(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 662 128**

(51) Int. CI.:

C12Q 1/68 (2008.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(86) Fecha de presentación y número de la solicitud internacional: **04.03.2013 PCT/US2013/028942**

(87) Fecha y número de publicación internacional: **12.09.2013 WO13134162**

(96) Fecha de presentación y número de la solicitud europea: **04.03.2013 E 13757482 (8)**

(97) Fecha y número de publicación de la concesión europea: **14.02.2018 EP 2823060**

(54) Título: **Determinación de cadenas de receptor inmunitario emparejadas a partir de la frecuencia de subunidades coincidentes**

(30) Prioridad:

05.03.2012 US 201261606617 P

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
05.04.2018

(73) Titular/es:

ADAPTIVE BIOTECHNOLOGIES CORPORATION (100.0%)
400 East Jamie Court, Suite 301
South San Francisco, CA 94080, US

(72) Inventor/es:

FAHAM, MALEK y
KLINGER, MARK

(74) Agente/Representante:

UNGRÍA LÓPEZ, Javier

DESCRIPCIÓN

Determinación de cadenas de receptor inmunitario emparejadas a partir de la frecuencia de subunidades coincidentes

Antecedentes de la invención

Muchas funciones inmunitarias cruciales están mediadas por receptores de células T (TCR), que comprenden subunidades α y β que juntas se unen a un complejo que consiste en un péptido antigénico y moléculas del complejo mayor de histocompatibilidad (MHC). Se cree que varias enfermedades importantes se producen por una función aberrante de las células T: Por ejemplo, se cree que hay cánceres que surgen por un fallo en la vigilancia inmunitaria, es decir, la función de células T de detección y destrucción de clones de células transformadas antes de que se desarrollen en tumores; y se cree que las enfermedades autoinmunitarias surgen por una respuesta hiperactiva o aberrante de las células T contra los auto-antígenos, Abbas et al, Cellular and Molecular Immunology, Cuarta Edición (W.B. Saunders Company, 2000). En consecuencia, se tiene interés en aprovechar las funciones de las células T para varias estrategias terapéuticas de tratamiento tanto del cáncer como de las enfermedades autoinmunitarias, por ejemplo, Molloy et al, Current Opinion in Pharmacology, 5: 438-443 (2005); Morgan et al, Science, 314: 126-129 (2006); Turcotte y Rosenberg, Adv. Surg., 45: 341-360 (2011). Un desafío común en dichas estrategias es identificar y aislar las subunidades TCRα y TCRβ que juntas van a formar un receptor completo capaz de unirse específicamente a una diana de interés. Normalmente, se identifica una célula T y se expande clónicamente para hacer posible el aislamiento y análisis de los ácidos nucleicos que codifican cada subunidad. Sin embargo, a menos de que el TCR sea específico para un antígeno de una enfermedad común, tal como MART-1 en el melanoma, el procedimiento de análisis de una única célula, clonación y aislamiento del receptor se debe repetir para cada paciente. Se ha descrito una técnica para emparejar los genes variables de cadena ligera y variables de cadena pesada de anticuerpo basándose en sus frecuencias relativas (Reddy et al., 2010. Nature Biotechnology 28, 965-969). Se ha descrito una técnica para análisis emparejados de cadenas TCRα y TCRβ a nivel de una única célula en ratones (Dash et al., 2011. J Clin. Invest. 121(i), 288-295).

Recientemente, se han propuesto aplicaciones diagnósticas y pronósticas que utilizan la secuenciación de ADN a gran escala ya que el coste por base de la secuenciación del ADN ha bajado y las técnicas de secuenciación se han convertido en más convenientes, por ejemplo, Welch et al, Hematology Am. Soc. Hematol. Educ. Program, 2011: 30-35; Cronin et al, Biomark Med., 5: 293-305 (2011); Palomaki et al, Genetics in Medicine (publicación en línea del 2 de febrero 2012). En particular, los perfiles de los ácidos nucleicos que codifican moléculas inmunitarias, tales como los receptores de células T o células B, o sus componentes, contienen una enormidad de información del estado de salud o enfermedad de un organismo, de manera que los indicadores diagnósticos y pronósticos basados en el uso de dichos perfiles se han desarrollado para una amplia variedad de afecciones, Faham y Willis, Publicación de Patente de EE. UU. 2010/0151471 y documento WO 2010/053587 A2; Freeman et al, Genome Research, 19: 1817-1824 (2009); Boyd et al, Sci. Transl. Med., 1(12): 12ra23 (2009); He et al, Oncotarget (8 de marzo de 2011). Los perfiles actuales que se basan en secuencias de repertorios inmunitarios consisten en ácidos nucleicos que codifican solamente cadenas sencillas de receptor; por lo tanto, no está disponible la información potencialmente útil de cadenas TCRα y TCRβ emparejadas correctamente o cadenas pesada y ligera de inmunoglobulina.

En vista de lo anterior, sería muy útil para el tratamiento del cáncer y enfermedades autoinmunitarias que hubiera métodos convenientes disponibles para la determinación de los receptores inmunitarios funcionales a partir de ácidos nucleicos que codifiquen subunidades que se hayan extraído o secuenciado por separado.

Sumario de la invención

La invención proporciona un método de determinación de cadenas de receptor inmunitario emparejadas en una muestra, comprendiendo el método las etapas de:

(a) dividir una muestra que contiene linfocitos que expresan parejas de cadenas de receptor inmunitario en una pluralidad de subconjuntos;
(b) determinar las secuencias de nucleótidos de una primera cadena de cada pareja de cadenas de receptor inmunitario de linfocitos que tienen dichas parejas en una parte de la pluralidad de subconjuntos;
(c) determinar las secuencias de nucleótidos de una segunda cadena de cada pareja de cadenas de receptor inmunitario de linfocitos que tienen dichas parejas en la misma parte de la pluralidad de subconjuntos;
(d) identificar como cadenas de receptor inmunitario emparejadas las parejas de primera cadena y segunda cadena (i) que, para cada subconjunto de la parte, o se encuentran juntas o no se encuentran y (ii) que se encuentran juntas en al menos un subconjunto de la parte y no se encuentran en la menos un subconjunto de la parte.

La presente invención se refiere a métodos para la determinación de receptores inmunitarios funcionales, tales como receptores de células T o receptores de células B, a partir de subunidades seleccionadas de entre bibliotecas separadas. La invención se ejemplifica en varias implementaciones y aplicaciones, algunas de las cuales se resumen posteriormente y a lo largo de la memoria descriptiva.

En un aspecto, la invención se refiere a un método de determinación de un número predeterminado de cadenas de receptor inmunitario emparejadas en una muestra, que comprende las etapas de: (a) dividir una muestra que contiene linfocitos que expresan parejas de cadenas de receptor inmunitario en una pluralidad de subconjuntos; (b) determinar las secuencias de nucleótidos de una primera cadena de cada pareja de cadenas de receptor inmunitario de linfocitos que tienen dichas parejas en una parte de la pluralidad de subconjuntos; (c) determinar secuencias de nucleótidos de una segunda cadena de cada pareja de cadenas de receptor inmunitario de linfocitos que tienen dichas parejas en la misma parte de la pluralidad de subconjuntos; (d) identificar como cadenas de receptor inmunitario emparejadas las parejas de primera cadena y segunda cadena (i) que, para cada subconjunto de la parte, o se encuentran juntas o no se encuentran y (ii) que se encuentran juntas en al menos un subconjunto de la parte y no se encuentran en al menos un subconjunto de la parte; (e) repetir las etapas de (a)-(d) para otra pluralidad de subconjuntos diferente de cualquier pluralidad anterior hasta que se obtenga el número predeterminado de receptores inmunitarios emparejados.

Se describe en el presente documento un método de determinación de cadenas de receptor de células T emparejadas de las células T de una muestra que comprende las etapas de: (a) obtener una muestra que contiene células T, expresando cada célula T una primera cadena de receptor inmunitario y una segunda cadena de receptor inmunitario; (b) determinar las secuencias de nucleótidos de la primera cadena de receptor inmunitario de células T de la muestra, teniendo cada primera cadena de receptor inmunitario una frecuencia con que se encuentra en la muestra; (c) determinar las secuencias de nucleótidos de las segundas cadenas de receptor inmunitario de célula T de la muestra, teniendo cada segunda cadena de receptor inmunitario una frecuencia con que se encuentra en la muestra; y (d) identificar las primeras cadenas de receptor inmunitario y las segundas cadenas de receptor inmunitario emparejadas que tienen la misma frecuencia en la muestra.

