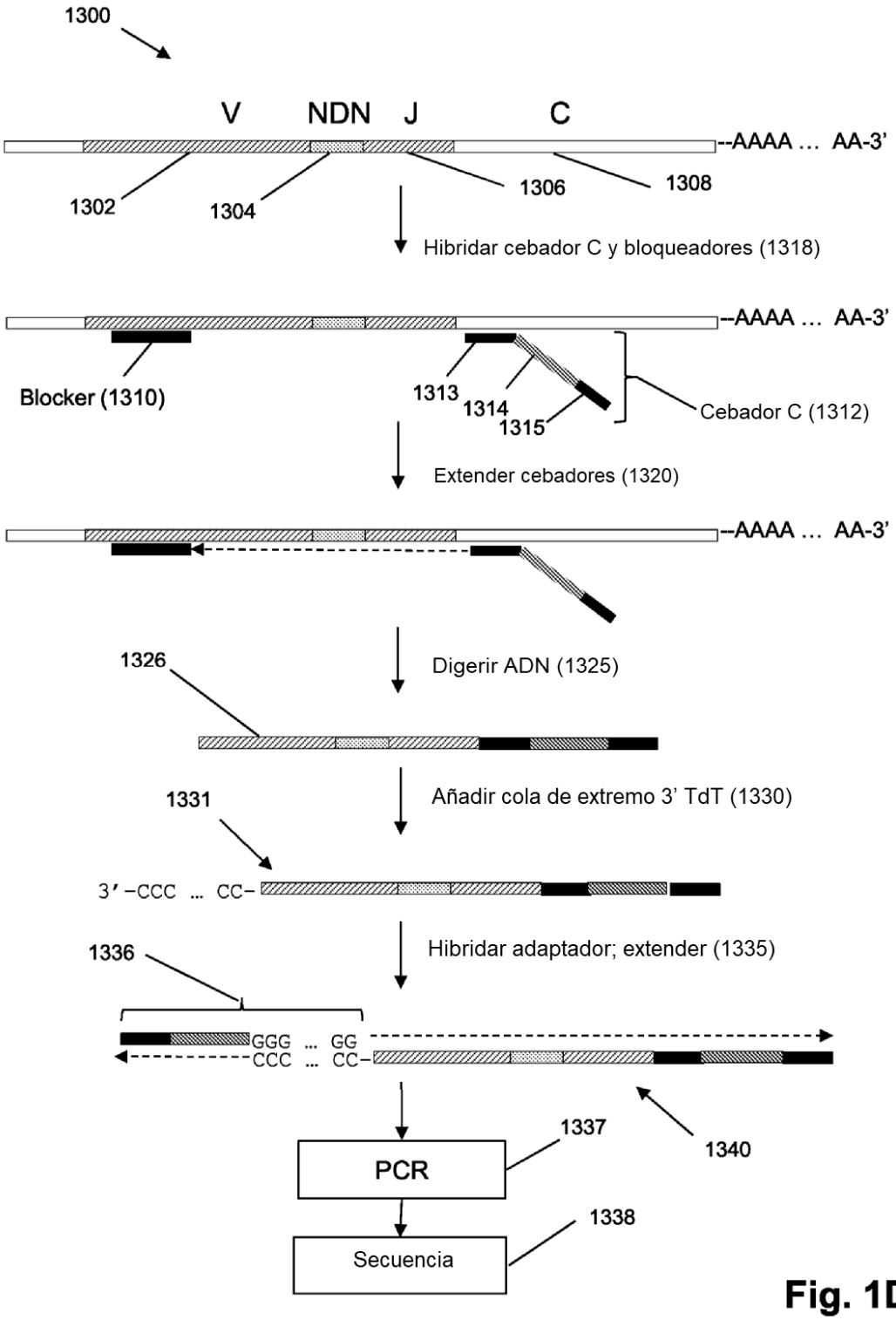

1300
V
NDN
J
C
--AAAA … AA-3'
1302
1304
1306
1308
Hibridar cebador C y bloqueadores (1318)
--AAAA … AA-3'
Blocker (1310)
1313
1314
1315
Cebador C (1312)
Extender cebadores (1320)
--AAAA … AA-3'
1326
Digerir ADN (1325)
1331
Añadir cola de extremo 3' TdT (1330)
3'-CCC … CC-
1336
Hibridar adaptador; extender (1335)
GGG … GG
CCC … CC-
1337
1340
PCR
1338
Secuencia

Fig. 1D

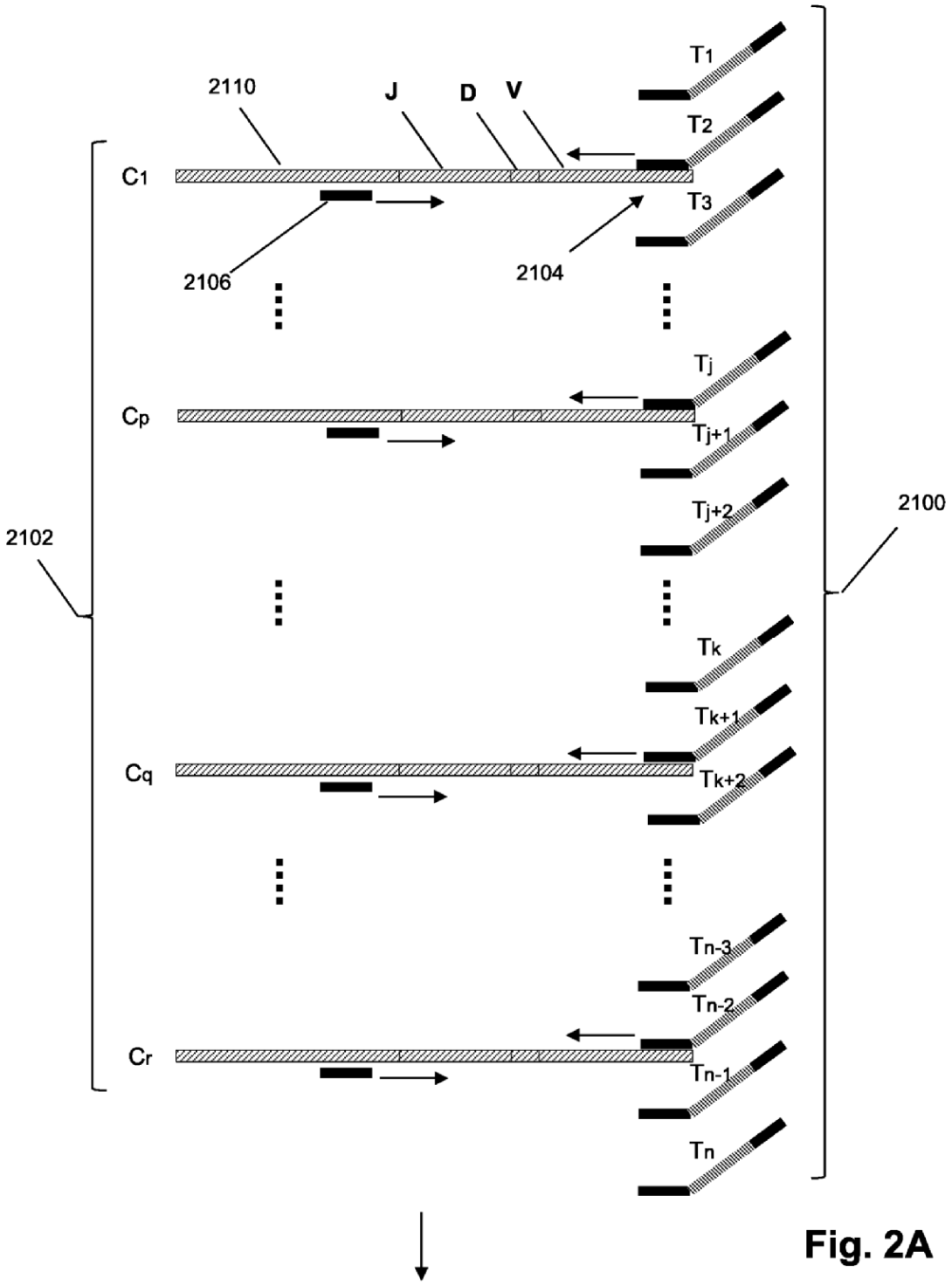

2110
J
D
V
T1
T2
T3
C1
2106
2104
Tj
Cp
Tj+1
Tj+2
2102
2100
Tk
Tk+1
Cq
Tk+2
Tn-3
Tn-2
Cr
Tn-1
Tn

Fig. 2A

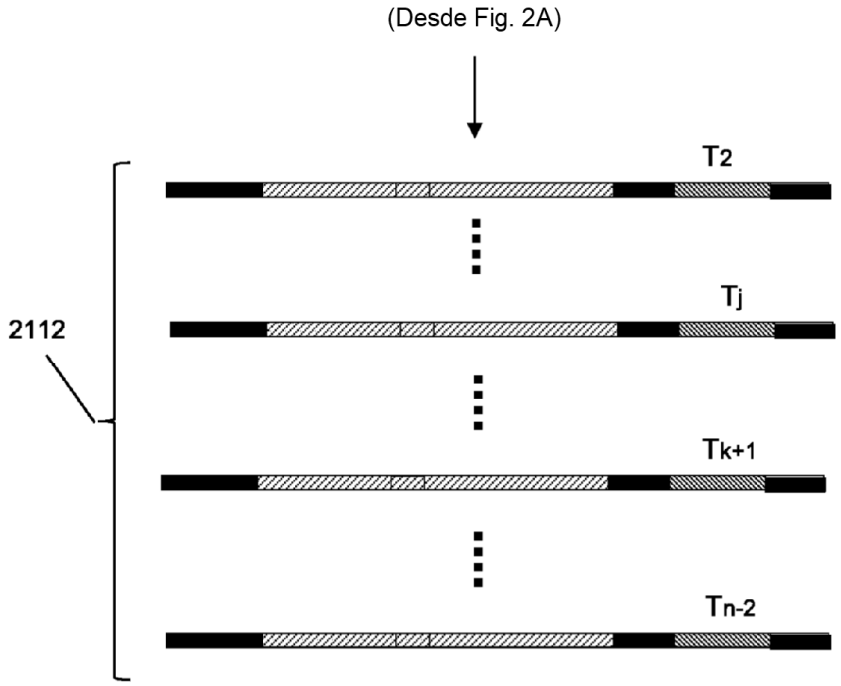


Fig. 2B

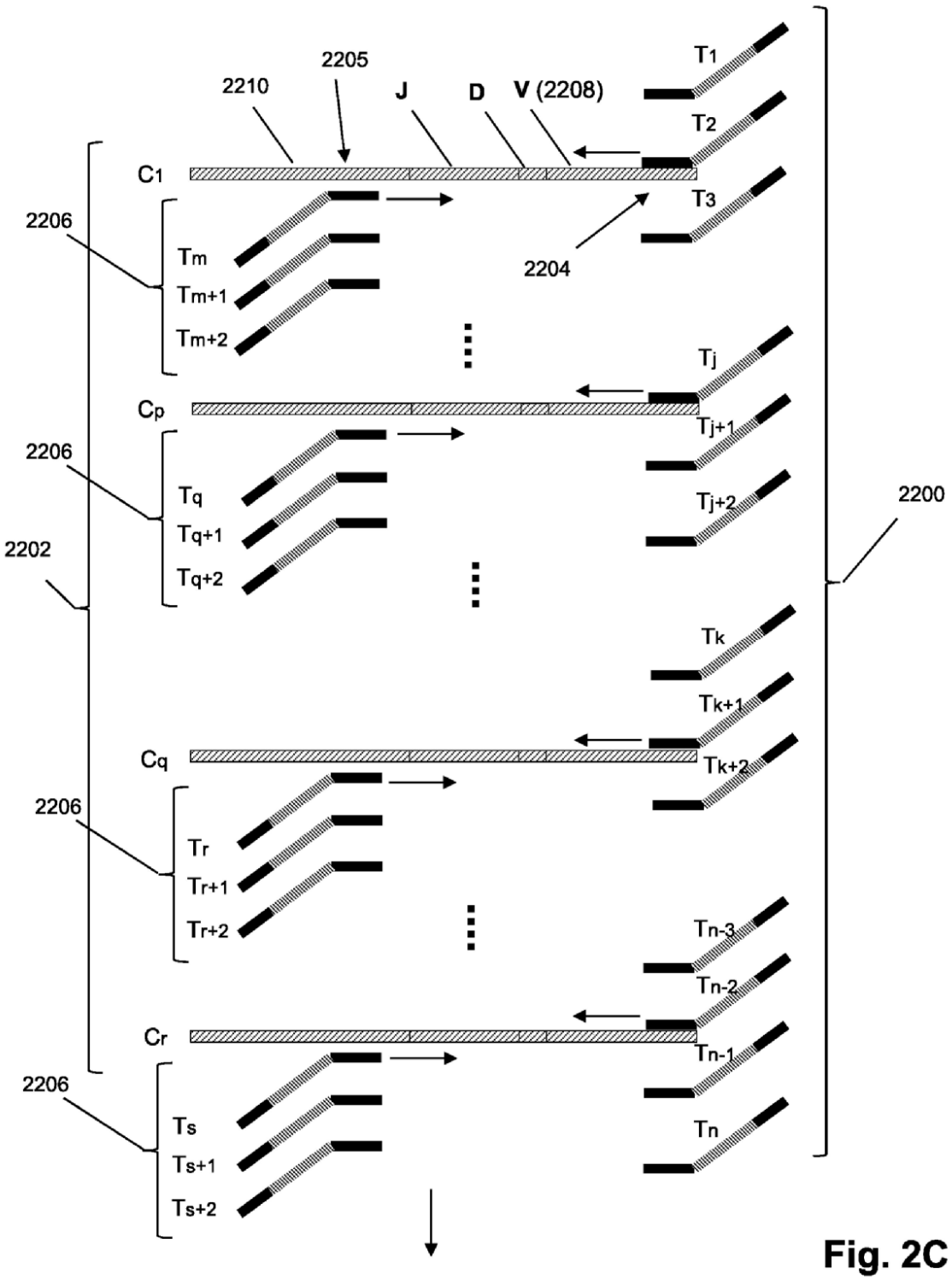

2205
2210
J
D
V (2208)
T1
T2
T3
C1
2206
Tm
Tm+1
Tm+2
2204
Tj
Cp
Tj+1
2206
Tq
Tq+1
Tq+2
Tj+2
2200
2202
Tk
Tk+1
Cq
Tk+2
2206
Tr
Tr+1
Tr+2
Tn-3
Tn-2
Cr
Tn-1
2206
Ts
Ts+1
Ts+2
Tn