Se describen en el presente documento los perfiles de clonotipos que se basan en las cadenas de receptor inmunitario emparejadas. Un método de generación de dichos perfiles comprende las etapas de (a) obtener una muestra que contenga células T y células B; (b) determinar las secuencias de nucleótidos de una primera cadena de receptor inmunitario de células T o células B de la muestra; (c) determinar las secuencias de nucleótidos de una segunda cadena de receptor inmunitario de células T o células B de la muestra; y (d) emparejar las secuencias de nucleótidos que codifican la primera y segunda cadenas de receptores inmunitarios que se expresan en la misma célula T y célula B para formar el perfil de cadenas de receptor inmunitario emparejadas.

Breve descripción de los dibujos

Las nuevas características de la invención se exponen con particularidad en las reivindicaciones adjuntas. Un mejor entendimiento de las características y ventajas de la presente invención se obtiene en referencia a la siguiente descripción detallada que expone realizaciones específicas, en las que se utilizan los principios de la invención, y los dibujos adjuntos en los que:

La FIG. 1A ilustra en forma de diagrama las etapas de un método de emparejamiento de cadenas TCRα y TCRβ a partir de moléculas secuenciadas por separado.

La FIG. 1B ilustra en forma de diagrama las etapas de una realización de la invención para la determinación de las cadenas TCRα y TCRβ que se originan de la misma célula T o la cadena pesada y ligera de inmunoglobulinas que se originan de la misma célula B.

La FIG. 2A-2C muestran un esquema de la PCR en dos etapas para la amplificación de genes de la TCRβ.

La FIG. 3A ilustra detalles de la determinación de una secuencia de nucleótidos del producto de la PCR de la Fig. 2C. La FIG. 3B ilustra los detalles de otro método de determinación de una secuencia de nucleótidos del producto de la PCR de la Fig. 2C.

La FIG. 4A ilustra un esquema de la PCR para generar tres matrices de secuenciación de una cadena IgH en una única reacción. Las FIG. 4B-4C ilustran esquemas de PCR para generar tres matrices de secuencia a partir de una cadena IgH en tres reacciones separadas después de lo cual se combinaron los amplicones resultantes para un PCR secundaria para añadir sitios de unión para los cebadores P5 y P7. La FIG. 4D ilustra las localizaciones de lecturas de secuencia generadas por una cadena IgH. La FIG. 4E ilustra el uso de la estructura del codón de las regiones V y J para mejorar la identificación de bases de la región NDN.

Descripción detallada de la invención

La práctica de la presente invención puede emplear, a menos de que se indique otra cosa, técnicas convencionales y descripciones de biología molecular (incluyendo técnicas recombinantes), bioinformática, biología celular, y bioquímica, que están en la experiencia de la técnica. Dichas técnicas convencionales incluyen, pero no se limitan a, muestreo y análisis de células sanguíneas, secuenciación y análisis de ácidos nucleicos, y similares. Se pueden tener ilustraciones específicas de técnicas adecuadas en referencia a los ejemplos posteriores en el presente documento. Sin embargo, también se pueden utilizar, por supuesto, otros procedimientos convencionales. Dichas

técnicas convencionales y descripciones se pueden encontrar en manuales de laboratorio convencionales tales como Genome Analysis: A Laboratory Manual Series (Vols. I-IV); PCR Primer: A Laboratory Manual; y Molecular Cloning: A Laboratory Manual (all from Cold Spring Harbor Laboratory Press); y similares.

La invención proporciona métodos para emparejar parejas de cadenas de receptor inmunitario a partir de poblaciones de sus ácidos nucleicos codificantes que se han secuenciado. De acuerdo con una realización de la invención, se secuencian las poblaciones de ácido nucleico que codifican repertorios de regiones variables de cadena pesada y regiones variables de cadena ligera de manera que se forman dos listas de secuencias sin ninguna correspondencia entre los miembros de cada lista. Esto se puede conseguir llevando a cabo operaciones, o ejecuciones, de secuenciación por separado de cada cadena, o se puede conseguir llevando a cabo una ejecución de secuencia única con los ácidos nucleicos marcados de acuerdo con la identidad del tipo de cadena que codifican. De acuerdo con otra realización de la invención, se secuencian las poblaciones de ácido nucleico que codifican repertorios de receptor alfa de células T (TCRα) y las cadenas de receptor beta de células T (TCRβ), de manera que se forman dos listas sin correspondencia entre los miembros de cada lista. De acuerdo con otra realización de la invención, se secuencian las poblaciones de ácido nucleico que codifican repertorios de cadenas de receptor gamma de células T (TCRγ) y cadenas de receptor delta de células T (TCRδ), de manera que se forman dos listas separadas sin ninguna correspondencia entre los miembros de cada lista. Como anteriormente, esto se puede conseguir llevando a cabo ejecuciones de secuenciación separadas para cada cadena, o se puede conseguir llevando a cabo la ejecución de una única secuencia con los ácidos nucleicos marcados de acuerdo con la identidad del tipo de cadena que codifican (es decir, sean TCRα y TCRβ, o TCRγ y TCRδ, respectivamente). En las últimas realizaciones, se pueden seguir dos estrategias para emparejar o hacer coincidir cadenas TCRα y TCRβ (o TCRγ y TCRδ) en cadenas que sean funcionales, por ejemplo, porque se originen de la misma célula T. En una primera estrategia, se determinan las secuencias de cada uno de los ácidos nucleicos codificantes y se emparejan las cadenas de TCRα y cadenas TCRβ cuyas secuencias de nucleótidos codificantes tengan las mismas frecuencias para formar un TCR funcional, o reconstituido. Las cadenas TCRγ y TCRδ se pueden emparejar por el mismo procedimiento. En una segunda estrategia, que es aplicable para emparejar los tres tipos de parejas de receptor inmunitario, se divide repetidamente una población de linfocitos en una pluralidad de subconjuntos. A partir de una porción o subpoblación, de los subconjuntos de ácidos nucleicos codificantes se extraen y secuencian las dos cadenas diferentes de receptor inmunitario, de manera que se forman dos listas separadas de secuencias sin ninguna correspondencia entre los miembros de cada lista. Como se ha descrito anteriormente, esto se puede conseguir llevando a cabo ejecuciones de secuenciación separadas de cada cadena, o puede conseguirse llevando a cabo una ejecución de cadena única con los ácidos nucleicos marcados de acuerdo con la identidad del tipo de cadena que codifican. Para ilustrarlo mediante un ejemplo, si se hacen alícuotas de una muestra que contiene células T o células B en 100 sub-muestras, de manera que de media cada alícuota contiene un subconjunto que contiene aproximadamente 1/100 del número total de células T o células B de la muestra original, entonces se pueden seleccionar aleatoriamente 20 de dichos subconjuntos como una parte del número total de subconjuntos. (Dicha parte podría ser un número mayor de uno y menor de 100, aunque como se describe más completamente posteriormente, un número en el intervalo de 10 a 20 es una buena marca entre la cantidad de secuenciación necesaria y la probabilidad de identificar parejas de receptor presentes con una frecuencia de interés). En una realización, una pluralidad de subconjuntos está en el intervalo de desde 20 a 2000 y una parte de subconjuntos de la misma está en el intervalo de entre 10 a 50. En otra realización, una parte de subconjuntos está en el intervalo de 10 a 20. Un ejemplo de la primera estrategia se ilustra en la Fig. 1A. Un ejemplo de la realización de la segunda estrategia se ilustra en la Fig. 1B.

Como se ilustra en la Fig. 1A, el ácido nucleico (que puede ser un ADN o ARN) se extrae de una muestra que contiene células T (100), después de lo cual en volúmenes de reacción separados, se combinan cebadores (102) específicos para los ácidos nucleicos que codifican la TCRα (o una parte de la misma) y los cebadores (104) específicos de los ácidos nucleicos que codifican la TCRβ (o una parte de la misma) en condiciones que permiten que se amplifiquen las respectivas poblaciones de ácidos nucleicos, por ejemplo, por una reacción en cadena de la polimerasa (PCR) en dos etapas, tal como se desvela por Faham y Willis (citado anteriormente). Las directrices y divulgaciones para seleccionar dichos cebadores y llevar a cabo dichas reacciones se describen extensamente en la bibliografía sobre inmunología molecular y posteriormente (para TCRβ e IgH) y en referencias tales como, Yao et al, Cellular and Molecular Immunology, 4: 215-220 (2007) (para TCRα). En un caso, los amplicones (106) y (108) producidos por una PCR de dos etapas están listos para el análisis de secuencia utilizando un secuenciador de última generación disponible en el mercado, tal como MiSeq Personal Sequencer (Illumina, San Diego, CA). Después de que se hayan determinado las secuencias de nucleótidos (107) y (109), se obtienen las bases de datos o tablas (110 y 112, respectivamente). Igualmente se pueden contar las secuencias y se construyen gráficos de frecuencia frente a secuencia (114 y 116). Se pueden determinar los TCR reconstituidos emparejando (118) TCRα y TCRβ con frecuencias idénticas o con frecuencias que tienen el mismo intervalo de ordenación. Claramente, este método funciona más eficazmente cuando las frecuencias de diferentes TCRα y TCRβ no están demasiado juntas, es decir son distintas, incluso teniendo en cuenta el error experimental.