Fig. 2C

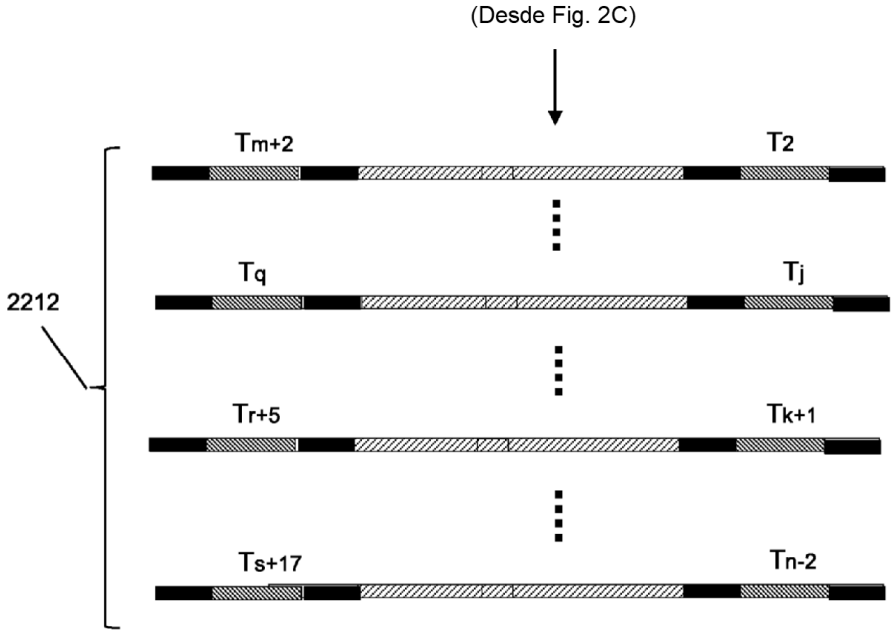


Fig. 2D

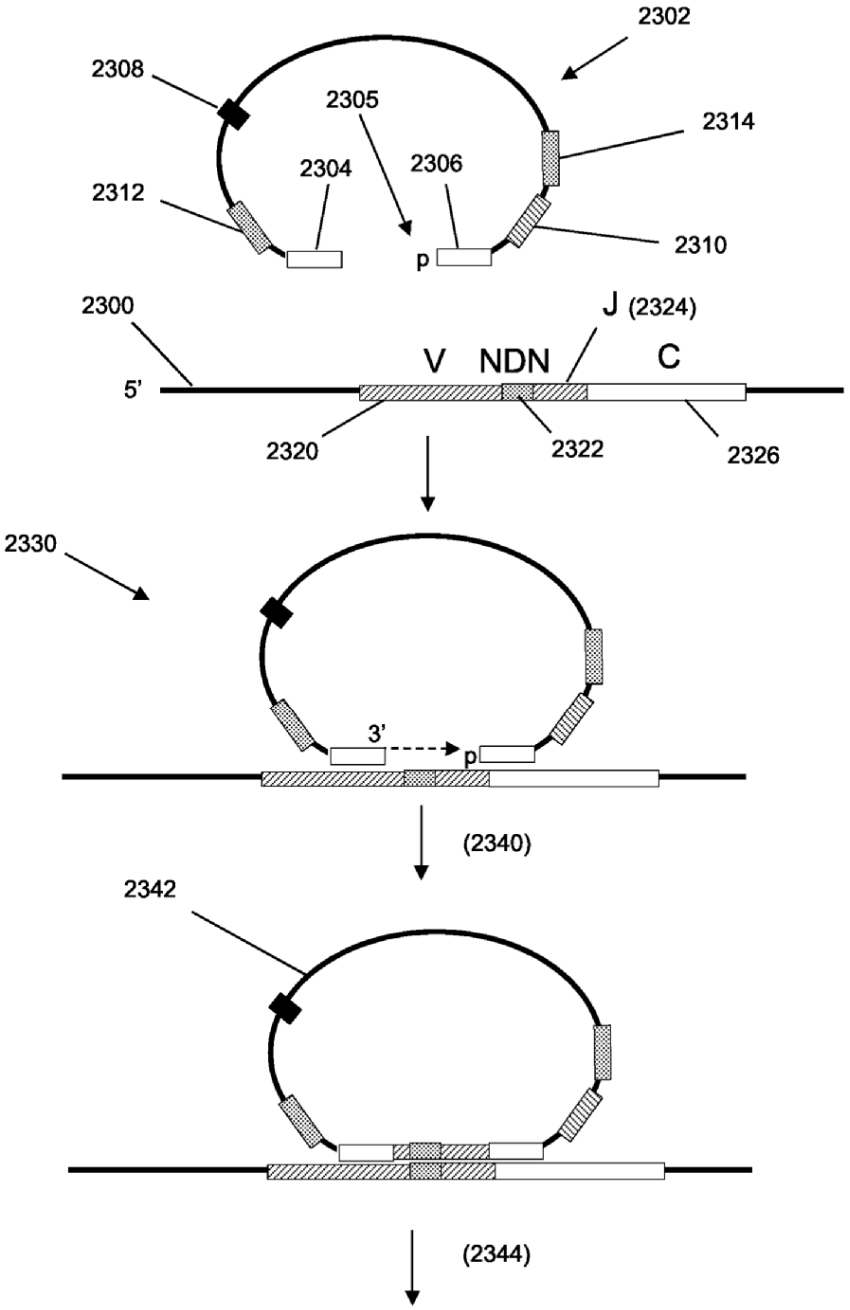

2302
2308
2305
2314
2304
2306
2312
2310
p
2300
J (2324)
V NDN C
5'
2320
2322
2326
2330
3'
p
(2340)
2342
(2344)

Fig. 2E

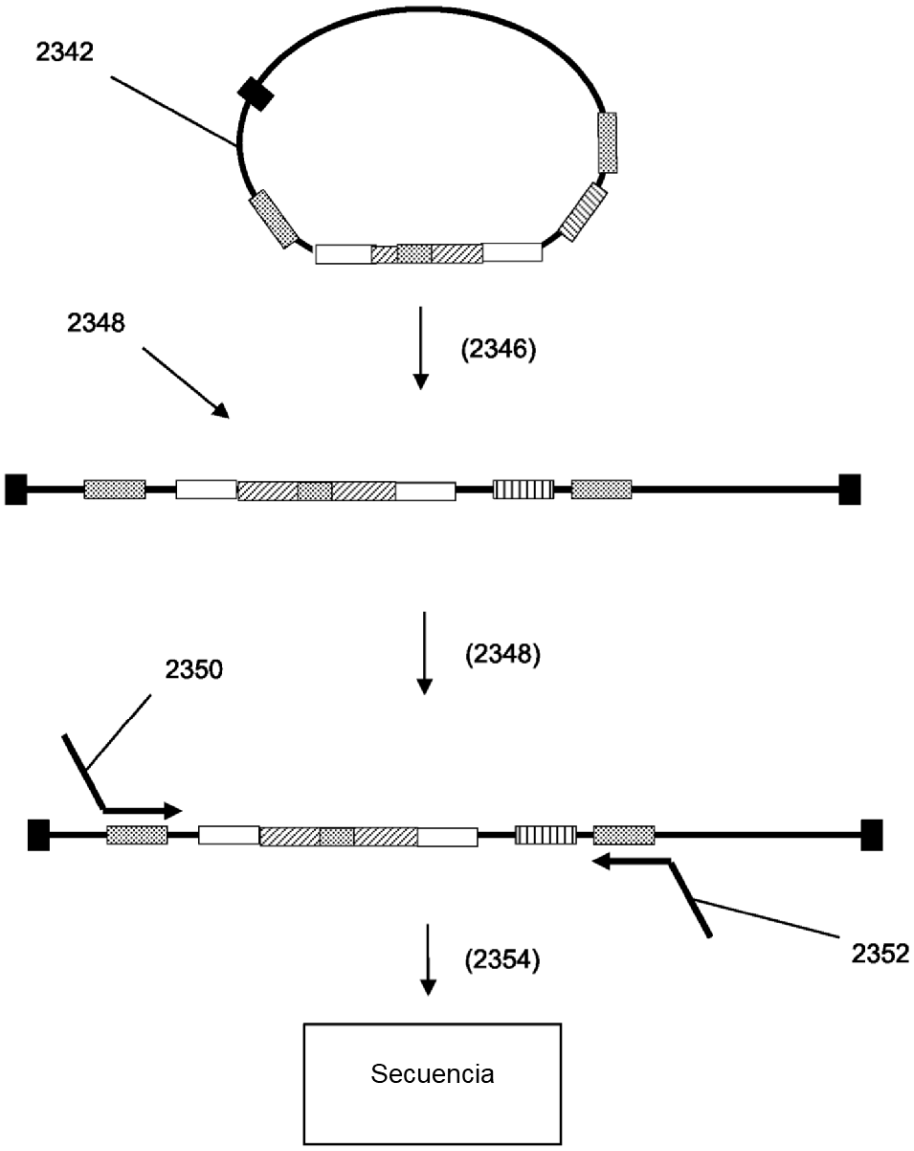


Fig. 2F

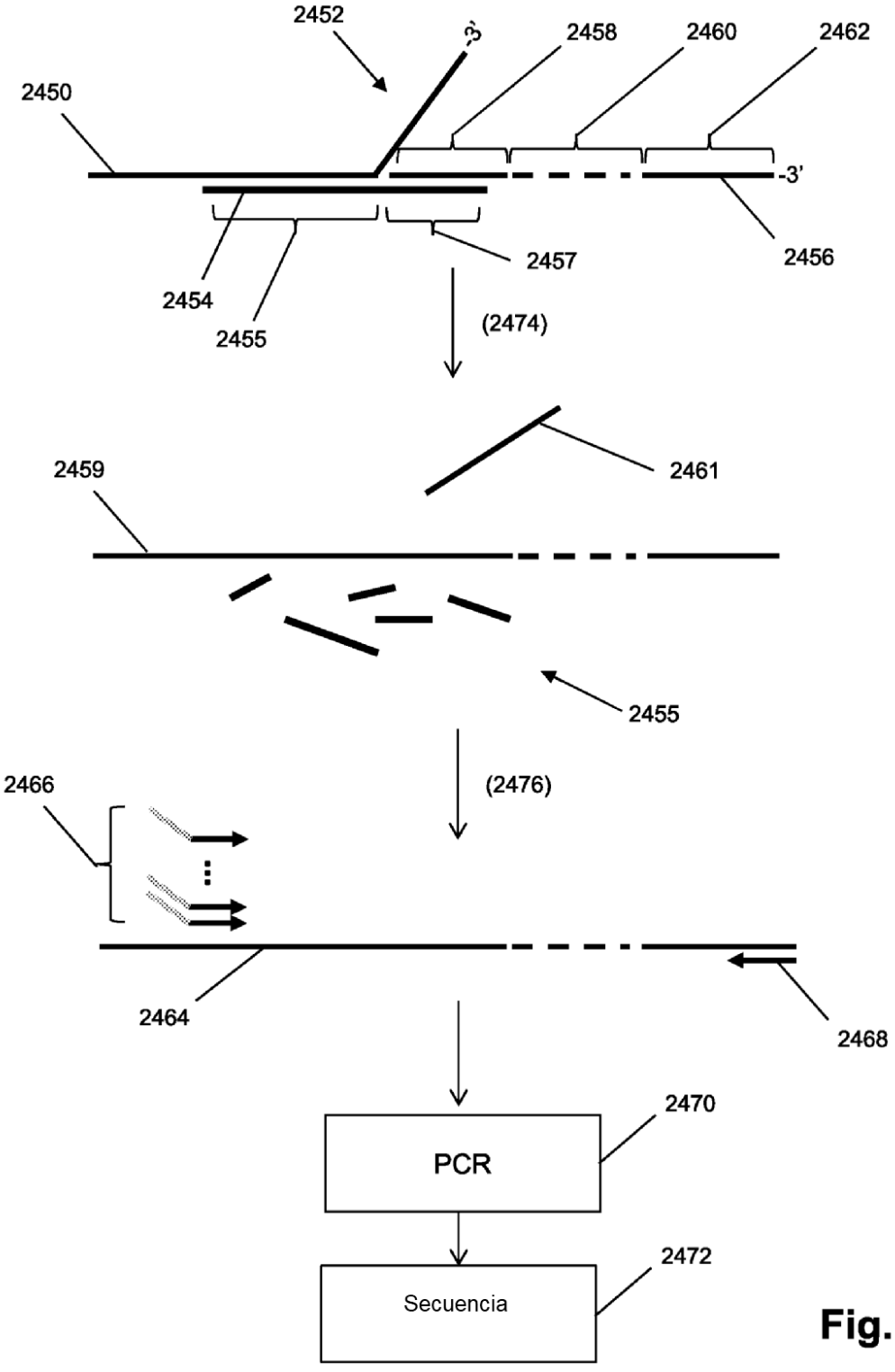

2452
3'
2458
2460
2462
2450
-3'
2457
2456
2454
2455
(2474)
2459
2461
2455
(2476)
2466
2464
2468
2470
PCR
2472
Secuencia