Una vez que se identifica un par de secuencias de clonotipo que tienen iguales frecuencias (o se clasifican igualmente) se pueden reconstruir secuencias de longitud completa que codifican cada cadena a partir de regiones constantes y variables utilizando técnicas convencionales de modificación y expresión genética, por ejemplo, Walchli et al, PLosOne, 6(11): e27930 (2011); o similares.

Se puede obtener una mayor precisión en la determinación de las frecuencias de cadenas de receptor con una variación del método anterior, que se puede ver en referencia a las Fig. 2A y 2B en las que el ARN que codifica TCRβ se amplifica en una PCR de dos etapas. Como se describe más completamente posteriormente, el cebador (202) y el conjunto de cebadores (212) se utilizan en una primera etapa de amplificación para adjuntar un sitio de unión al cebador (214) común a todos los ácidos nucleicos que codifiquen la TCRβ. La Fig. 2B ilustra los componentes de una segunda etapa de amplificación para generar más material y para adjuntar sitios de unión del cebador P5 (222) y P7 (220) que se utilizan en la formación de agrupamientos (mediante una PCR puente) en el protocolo de secuenciación basado en Solexa. El cebador P7 (220) también puede incluir un marcador de muestra (221) para multiplexar hasta 96 muestras para secuenciación concurrente en la misma ejecución, por ejemplo, la nota de la aplicación Illumina 770-2008-011 (2008). Se puede utilizar un tipo diferente de marcador en el mismo cebador para aumentar la precisión de la determinación de las frecuencias de la cadena del receptor. En este método, el cebador P7 se modifica para que incluya un conjunto de marcadores, de manera que, en vez de 96 marcadores, el cebador P7 se modifica para que tenga 10.000 marcadores distintos, o más. En otras palabras, el cebador P7 es una mezcla de 10.000 o más oligonucleótidos distintos que tiene cada uno una región de unión idéntica en la matriz, una secuencia de marcador distinta, y una parte de la cola 5' idéntica (por ejemplo, (223) en la Fig. 2B). Con esta disposición, cualquier subconjunto de ácidos nucleicos que codifique la misma cadena de receptor (por ejemplo, menos de 100) recibirá con alta probabilidad un marcador diferente. Dicho procedimiento de emparejamiento de miembros de un pequeño conjunto de ácidos nucleicos con un conjunto de marcadores mucho más grande con fines de recuento, marcado, clasificación se conoce bien y se desvela de distintas formas en las siguientes referencias, Brenner, Patente de EE. UU. 6,172,214; Brenner et al, Patente de EE. UU. 7,537,897; y Macevicz, Publicación de patente internacional WO US2005/111242; Brenner et al, Proc. Natl. Acad. Sci., 97: 1665-1670 (2000); Casbon et al, Nucleic Acids Research, 39(12): e81 (2011); Fu et al, Proc. Natl. Acad. Sci., 108: 9026-9031 (2011). La construcción de conjuntos de oligonucleótidos marcadores mínimamente hibridados, o marcadores con otras propiedades útiles, se desvela en las siguientes referencias ejemplares: Brenner, Patente de EE. UU. 6.172.214; Morris et al, Publicación de Patente de EE. UU. 2004/0146901; Mao et al, Publicación de Patente de EE. UU. 2005/0260570; y similares. Preferentemente, el conjunto de marcadores debería ser al menos 100 veces (o más) el tamaño del conjunto de ácidos nucleicos que se van a marcar si todos los ácidos nucleicos van a recibir un único marcador con alta probabilidad. Para las cadenas de receptor inmunitario, en un caso, el número de marcadores distintos está en el intervalo de desde 10.000 a 100.000; en otro caso, el número de distintos marcadores está en el intervalo de desde 10.000 a 50.000; y en otro caso, el número de marcadores distintos está en el intervalo de desde 10.000 a 50.000; y en otro caso, el número de marcadores distintos está en el intervalo de 10.000 a 20.000. Como se desvela en Brenner, Patente de EE. UU. 6.172.214, dichas grandes mezclas de oligonucleótidos marcadores se pueden sintetizar por métodos combinatorios; de manera alternativa, los cebadores que contienen marcadores únicos se pueden sintetizar individualmente por métodos no combinatorios, tales como se desvela por Cleary et al, Nature Methods, 1: 241-248 (2004); York et al, Nucleic Acids Research, 40(1): e4 (2012); LeProust et al, Nucleic Acids Research, 38(8): 2522-2540 (2010); y similares.

El método anterior se puede llevar a cabo por las siguientes etapas: (a) obtener una muestra que contenga células T; (b) determinar las secuencias de nucleótidos de las cadenas TCRα de células T de la muestra, teniendo cada cadena TCRα una frecuencia que se encuentra en la muestra; (c) determinar las secuencias de nucleótidos de cadenas TCRβ de las células T de la muestra, teniendo cada cadena TCRβ una frecuencia que se encuentra en la muestra; y (d) identificar las cadenas TCRα y cadenas TCRβ como las que tienen la misma frecuencia en la muestra. Las frecuencias de las respectivas cadenas TCRα y cadena TCRβ se pueden determinar a partir de las tabulaciones de ácidos nucleicos codificantes, o clonotipos. De manera alternativa, las frecuencias de las respectivas cadenas TCRα y cadenas TCRβ se puede determinar a partir de las tabulaciones de polipéptidos codificados por los clonotipos. Como se ha mencionado anteriormente se pueden determinar las frecuencias de clonotipos haciendo el recuento de los clonotipos directamente o indirectamente utilizando un esquema de marcación como se ha descrito anteriormente.

La Fig. 1B ilustra una realización para la identificación de la coincidencia de subunidades de receptor que se puede aplicar a los TCR y BCR y que se pueden utilizar incluso cuando las frecuencias entre las cadenas de las subunidades del receptor están próximas o son indistinguibles, sea debido al error experimental u otros. Comenzando con una muestra que contienen linfocitos (149), que pueden ser células T o células B, se forman subconjuntos separando o dividiendo la muestra en una pluralidad de subconjuntos (152), 1 a K (en la figura). En algunas realizaciones, solamente se analiza una parte del subconjunto K, por lo tanto, no es necesario formar actualmente todos los subconjuntos K. Se pueden formar subconjuntos de solo la parte que se analiza actualmente. Por ejemplo, si la muestra tiene un volumen de 100 µl y K = 100, pero solo se va a analizar una parte que consiste en 20 subconjuntos, entonces solamente se necesitan formar veinte subconjuntos de 1 µl. Para cada subconjunto (152) se secuencian los ácidos nucleicos que codifican cada cadena diferente de receptor inmunitario (mostrándose la TCRα y TCRβ en el subconjunto 1), formando de esta manera pares de listas, por ejemplo, (162), (164), (166) y (168) para los subconjuntos 1, 2… K-1, K, respectivamente. Cada par de dichas listas contienen una primera lista de secuencias de nucleótidos de una primera cadena de receptor inmunitario, por ejemplo, la lista (154) para TCRα del subconjunto 1, y una segunda lista de secuencias de nucleótidos de una segunda cadena de receptor inmunitario, por ejemplo, la lista (156) para TCRβ del subconjunto 1. En una realización, el número de subconjunto, K, es un número en el intervalo de desde 5 a 500; en otra realización, K es un número en el intervalo de desde 10 a 100; en otra realización, K es un número en el intervalo de desde 20 a 50. En algunas realizaciones una porción de

subconjuntos que se analiza es de 10 o menos subconjuntos; en otras realizaciones, una parte de subconjuntos que se analiza es de 20 o menos subconjuntos; en otras realizaciones, una parte de subconjuntos que se analiza es al menos del cinco por ciento de los subconjuntos; en otras realizaciones, una parte de los subconjuntos que se analiza es al menos del diez por ciento de los subconjuntos; en otras realizaciones, una parte de subconjuntos que se analiza es al menos del veinte por ciento de los subconjuntos.