Fig. 2G

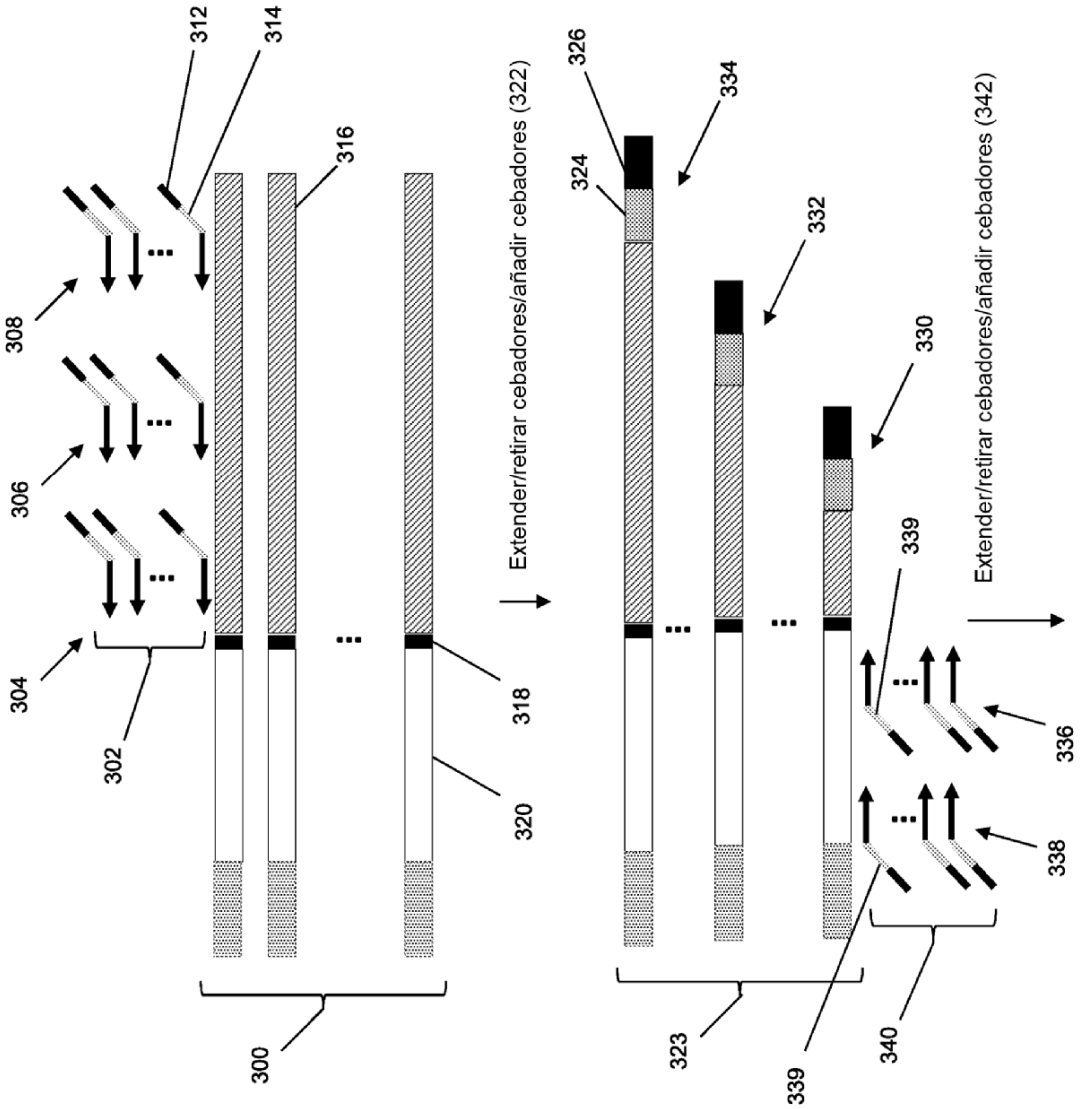
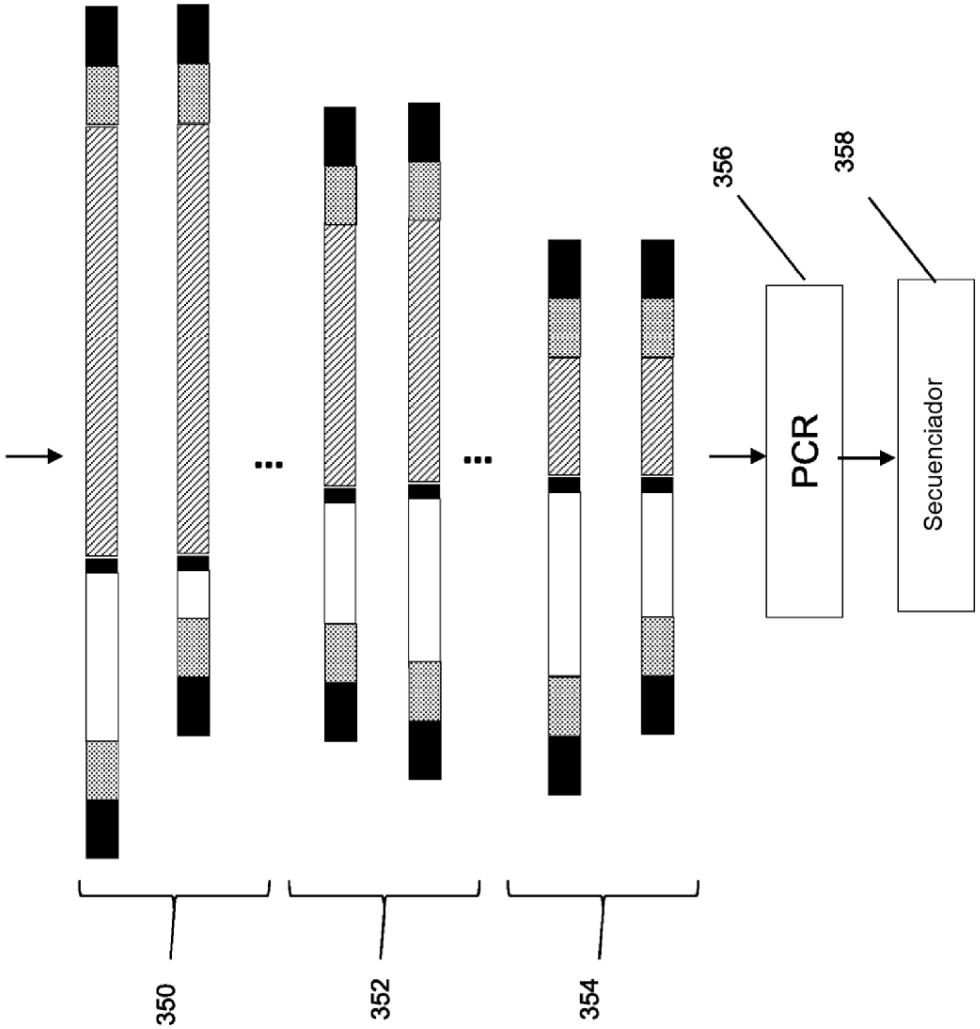

312
314
316
308
306
304
302
318
320
300
Extender/retirar cebadores/añadir cebadores (322)
326
334
324
332
330
339
Extender/retirar cebadores/añadir cebadores (342)
336
338
323
339
340