Cada tipo de linfocitos de la muestra, por ejemplo, linfocitos (150), está presente en la muestra con una frecuencia particular. La distribución de linfocitos en los subconjuntos se aproxima fácilmente por un modelo binómico; por lo tanto, para un linfocito arbitrario (por ejemplo, (150)) que tiene un clonotipo particular, (a) su frecuencia en la muestra, (b) el número total de linfocitos en la muestra, y (c) el número de subconjuntos se pueden relacionar con la expectativa de encontrar al menos uno de los linfocitos particulares en una fracción predeterminada de subconjuntos. Esta relación se puede expresar de la siguiente manera: $r = (1-f)^{(N/K)}$, donde r es la fracción de subconjuntos que contienen al menos uno del linfocito particular, f es la frecuencia del linfocito particular en la muestra, N es el número total de linfocitos en la muestra, y K es el número de subconjuntos. Por lo tanto, si se establece que r = 1/2 y se toma K como una constante, entonces se pueden seleccionar valores sucesivos de K de manera que los linfocitos de diferentes frecuencias estén presentes en aproximadamente la mitad de los subconjuntos. Se podrían seleccionar otros valores de r, pero r = 1/2 proporciona los resultados con la mayor potencia estadística, por lo tanto, se prefiere una r ~ 1/2. Una vez que se obtienen dichas listas se examinan para identificar parejas de la primera y segunda secuencias de nucleótidos que existen juntas en un subconjunto o no. A modo de ejemplo, los miembros de la pareja (158) aparecen en las listas (164) del subconjunto 2 y en las listas (166) del subconjunto K-1, pero ninguno de los miembros de la pareja aparece en las listas (162) o (168) del subconjunto 2 y K-1, pero están ausentes en los subconjuntos 1 y K, tal como un linfocito (150). Dicho patrón confirma que los miembros de la pareja (158) van juntas y corresponden a las cadenas de un receptor inmunitario funcional. Pueden estar presentes otros linfocitos en la muestra (149) con aproximadamente la misma frecuencia, tal como el linfocito (153). Sin embargo, la probabilidad de que al menos uno de los linfocitos (153) se encuentre en exactamente los mismos subconjuntos que el linfocito (150) es extremadamente baja, especialmente si r es aproximadamente la mitad y la parte de los subconjuntos K analizados están en el intervalo de desde 10 a 20, o más.

En un aspecto de la invención, se puede determinar la primera y segunda cadena emparejadas de linfocitos de entre una sucesión de clases de frecuencia llevando a cabo el procedimiento anterior repetidamente para diferentes valores de K. Por ejemplo, una muestra de 1 ml de sangre periférica de un individuo normal contiene aproximadamente de 1-4,8 x 10^6 linfocitos de los que un 10-15 por ciento son células B, aproximadamente un 70-85 por ciento son células T y aproximadamente un 10 por ciento son células NK; por lo tanto, la muestra de 1 ml puede contener desde aproximadamente 7 x 10^5 a aproximadamente 4 x 10^6 células T. Si el número de linfocitos T en una muestra de 1 ml es N = 10^6, entonces el emparejamiento de cadenas de TCR de células T de las siguientes frecuencias se empareja identificando las que aparecen juntas en un cincuenta por ciento de los subconjuntos y en ninguno del otro cincuenta por ciento de los subconjuntos:

Frecuencia	Número de Subconjuntos	Volumen (µl)
0,001	1443	0,7
0,0005	722	1,4
0,0001	144	6,9
0,00005	72	13,9

Como se ha mencionado anteriormente, no es necesario analizar todos los subconjuntos con una frecuencia particular. Si hay un gran número de linfocitos que tienen frecuencias con o muy cerca de una frecuencia seleccionada, por ejemplo, f = 0,001, se pueden resolver todas tomando una porción cada vez mayor del número total de subconjuntos hasta que cada pareja aparece unida en el cincuenta por ciento de los subconjuntos se pueda distinguir de cada una de las otras parejas con la misma frecuencia. Esto es debido a que la probabilidad de que existan dos linfocitos diferentes en los mismos subconjuntos del cincuenta por ciento es infinitesimal según aumenta la porción de subconjuntos.

Usos de los TCR reconstituidos

Los receptores de células T reconstituidos tienen una variedad de usos tanto individualmente como en grupo, incluyendo, pero sin limitarse a, como compuestos de unión para inmunoterapia, como componentes de células T transfectadas para inmunoterapia adoptiva, como fuentes de antígenos para vacunas, y en indicadores del estado inmunitario. Las cadenas de TCR emparejadas en formato soluble se pueden utilizar como compuestos de unión de alta afinidad unidos a agentes de captura de células T para agentes terapéuticos anti-cáncer únicos, por ejemplo, como enseña Jakobsen et al, Patentes de EE. UU. 7.329.731 y 7.666.604. Las cadenas de TCR emparejadas pueden utilizarse para construir vectores que pueden, a su vez, utilizarse para transferir células T autólogas para inmunoterapia adoptiva de un paciente. En una realización de la presente solicitud, los TCR se analizan a partir de

las muestras tomadas antes y después de que el paciente se haya inmunizado con un antígeno canceroso, de manera que las cadenas de TCR elevadas se emparejan y se seleccionan fácilmente. Las referencias que desvelan dichas aplicaciones incluyen Turcotte et al, Adv. Surg., 45: 341-360 (2011); Morgan et al, Science, 314: 126-129 (2006); Walchli et al, PlosOne, 6: e27930 (2011); Robbins et al, Publicación de Patente de EE. UU. 2010/0034834;

Una población de TCR emparejados o reconstituidos de una muestra comprenden un único perfil del sistema inmunitario de un individuo, que contiene mucha más información que los perfiles de clonotipos de secuencia única. Es decir, una población de cadenas de TCR emparejadas o cadenas pesadas y ligeras de inmunoglobulinas emparejadas comprende un perfil de clonotipos en el que los clonotipos son parejas de secuencias de nucleótidos que codifican parejas de cadenas de TCR expresadas en la misma célula T o parejas de cadenas pesadas y ligeras de inmunoglobulina expresadas en la misma célula B. En ambos casos, dichas parejas se pueden relacionar directamente con la función de la célula T, por ejemplo, por interacción con conjuntos de complejos de péptidos tetraédricos del MHC, por ejemplo, Palmowski et al, Immunol. Rev., 188: 155-163 (2002); Hadrup et al, Nature Methods, 6: 520-526 (2009), o la función de la célula B, por ejemplo, por ELISA, por ejemplo, Reddy et al, Nature Biotechnology, 28(9): 965-969 (2010). En una realización, los perfiles de clonotipos de cadenas de receptor inmunitario emparejadas comprenden al menos 100 parejas de clonotipos, en el que cada clonotipo de la pareja comprende una secuencia de desde 30 a 300 nucleótidos. En otras realizaciones, los perfiles de clonotipos de las cadenas de receptor inmunitario emparejadas comprenden al menos 500 pares de clonotipos, en el que cada clonotipo de la pareja comprende una secuencia de desde 30 a 300 nucleótidos. En otra realización, los perfiles de clonotipos de las cadenas de receptor inmunitario emparejadas comprenden al menos 1000 pares de clonotipos, en los que cada clonotipo de la pareja comprende una secuencia de desde 30 a 300 nucleótidos. En otra realización más, dichos perfiles de clonotipos de cadenas de receptor inmunitario emparejadas comprenden pares de clonotipos de TCRα y TCRβ. En otra realización, dichos perfiles de clonotipos de cadenas de receptor inmunitario emparejadas comprenden clonotipos de TCRγ y TCRδ.