Fig. 3A

Fig. 3B



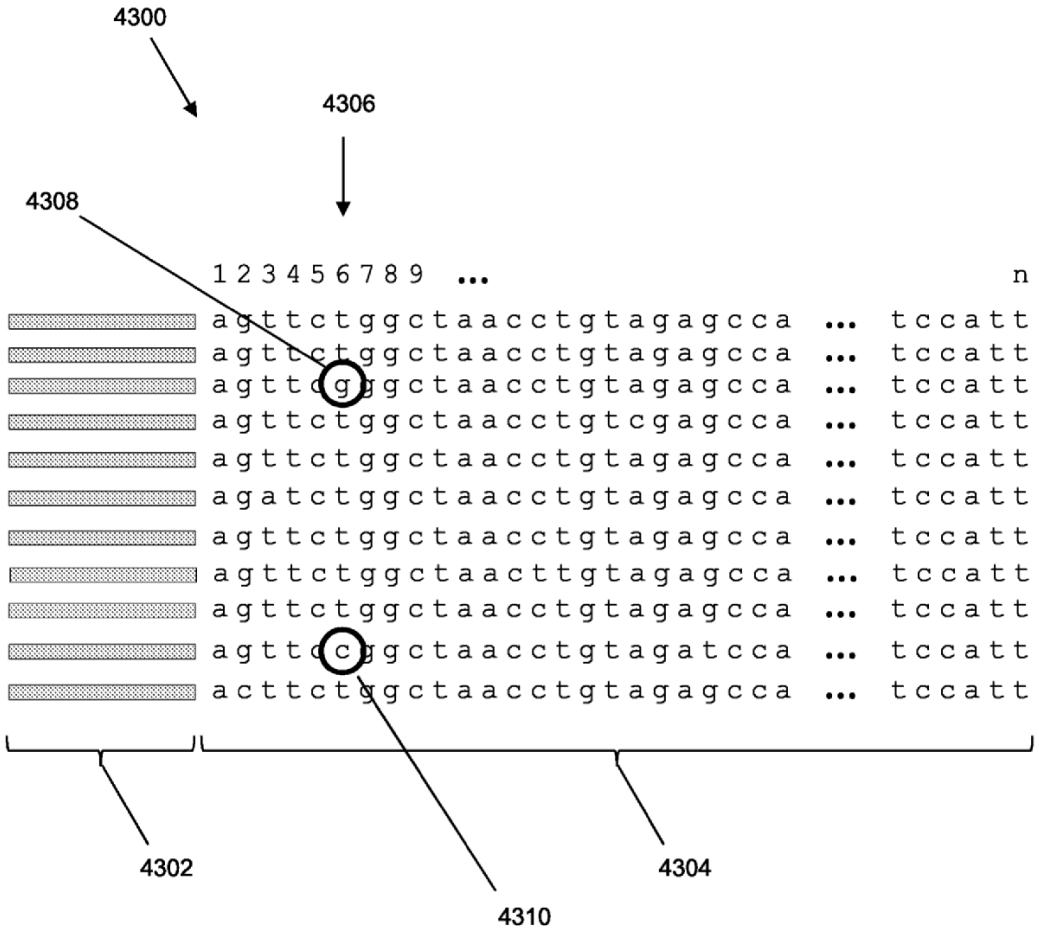

4300
4306
4308
1 2 3 4 5 6 7 8 9 ... n
agttctggctaacctgtagagcca ... tccatt
agttctggctaacctgtagagcca ... tccatt
agttctggctaacctgtagagcca ... tccatt
agttctggctaacctgtcgagcca ... tccatt
agttctggctaacctgtagagcca ... tccatt
agatctggctaacctgtagagcca ... tccatt
agttctggctaacctgtagagcca ... tccatt
agttctggctaacttgtagagcca ... tccatt
agttctggctaacctgtagagcca ... tccatt
agttctggctaacctgtagatcca ... tccatt
acttctggctaacctgtagagcca ... tccatt
4302
4304
4310

Fig. 4A

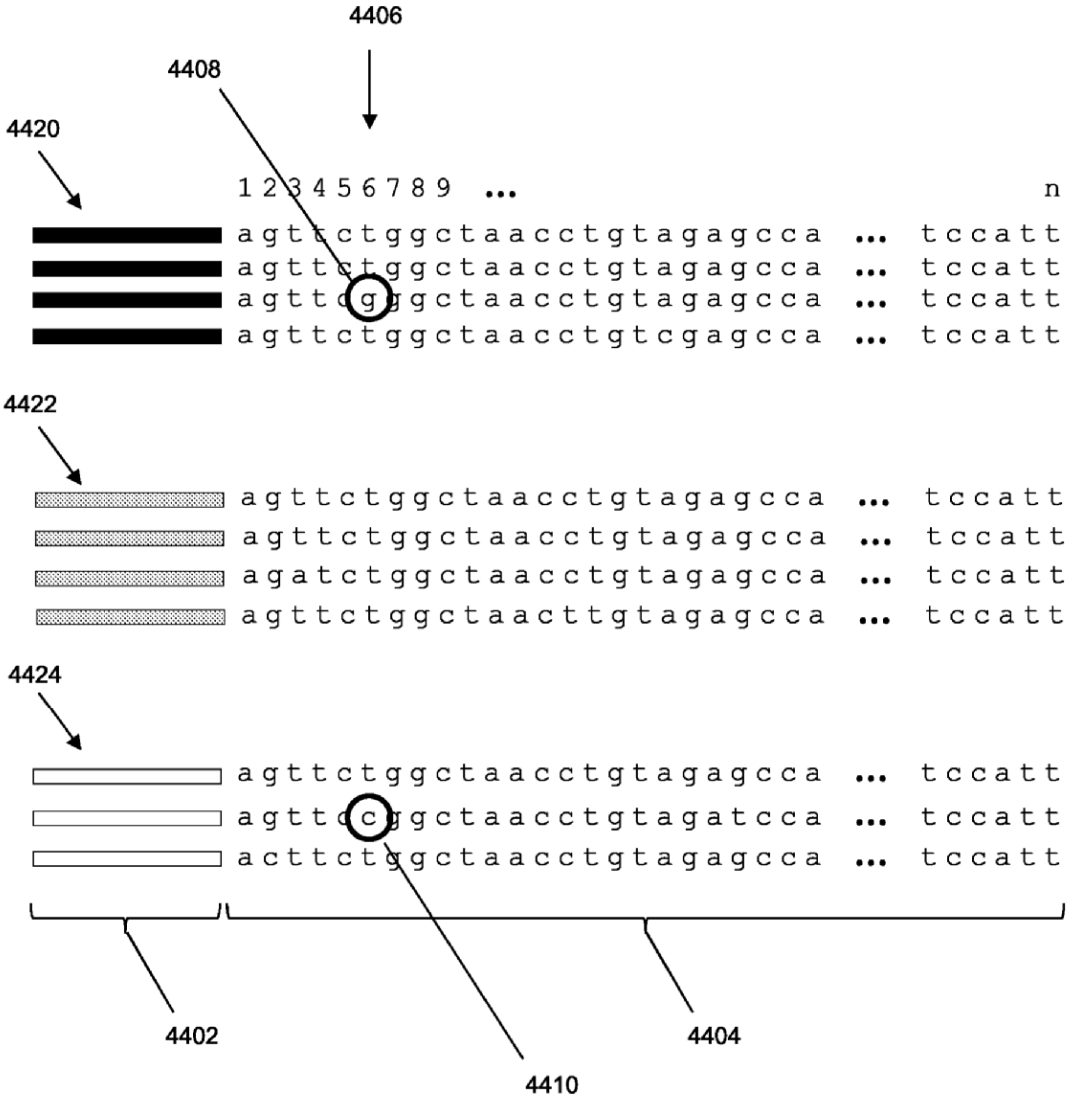

4406
4408
4420
1 2 3 4 5 6 7 8 9 ... n
agttctggctaacctgtagagcca ... tccatt
agttctggctaacctgtagagcca ... tccatt
agttcgggctaacctgtagagcca ... tccatt
agttctggctaacctgtcgagcca ... tccatt
4422
agttctggctaacctgtagagcca ... tccatt
agttctggctaacctgtagagcca ... tccatt
agatctggctaacctgtagagcca ... tccatt
agttctggctaacttgtagagcca ... tccatt
4424
agttctggctaacctgtagagcca ... tccatt
agttccggctaacctgtagatcca ... tccatt
acttctggctaacctgtagagcca ... tccatt
4402
4410
4404