Muestras

Los perfiles de clonotipos se pueden obtener a partir de muestras de células inmunitarias. Por ejemplo, las células inmunitarias pueden incluir células T y/o células B. Las células T (linfocitos T) incluyen, por ejemplo, células que expresan receptores de célula T. Las células T incluyen las células T auxiliares (células T efectoras o células Th), células T citotóxicas (CTL), células T de memoria, y células T reguladoras. En un aspecto, una muestra de células T incluye al menos 1.000 células T; pero más normalmente, una muestra incluye al menos 10.000 células T y más normalmente, al menos 100.000 células T. En otro aspecto, una muestra incluye una cantidad de células T en el intervalo de desde 1000 a 1.000.000 células. Una muestra de células inmunitarias también puede comprender células B. Las células B incluyen, por ejemplo, células B plasmáticas, células B de memoria, células B1, células B2, células B de la zona marginal, y células B foliculares. Las células B pueden expresar inmunoglobulinas (anticuerpos, receptores de células B). Como anteriormente, en un aspecto una muestra de células B incluye al menos 1.000 células B; pero más normalmente, una muestra incluye al menos 10.000 células B, y más normalmente, al menos 100.000 células B. En otro aspecto, una muestra incluye una cantidad de células B en el intervalo de desde 1000 a 1.000.000 células B

Las muestras que se utilizan en los métodos de la invención pueden provenir de una variedad de tejidos, incluyendo, por ejemplo, un tejido tumoral, sangre y plasma sanguíneo, líquido linfático, líquido cefalorraquídeo que rodea el cerebro y la médula espinal, líquido sinovial que rodea las articulaciones entre los huesos, y similares. En una realización, la muestra es una muestra de sangre. La muestra de sangre puede tener 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 ml. La muestra puede ser una biopsia tumoral. La biopsia puede ser, por ejemplo, de un tumor del cerebro, hígado, pulmón, corazón, colon, riñón o medula ósea. Se puede utilizar cualquier técnica de biopsia utilizada por los expertos en la técnica para aislar una muestra de un sujeto. Por ejemplo, una biopsia puede ser una biopsia abierta, en la que se utiliza anestesia general. La biopsia puede ser una biopsia cerrada, en la cual se hace un corte más pequeño que en una biopsia abierta. La biopsia puede ser una biopsia nuclear o una biopsia incisional, en la cual se retira una parte del tejido. La biopsia puede ser una biopsia excisional, en la que se hace un intento de retirar la lesión completa. La biopsia puede ser una biopsia por aspiración de aguja fina, en la cual la muestra de tejido o de fluido se retira con una aguja.

La muestra puede ser una biopsia, por ejemplo, una biopsia de piel. La biopsia puede ser, por ejemplo, de cerebro, hígado, pulmón, corazón, colon, riñón o medula ósea. Se puede utilizar cualquier técnica de biopsia utilizada por los expertos en la técnica para aislar una muestra de un sujeto. Por ejemplo, una biopsia puede ser una biopsia abierta, en la que se utiliza anestesia general. La biopsia puede ser una biopsia cerrada, en la cual se hace un corte más pequeño que en una biopsia abierta. La biopsia puede ser una biopsia nuclear o una biopsia incisional, en la cual se retira una parte del tejido. La biopsia puede ser una biopsia excisional, en la que se hace un intento de retirar la lesión completa. La biopsia puede ser una biopsia por aspiración de aguja fina, en la cual la muestra de tejido o de fluido se retira con una aguja.

La muestra se puede obtener a partir de material corporal que es desechado por el sujeto. Dicho material desechado puede incluir heces humanas. El material desechado también puede incluir células de la piel desprendidas, sangre, dientes o pelo.

La muestra puede incluir un ácido nucleico, por ejemplo, ADN (por ejemplo, ADN genómico) o ARN (por ejemplo, ARN mensajero). El ácido nucleico puede ser un ADN o ARN libre de células, por ejemplo, extraído del sistema circulatorio, Vlassov et al, Curr. Mol. Med., 10: 142-165 (2010); Swarup et al, FEBS Lett., 581: 795-799 (2007). En los métodos de la invención proporcionada, la cantidad de ARN o ADN de un sujeto que se puede analizar incluye, por ejemplo, tan poco como una única célula en algunas aplicaciones (por ejemplo, en un ensayo de calibración) y como mucho 10 millones de células o más traducido a un intervalo de ADN de 6 pg – 60 ug, y ARN de aproximadamente 1 pg – 10 ug.

Como se expone más completamente posteriormente (Definiciones), una muestra de linfocitos será suficientemente grande como para que esté representada sustancialmente cada célula T o célula B con un clonotipo distinto en la misma, formando de esta manera un repertorio (como el término se utiliza en el presente documento). En una realización, una muestra contiene con una probabilidad del noventa y nueve por ciento cada clonotipo de una población presente con una frecuencia de 0,001 o mayor. En otra realización, una muestra contiene con una probabilidad del noventa y nueve por ciento cada clonotipo de una población presente con una frecuencia de un 0,0001 por ciento o mayor. En una realización, una muestra de células B o células T incluye al menos medio millón de células, y en otra realización dicha muestra incluye al menos un millón de células.

Cuando una fuente de material de la que se toma una muestra es escasa, tal como, muestras de estudios clínicos o similares, el ADN del material se puede amplificar mediante una técnica sin desviación, tal como la amplificación del genoma completo (WGA), amplificación de desplazamiento múltiple (MDA); o una técnica similar, por ejemplo, Hawkins et al, Curr. Opin. Biotech., 13: 65-67 (2002); Dean et al, Genome Research, 11: 1095-1099 (2001); Wang et al, Nucleic Acids Research, 32: e76 (2004); Hosono et al, Genome Research, 13: 954-964 (2003); y similares.

Las muestras de sangre son de particular interés y se pueden obtener utilizando técnicas convencionales, por ejemplo, Innis et al, editores, PCR Protocols (Academic Press, 1990); y similares. Por ejemplo, se pueden separar los glóbulos blancos de las muestras de sangre utilizando técnicas convencionales, por ejemplo, el kit RosetteSep (Stem Cell Technologies, Vancouver, Canadá). Las muestras de sangre pueden variar de volumen desde 100 µl a 10 ml; en un aspecto, los volúmenes de la muestra de sangre están en el intervalo de desde 200 100 µl a 2 ml. Entonces se puede extraer el ADN y/o ARN de dicha muestra de sangre utilizando técnicas convencionales para su uso en los métodos de la invención, por ejemplo, el kit DNeasy Blood & Tissue (Qiagen, Valencia, CA). Opcionalmente, los subconjuntos de glóbulos blancos, por ejemplo, linfocitos, se pueden aislar adicionalmente utilizando técnicas convencionales, por ejemplo, por clasificación celular activada fluorescentemente (FACS) (Becton Dickinson, San Jose, CA), clasificación celular activada magnéticamente (MACS) (Miltenyi Biotec, Auburn, CA), y similares.

Como las recombinaciones de identificación están presentes en el ADN de cada una de las células de inmunidad adaptativa del individuo, así como en sus transcripciones de ARN, se puede secuenciar el ARN o el ADN en los métodos de la invención que se proporciona. Se hace referencia a una secuencia recombinada de una célula T o célula B que codifica un receptor de célula T o molécula de inmunoglobulina, o una parte de la misma, como un clonotipo. El ADN o ARN puede corresponderse con secuencias genéticas del receptor de células T (TCR) o genes de inmunoglobulina (Ig) que codifican anticuerpos. Por ejemplo, el ADN y ARN puede corresponderse con secuencias que codifican cadenas α, β, γ, o δ de un TCR. En una mayoría de las células T, el TCR es un heterodímero que consiste en una cadena α, y una cadena β. La cadena de TCRα se genera por una recombinación VJ, y la cadena β del receptor se genera por la recombinación V(D)J. Para la cadena TCRβ, en los seres humanos hay 48 segmentos V, 2 segmentos D, y 13 segmentos J. Varias bases se pueden eliminar y otras se añaden (llamados nucleótidos N y P) en cada una de las dos uniones. En una minoría de células T, los TCR consisten en cadenas γ y δ delta. La cadena TCRγ se genera por una recombinación VJ, y la cadena de TCR δ se genera por recombinación V(D)J (Kenneth Murphy, Paul Travers, y Mark Walport, Janeway's Immunology 7ª edición, Garland Science, 2007).

El ADN y ARN analizado en los métodos de la invención pueden corresponderse con secuencias que codifican la cadena pesada de inmunoglobulinas (IgH) con regiones constantes (α, δ, ε, γ, o µ) o la cadena ligera de inmunoglobulinas (IgK o IgL) con regiones constantes λ o κ. Cada anticuerpo tiene dos cadenas ligeras idénticas y dos cadenas pesadas idénticas. Cada cadena está compuesta por una región constante (C) y una región variable. Para la cadena pesada, la región variable está compuesta por los segmentos variable (V), de diversidad (D), y de unión (J). Varias secuencias distintas que codifican cada tipo de estos segmentos están presentes en el genoma. Un evento de recombinación VDJ específica se produce durante el desarrollo de una célula B, marcando esa célula para generar una cadena pesada específica. La diversidad en la cadena ligera se genera de manera similar excepto que no hay una región D por lo que solo hay una recombinación VJ. A menudo se produce una mutación somática cerca del sitio de la recombinación, causando la adición o eliminación de varios nucleótidos., aumentando adicionalmente la diversidad de cadenas pesadas y ligeras generadas por las células B. La posible diversidad de los anticuerpos generados por una célula B es entonces el producto de diferentes cadenas pesadas y ligeras. Las regiones variables de las cadenas pesadas y ligeras contribuyen a la formación de la región o sitio de unión de reconocimiento (o unión). Se añade a esta diversidad un proceso de hipermutación somática que puede producirse después de que se monte una respuesta específica contra algún epítopo.