Fig. 4B

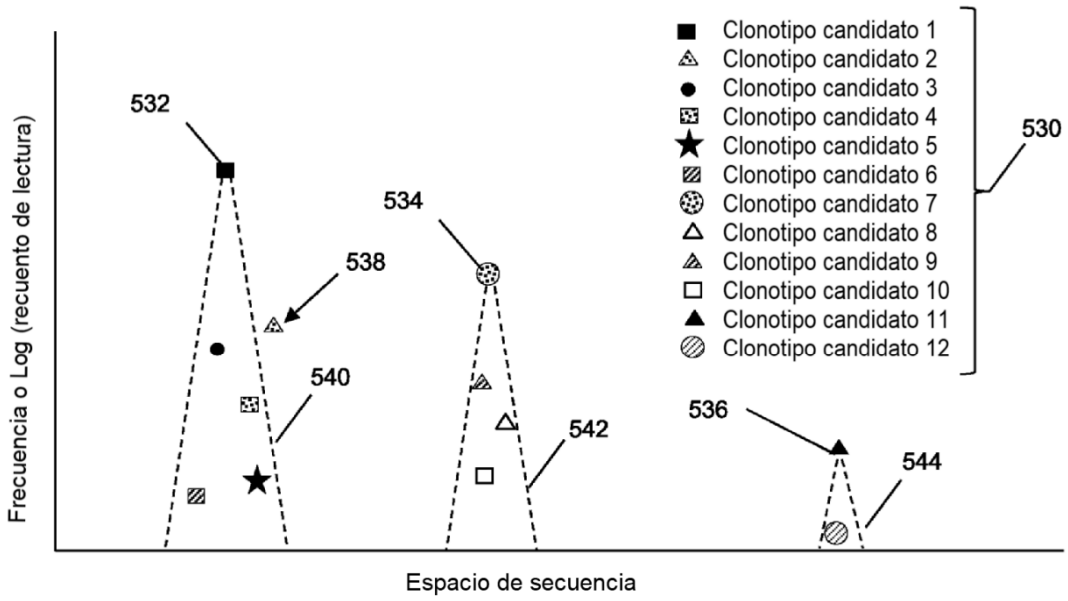

Clonotipo candidato 1
Clonotipo candidato 2
Clonotipo candidato 3
Clonotipo candidato 4
Clonotipo candidato 5
Clonotipo candidato 6
Clonotipo candidato 7
Clonotipo candidato 8
Clonotipo candidato 9
Clonotipo candidato 10
Clonotipo candidato 11
Clonotipo candidato 12
530
532
534
536
538
540
542
544
Frecuencia o Log (recuento de lectura)
Espacio de secuencia

Fig. 5A

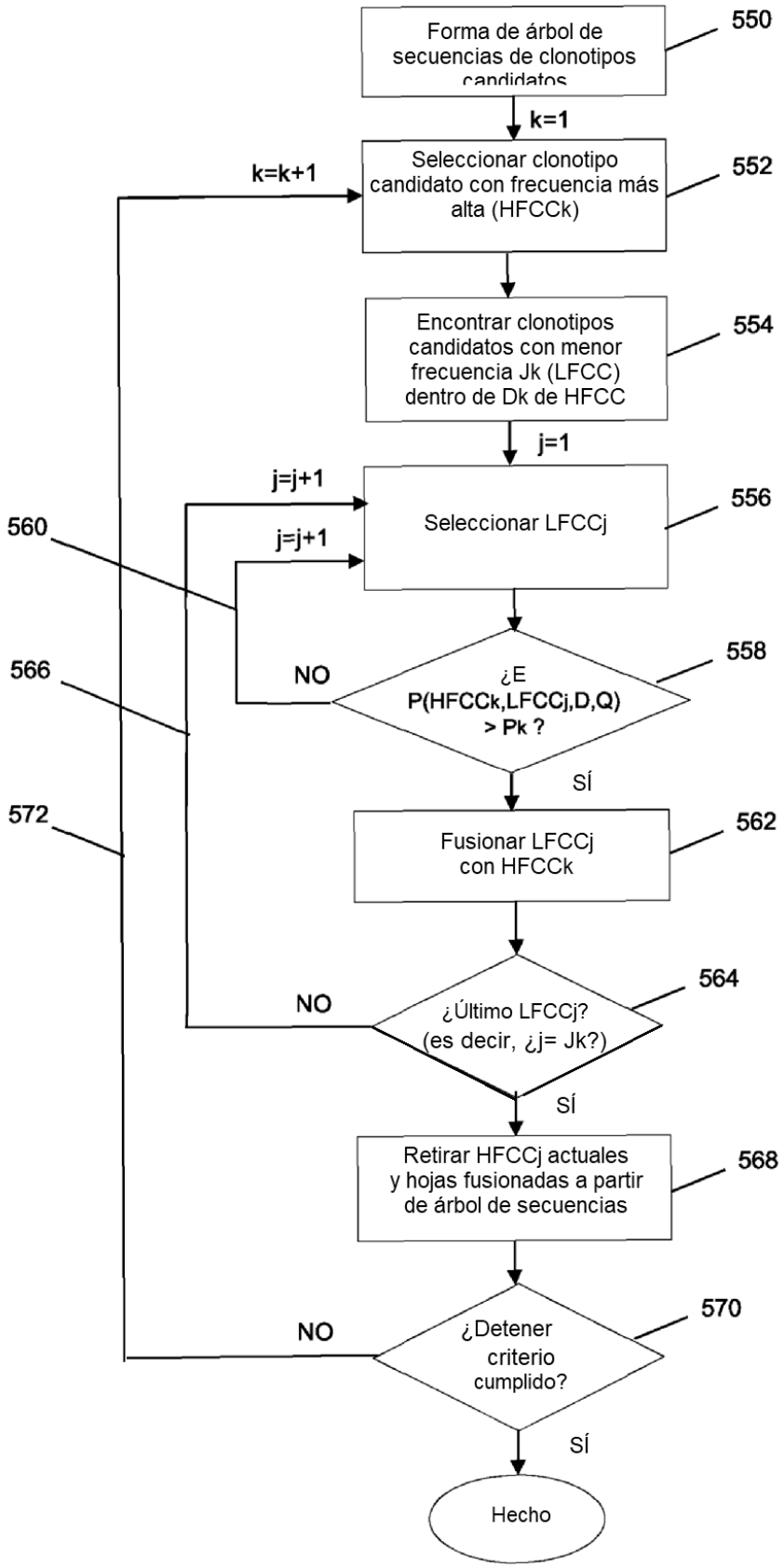

Forma de árbol de secuencias de clonotipos candidatos
550
k=1
Seleccionar clonotipo candidato con frecuencia más alta (HFCCk)
552
k=k+1
Encontrar clonotipos candidatos con menor frecuencia Jk (LFCC) dentro de Dk de HFCC
554
j=1
Seleccionar LFCCj
556
j=j+1
j=j+1
560
¿E P(HFCCk,LFCCj,D,Q) > Pk ?
558
NO
566
SÍ
Fusionar LFCCj con HFCCk
562
572
¿Último LFCCj? (es decir, ¿j= Jk?)
564
NO
SÍ
Retirar HFCCj actuales y hojas fusionadas a partir de árbol de secuencias
568
¿Detener criterio cumplido?
570
NO
SÍ
Hecho

Fig. 5B

Función ejemplar para determinar si fusionar clonotipos candidatos dependiendo de los recuentos de lectura, diferencias de base y puntuaciones Q