Como se ha mencionado anteriormente, de acuerdo con la invención, se pueden seleccionar cebadores para generar amplicones de subconjuntos de ácidos nucleicos recombinados extraídos de los linfocitos. Se puede hacer referencia a dichos subconjuntos en el presente documento como “regiones reordenadas somáticamente”. Las regiones reordenadas pueden comprender ácidos nucleicos del desarrollo o de linfocitos completamente desarrollados, donde los linfocitos en desarrollo son células en los que el reordenamiento de los genes inmunitarios no se ha completado para formar moléculas que tengan regiones V(D)J completas. Las regiones reordenadas somáticamente incompletas ejemplares incluyen moléculas de IgH incompletas (tal como, moléculas que contienen solo regiones D-J), moléculas de TCRδ incompletas (tales como moléculas que solo contienen regiones D-J),e IgK inactivas (por ejemplo, que comprenden regiones Kde-V).

El muestreo adecuado de las células es un aspecto importante de interpretación de los datos del repertorio, como se describe adicionalmente posteriormente en las definiciones de “clonotipo” y “repertorio”. Por ejemplo, comenzando con 1.000 células se crea una frecuencia mínima a la que es sensible el ensayo independientemente de cuantas lecturas de secuenciación se obtengan. Por lo tanto, un aspecto de la presente invención es el desarrollo de métodos para cuantificar el número de moléculas de receptor inmunitario de entrada. Esto se ha implementado para las secuencias de TCRβ e IgH. En cualquier caso, se utiliza el mismo conjunto de cebadores que son capaces de amplificar todas las secuencias diferentes. Con el fin de obtener un número de copias absoluto, se llevó a cabo una PCR en tiempo real con múltiples cebadores junto con una convencional con un número conocido de copias del receptor inmunitario. Esta medición por la PCR en tiempo real se puede hacer a partir de la reacción de amplificación que se secuenciará posteriormente o se puede hacer con una alícuota separada de la misma muestra. En el caso del ADN, el número absoluto de moléculas de receptor inmunitario reordenadas se pueden fácilmente convertir en número de células (por 2 veces ya que algunas células tendrán 2 copias reordenadas del receptor inmunitario específico evaluado y otras tendrán una). En el caso del ADNc el número total medido de moléculas reordenadas en la muestra de tiempo real se puede extrapolar para definir el número total de estas moléculas utilizadas en otra reacción de amplificación de la misma muestra. Además, este método se puede combinar con un método para determinar la cantidad total de ARN para definir el número de moléculas de receptor inmunitario reordenado en una cantidad unitaria (digamos 1 µg) de ARN asumiendo una eficacia específica de síntesis de ADNc. Si se mide la cantidad total de ADNc entonces no es necesario considerar la eficacia de síntesis de ADNc. Si el número de células también es conocido entonces las copias de receptor inmunitario reordenado por célula se puede computar. Si el número de célula no se conoce, se puede estimar a partir del ARN total ya que las células de un tipo específico habitualmente generan una cantidad de ARN comparable. Por lo tanto, a partir de las copias de moléculas de receptor inmunitario reordenada por 1 µg se puede estimar el número de estas moléculas por célula.

Una desventaja de hacer una PCR en tiempo real de la reacción que se procesaría para la secuenciación es que podría haber efectos inhibidores que son diferentes en la PCR en tiempo real que en la otra reacción, ya que se pueden utilizar diferentes enzimas, ADN de entrada, y otras condiciones. El procesamiento de los productos de la PCR en tiempo real para la secuenciación mejoraría este problema. Sin embargo, el bajo número de copias utilizando una PCR en tiempo real se puede deber a un bajo número de copias o a efectos inhibidores, u otras condiciones subóptimas en la reacción.

Otra estrategia que se puede utilizar es añadir una cantidad conocida de moléculas de receptor inmunitario reordenadas únicas con una secuencia conocida, es decir, cantidades conocidas de referencias internas, para el ADNc o ADN genómico de una muestra de cantidad desconocida. Contando el número relativo de moléculas que se obtienen de la secuencia conocida añadida en comparación con el resto de las secuencias de la misma muestra, se puede estimar el número de moléculas de receptor inmunitario reordenadas en la muestra de ADNc inicial. (Dichas técnicas para el recuento moléculas se conocen bien, por ejemplo, Brenner et al, Patente de EE. UU. 7.537.897). Los datos de la secuenciación de la secuencia única añadida se pueden utilizar para distinguir las diferentes posibilidades si se va a utilizar también una calibración por PCR en tiempo real. El bajo número de copias del receptor inmunitario reordenado en el ADN (o ADNc) crearía una relación alta entre el número de moléculas respecto a la secuencia añadida en comparación con el resto de las secuencias de la muestra. Por otra parte, si la medición de un número bajo de copias por PCR en tiempo real es debido a la ineficacia de la reacción, la relación no debería ser alta.

Amplificación de poblaciones de ácidos nucleicos

Se pueden generar amplicones de ácidos nucleicos de poblaciones diana mediante una variedad de técnicas de amplificación. En un aspecto de la invención, se utiliza una PCR múltiple para amplificar los miembros de una mezcla de ácidos nucleicos, particularmente mezclas que comprenden moléculas inmunitarias recombinadas tales como receptores de células T, o partes de las mismas. La directriz para llevar a cabo PCR múltiples de dichas moléculas inmunitarias se encuentra en las siguientes referencias: Morley, Patente de EE. UU. 5.296.351; Gorski, Patente de EE. UU. 5.837.447; Dau, Patente de EE. UU. 6.087.096; Von Dongen et al, Publicación de Patente de EE. UU. 2006/0234234; Publicación de Patente Europea EP 1544308B1; y similares.

Después de la amplificación del ADN del genoma (o amplificación de ácidos nucleicos en forma de ADNc por transcripción inversa de ARN), se pueden aislar moléculas de ácido nucleico individuales, re-amplificarse opcionalmente, y entonces secuenciarse individualmente. Los protocolos de amplificación ejemplares se pueden

encontrar en van Dongen et al, Leukemia, 17: 2257-2317 (2003) o van Dongen et al, Publicación de Patente de EE. UU. 2006/0234234. En resumen, un protocolo ejemplar es el siguiente: Tampón de reacción: tampón ABI II o Tampón ABI Gold (Life Technologies, San Diego, CA); 50 µl de volumen de reacción final; 100 ng de ADN de muestra; 10 pmol de cada cebador (sometido a ajustes para equilibrar la amplificación como se describe posteriormente); dNTP a 200 µM de concentración final; MgCl2 a 1,5 mM de concentración final (sometido a optimización dependiendo delas secuencias diana y la polimerasa); polimerasa Taq (1-2 U/tubo); condiciones de ciclado: pre-activación 7 min a 95 ºC; hibridación a 60 ºC; tiempos de ciclado: desnaturalización 30 s; hibridación 30 s; extensión 30 s. Las polimerasas que se pueden utilizar para la amplificación en los métodos de la invención están disponibles en el mercado e incluyen, por ejemplo, polimerasa Taq, polimerasa AccuPrime, o Pfu. La elección de polimerasa que se va a utilizar se puede basar en si se prefiere la fidelidad o la eficacia.

Se pueden utilizar mediciones por PCR en tiempo real, tinción picogreen, electroforesis de nanofluídica (por ejemplo, LabChip) o absorción de UV en una etapa inicial para juzgar la cantidad funcional de material amplificable.

En un aspecto se llevan a cabo amplificaciones múltiples de manera que las cantidades relativas de secuencias en una población de partida sea sustancialmente la misma que en la población amplificada, o amplicón. Es decir, las amplificaciones múltiples se llevan a cabo con una mínima tendencia de amplificación entre las secuencias miembros de una población de muestras. En una realización, dichas cantidades relativas son sustancialmente las mismas si cada cantidad relativa en un amplicón es de cinco veces su valor en la muestra de partida. En otra realización, dichas cantidades relativas son sustancialmente las mismas si cada cantidad relativa en un amplicón es de dos veces su valor en la muestra de partida. Como se expone más completamente posteriormente, la tendencia de amplificación en la PCR se puede detectar y corregir utilizando técnicas convencionales de manera que se puede seleccionar un conjunto de cebadores de PCR para un repertorio predeterminado que proporcione una amplificación no desviada de cualquier muestra.

Con respecto a muchos repertorios basados en secuencias de TCR o BCR, una amplificación múltiple utiliza opcionalmente todos los segmentos V. La reacción se optimiza para intentar tener la amplificación que mantenga la abundancia relativa de las secuencias amplificadas por diferentes cebadores del segmento V. Algunos de los cebadores están relacionados, y por lo tanto muchos de los cebadores pueden “intercomunicarse”, amplificando matrices que no coinciden perfectamente con ellos. Las condiciones se optimizan de manera que cada matriz se pueda amplificar de una manera similar independientemente del cebador que lo amplifique. En otras palabras, si hay dos matrices, entonces después de una amplificación de 1.000 veces ambas matrices se pueden amplificar aproximadamente 1.000 veces, y no importa que para una de las matrices la mitad de los productos amplificados lleven un cebador diferente debido a la intercomunicación. En análisis posteriores de los datos de secuenciación la secuencia de cebador se elimina del análisis y por tanto no importa el cebador que se utilice en la amplificación a condición de que la matriz se amplifique de manera igual.

En una realización, la tendencia de amplificación puede evitarse llevando a cabo una amplificación de dos etapas (como se describe en Faham y Willis, citado anteriormente) en la que se implementa un pequeño número de ciclos en una primera, o etapa primaria, utilizando cebadores que tienen colas no complementarias con las secuencias diana. Las colas incluyen sitios de unión al cebador que se han añadido a los extremos de las secuencias del amplicón primario de manera que dichos sitios se utilizan en una segunda etapa de amplificación utilizando solo un único cebador directo y un único cebador inverso, eliminando de esta manera una causa primaria de tendencia de amplificación. Preferentemente, la PCR primaria tendrá un número lo suficientemente pequeño de ciclos (por ejemplo, de 5-10) para minimizar la amplificación diferencial de los diferentes cebadores. La amplificación secundaria se hace con un par de cebadores y por lo tanto el problema de la amplificación diferencial es mínimo. Un porcentaje de la PCR primaria se toma directamente para la PCR secundaria. Eran suficientes treinta y cinco ciclos (equivalente a ~ 28 ciclos sin la etapa de dilución de 100 veces) utilizados entre las dos amplificaciones para mostrar una amplificación robusta independientemente de si la repartición era: un ciclo primario y 34 secundarios o 25 primarios y 10 secundarios. Incluso aunque idealmente haciendo solo un ciclo en la PCR primaria se pueda disminuir la tendencia de amplificación, hay otras consideraciones. Un aspecto de esto es la representación. Esto tiene un papel cuando la cantidad entrante de partida no excede el número de lecturas obtenidas en último término. Por ejemplo, si se obtienen 1.000.000 de lecturas y se inició con 1.000.000 de moléculas entrantes entonces al tomar solo una representación de 100.000 moléculas de la amplificación secundaria se degradaría la precisión de la estimación de la abundancia relativa de las diferentes especies en la muestra original. Una dilución de 100 veces entre las 2 etapas significa que la representación se reduce a menos que la amplificación por PCR primaria genere significativamente más de 10 moléculas. Esto indica que se puede utilizar un mínimo de 8 ciclos (256 veces), pero más confortablemente 10 ciclos (~ 1.000 veces). La alternativa a esto es tomar más del 1 % de la PCR primaria en la secundaria, pero debido a la alta concentración del cebador utilizado en la PCR primaria, se puede utilizar un factor de dilución grande para asegurar que estos cebadores no interfieran en la amplificación y empeoren la tendencia de amplificación entre las secuencias. Otra alternativa es añadir una etapa de purificación o enzimática para eliminar los cebadores de la PCR primaria para permitir una dilución menor de esta. En este ejemplo, la PCR tenía 10 ciclos y la segunda 25.

Generación de lecturas de secuencias por clonotipos

Se puede utilizar cualquier técnica de alto rendimiento para la secuenciación de ácidos nucleicos en el método de la invención. Preferentemente, dicha técnica tendrá la capacidad de generación de manera barata un volumen de datos de secuencias de las que se pueden determinar al menos 1000 clonotipos, y preferentemente, de las que se pueden determinar al menos 10.000 a 1.000.000 de clonotipos. Las técnicas de secuenciación de ADN incluyen las reacciones de secuenciación didesoxi clásicas (Método de Sangre) que utiliza terminadores o cebadores marcados y separación en gel en plancha o capilaridad, secuenciación por síntesis utilizando nucleótidos marcados terminados reversiblemente, pirosecuenciación, secuenciación 454, hibridación de alelos específicos a una biblioteca de sondas de oligonucleótido marcadas, secuenciación por síntesis utilizando hibridación de alelos específicos a una biblioteca de clones marcados que continúa con la ligadura, control en tiempo real de la incorporación de nucleótidos marcados durante una etapa de polimerización, secuenciación con polonio y secuenciación SOLiD. La secuenciación de moléculas separadas se ha demostrado recientemente más por reacciones de extensión secuencial o sencilla utilizando polimerasas o ligasas, así como por hibridaciones diferenciales secuenciales o sencillas con bibliotecas de sodas. Estas reacciones se han llevado a cabo en muchas secuencias clónicas en paralelo incluyendo las demostraciones en aplicaciones actuales en el mercado de más de 100 millones de secuencias en paralelo. Estas estrategias de secuenciación se pueden utilizar de esta manera para estudiar el repertorio de un receptor de células T (TCR) y/o receptor de células B (BCR). En un aspecto de la invención , se emplean métodos de secuenciación de altas prestaciones que comprenden una etapa de asilamiento espacial de moléculas en una superficie sólida en la que se secuencian en paralelo. Dichas superficies sólidas pueden incluir superficies no porosas (tales como en la secuenciación Solexa, por ejemplo, Bentley et al, Nature, 456: 53-59 (2008) o secuenciación Genómica Completa, por ejemplo, Drmanac et al, Science, 327: 78-81 (2010)), matrices de pocillos, que pueden incluir matrices unidas a perlas o partículas (tales como en 454, por ejemplo, Margulies et al, Nature, 437: 376-380 (2005) o secuenciación de Ion Torrent, Publicación de Patente de EE. UU. 2010/0137143 o 2010/0304982, membranas microfabricadas (tales como en la secuenciación SMRT, por ejemplo, Eid et al, Science, 323: 133-138 (2009)), o matrices de perlas como en la secuenciación SOLiD o la secuenciación con polonio, por ejemplo, Kim et al, Science, 316: 1481-1414 (2007)). En otro aspecto, dichos métodos comprenden la amplificación de moléculas asiladas antes o después de aislarse espacialmente en una superficie sólida. La amplificación anterior puede comprender una amplificación basada en una emulsión, tal como una emulsión PCR, o amplificación de círculo rodante. De particular interés es la secuenciación basada en Solexa en la que las moléculas de matriz individuales se aíslan espacialmente en una superficie sólida, tras lo cual se amplifican en paralelo por una PCR de puente para formar poblaciones clónicas separadas, o agrupamientos, y entonces se secuencian como se describe en Bentley et al (citado anteriormente) y en las instrucciones del fabricante (por ejemplo, el kit de preparación de muestra TruSeq™ y hoja de datos, Illumina Inc., San Diego, CA, 2010); y adicionalmente en las siguientes referencias: Patentes de EE. UU. 6.090.592; 6.300.070; 7.115.400; y documento EP 0972081B1. En una realización, las moléculas individuales dispuestas y amplificadas en una superficie sólida forma agrupamientos con una densidad de al menos 10^5 agrupamientos por cm^2; o con una densidad de al menos 5 x 10^5 por cm^2; o con una densidad de 10^6 por cm^2. En una realización se emplean secuenciaciones químicas que tienen relativamente altas tasas de error. En dichas realizaciones, los valores de calidad media producida por dicha química son funciones que disminuyen monótonamente las longitudes de secuencia leídas. En una realización, dicha disminución se corresponde con un 0,5 por ciento de lecturas de secuencia que tienen al menos un error en la posición 1-75; un 1 por ciento de lecturas de secuencias tienen al menos un error en las posiciones 76-100; y un 2 por ciento de lecturas de secuencia tienen al menos un error en las posiciones 101-125.