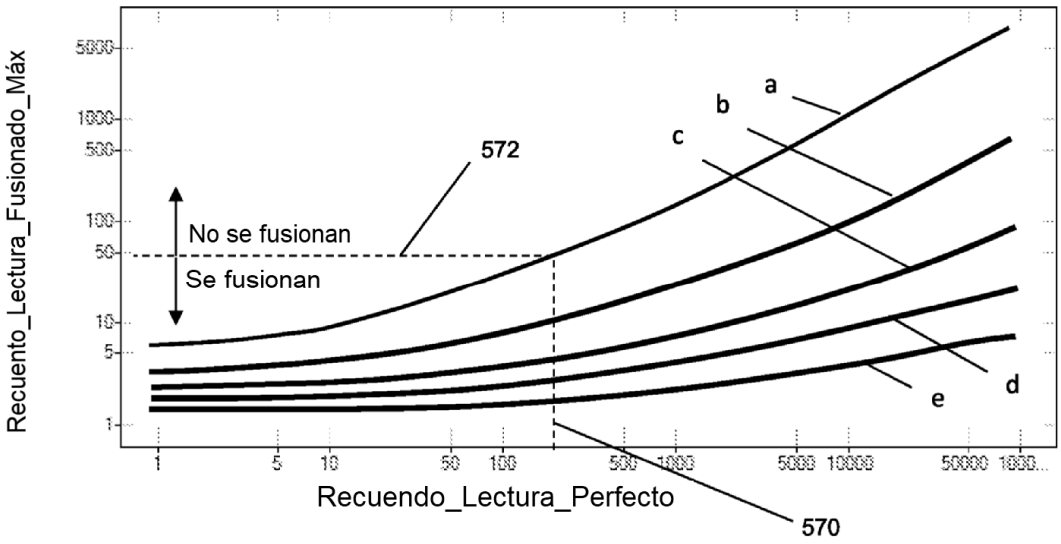


Fig. 5C

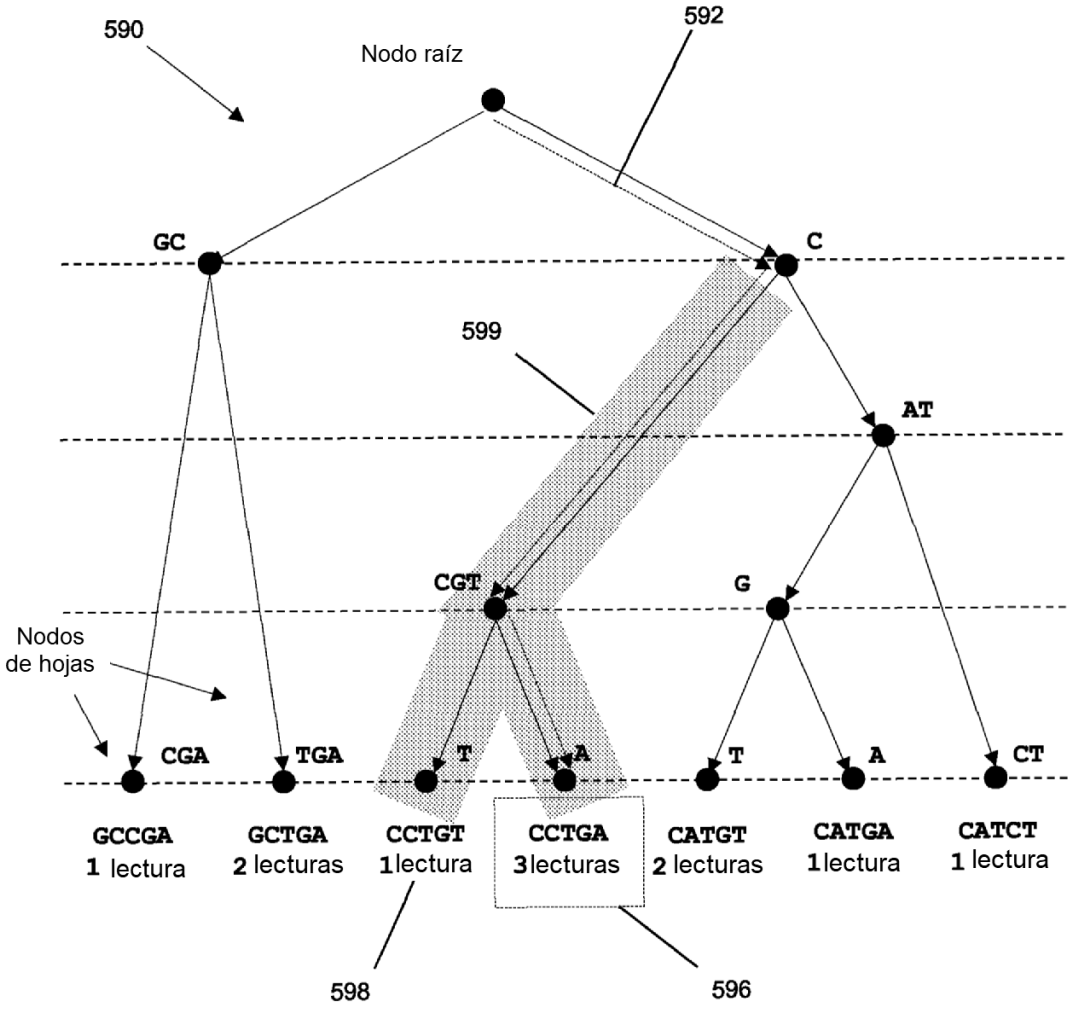

590
Nodo raíz
592
GC
C
599
AT
CGT
G
Nodos
de hojas
CGA
TGA
T
A
T
A
CT
GCCGA
1 lectura
GCTGA
2 lecturas
CCTGT
1lectura
CCTGA
3lecturas
CATGT
2 lecturas
CATGA
1lectura
CATCT
1 lectura
598
596

Fig. 5D

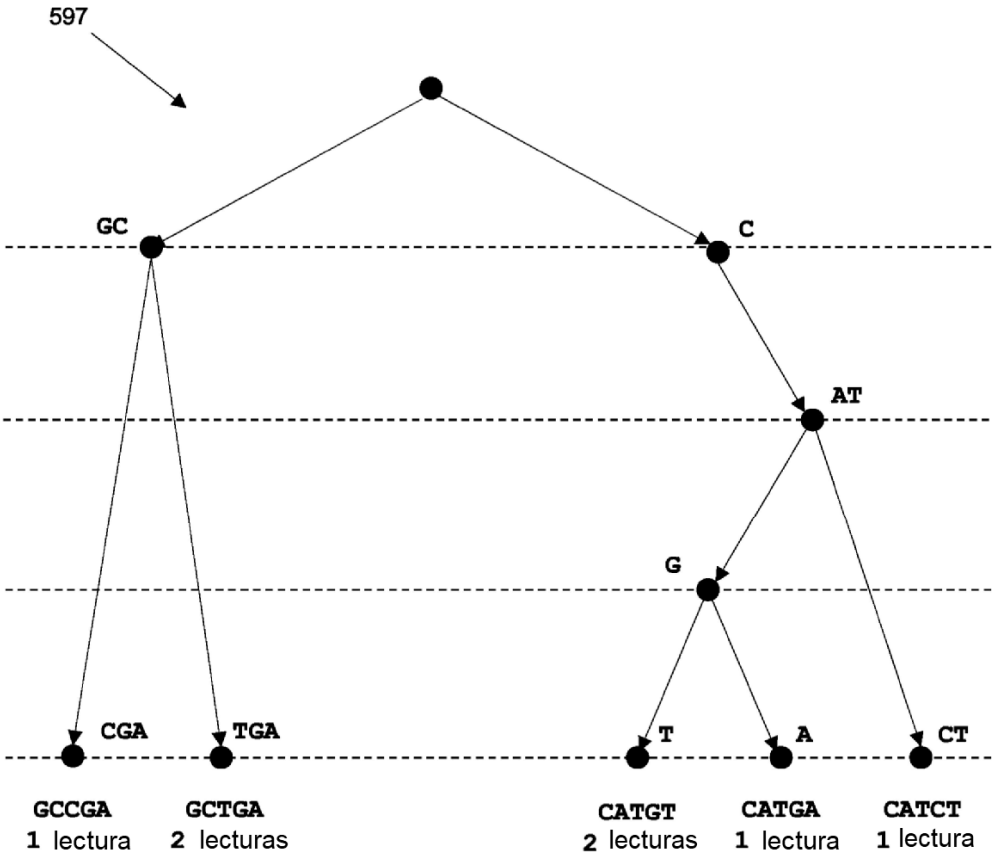

597
GC
C
AT
G
CGA
TGA
T
A
CT
GCCGA
1 lectura
GCTGA
2 lecturas
CATGT
2 lecturas
CATGA
1 lectura
CATCT
1 lectura

Fig. 5E