Un perfil de clonotipos basado en la secuencia de un individuo se puede obtener utilizando las siguientes etapas: (a) obtener una muestra de ácido nucleico de las células T y/o células B del individuo; (b) aislar espacialmente las moléculas individuales derivadas de dicha muestra de ácido nucleico, comprendiendo las moléculas individuales al menos una matriz generada a partir de un ácido nucleico de la muestra, cuya matriz comprende una región reordenada somáticamente o una parte de la misma, siendo capaz cada molécula individual de producir al menos una lectura de secuencia; (c) secuenciar dichas moléculas aisladas espacialmente; y (d) determinar la abundancia de diferentes secuencias de las moléculas de ácido nucleico de la muestra de ácido nucleico para generar el perfil de clonotipos. En un caso, cada una de las regiones reordenadas somáticamente comprenden una región V y una región J. En otro caso, la etapa de secuenciación comprende la secuenciación bidireccional de cada una de las moléculas individuales aisladas espacialmente para producir al menos una lectura de secuencia directa y al menos una lectura de secuencia inversa. Adicionalmente al último caso, al menos una de las lecturas de secuencia directa y al menos una de las lecturas de secuencia inversa, tiene una región solapada con las bases, de manera que las bases de dicha región solapada están determinadas por una relación complementariamente inversa entre dichas lecturas de secuencia. En otro caso más, cada una de las regiones reordenadas comprende una región V y una región J y la etapa de secuenciación incluyen adicionalmente la determinación de una secuencia de cada una de las moléculas de ácido nucleico individuales de una o más de sus lecturas de secuencia directa y al menos una lectura de secuencia inversa que comienza en una posición en una región J y se extiende en dirección de su región V asociada. En otro caso, las moléculas individuales comprenden ácidos nucleicos seleccionados de entre el grupo que consiste en moléculas de IgH completa, moléculas de IgH incompleta, moléculas de IgK completa, moléculas de IgK inactiva, moléculas de TCRβ, moléculas de TCRγ, moléculas de TCRδ completa, y moléculas de TCRδ incompleta. En otro caso, la etapa de secuenciación comprende la generación de lecturas de secuencia que tienen valoraciones de calidad que disminuyen monótonamente. Además del último caso, las valoraciones de calidad que

disminuyen monótonamente son de tal manera que las lecturas de secuencia tienen tasas de error no mejores de las siguientes: un 0,2 por ciento de las lecturas de secuencia contienen al menos un error en las posiciones de bases 1 a 50, un 0,2 a 1,0 por ciento de las lecturas de secuencias contienen al menos un error en las posiciones 51-75, un 0,5 a 1,5 por ciento de lecturas de secuencia contienen al menos un error en las posiciones 76-100. En otro caso, el método anterior comprende las siguientes etapas: (a) obtener una muestra de ácido nucleico a partir de las células T y/o células B del individuo; (b) aislar espacialmente moléculas individuales derivadas de dicha muestra de ácido nucleico, comprendiendo las moléculas individuales conjuntos anidados de matrices cada uno generados de un ácido nucleico de la muestra y que cada uno contiene una región reordenada somáticamente o una parte de la misma, siendo capaz cada conjunto anidado de producir una pluralidad de lecturas de secuencia que se extiende cada una en la misma dirección y que se inicia cada una en una posición diferente del ácido nucleico a partir de la cual se genera el conjunto anidado; (c) secuenciar dichas moléculas aisladas espacialmente individuales; y (d) determinar las abundancias de diferentes secuencias de las moléculas de ácido nucleico de la muestra de ácido nucleico para generar el perfil de clonotipos. En un caso, la etapa de secuenciación incluye la producción de una pluralidad de lecturas de secuencia para cada uno de los grupos anidados. En otro caso cada una de las regiones reordenadas somáticamente comprenden una región V y una región J, y cada una de la pluralidad de lecturas de secuencia comienza a partir de una posición diferente en la región V y se extiende en la dirección de su región J asociada.

En un aspecto, para cada muestra de un individuo, la técnica de secuenciación utilizada en los métodos de la invención genera secuencias de al menos 1000 clonotipos por ejecución; en otro aspecto, dicha técnica genera secuencias de al menos 10.000 clonotipos por ejecución; en otro aspecto, dicha técnica genera secuencias de al menos 100.000 clonotipos por ejecución; en otro aspecto, dicha técnica genera secuenciadas de al menos 500.000 clonotipos por ejecución; y en otro aspecto, dicha técnica genera secuencias de al menos 1.000.000 clonotipos por ejecución. En otro aspecto más, dicha técnica genera secuencias de entre 100.000 a 1.000.000 clonotipos por ejecución por cada muestra individual.

La técnica de secuenciación utilizada en los métodos de la invención proporcionada puede 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 del clonotipo a partir de los datos de secuencia

La construcción de clonotipos a partir de los datos de lectura de secuencia depende en parte del método de secuenciación que se utilice para generar dichos datos, ya que los diferentes métodos tienen diferentes expectativas de longitud de lectura y calidad de los datos. En una estrategia, se emplea un secuenciador Solexa para generar los datos de lectura de secuencia para el análisis. En una realización, una muestra proporciona al menos 0,5 – 1,0 x 10^6 linfocitos para producir al menos 1 millón de moléculas de matriz, que después de la amplificación opcional puede producir un millón o más de las correspondientes poblaciones clónicas de moléculas de matriz (o agrupamientos). Para las estrategias de secuenciación de mayor alto rendimiento, incluyendo la estrategia Solexa, es deseable dicho sobre-muestreo a nivel de agrupamiento ya que cada secuencia de matriz se determina con un alto grado de redundancia para aumentar la exactitud de la determinación de secuencia. Para las implementaciones basadas en Solexa, se determina preferentemente la secuencia de cada matriz independiente 10 veces o más. Para otras estrategias de secuenciación con diferentes expectativas de longitudes de lectura y calidad de datos, se pueden utilizar diferentes niveles de redundancia para una exactitud comparable de la determinación de secuencia. Los expertos en la técnica reconocen que los parámetros anteriores, por ejemplo, tamaño de muestra, redundancia, y similares, son elecciones de diseño relacionadas con las aplicaciones particulares.

En un aspecto de la invención, las secuencias de clonotipos (incluyendo, pero sin limitarse a las derivadas de IgH, TCRα, TCRβ, TCRγ, TCRδ, y/o IgLκ (IgK)) se pueden determinar combinando la información de una o más lecturas de secuencias, por ejemplo, a lo largo con las regiones V(D)J de las cadenas seleccionadas. En otro aspecto, las secuencias de clonotipos se determinan combinando la información de una pluralidad de lecturas de secuencia. Dichas pluralidades de lecturas de secuencia pueden incluir una o más lecturas de secuencia a lo largo de una cadena en sentido (es decir, lecturas de secuencia “directas” y una o más lecturas de secuencia a lo largo de su cadena complementaria (es decir, lecturas de secuencia “inversas”). Cuando se generan múltiples lecturas de secuencia a lo largo de la misma cadena, se generan primero matrices separadas amplificando moléculas de muestra con cebadores seleccionadas de diferentes posiciones de las lecturas de secuencia. Este concepto se ilustra en la Fig. 4A en la que se emplean cebadores (404, 406 y 408) para generar amplicones (410, 412, y 414, respectivamente) en una única reacción. Dichas amplificaciones se pueden llevar a cabo en la misma reacción o en reacciones separadas. En un aspecto, cuando sea que se emplee la PCR, se utilizan reacciones de amplificación separadas para generar matrices separadas que, a su vez, se combinan y se utilizan para generar múltiples lecturas de secuencia a lo largo de la misma cadena. Esta última estrategia es preferible para evitar la necesidad de equilibrar las concentraciones de cebador (y/u otros parámetros de reacción) para asegurar una amplificación igual

de las múltiples matrices (a lo que se hace referencia a veces en el presente documento como “amplificación equilibrada” o “amplificación sin desviación”). La generación de matrices en reacciones separadas se ilustra en las Fig. 4B-4C. En las que una muestra que contiene IgH (400) se divide en tres partes (472, 474, y 476) que se añaden a distintas PCR utilizando cebadores de la región J (401) y cebadores de la región V (404, 406, y 408, respectivamente) para producir amplicones (420, 422, y 424, respectivamente). Los últimos amplicones se combinan entonces (478) en una PCR secundaria (480) utilizando los cebadores P5 y P7 para preparar las matrices (482) para la PCR de puente y la secuenciación en un secuenciador GA de Illumina, o un instrumento similar.

Las lecturas de secuencia producidas por los métodos de la invención pueden tener una amplia variedad de longitudes, dependiendo en parte de la técnica de secuenciación que se emplee. Por ejemplo, para algunas técnicas, pueden producirse varias compensaciones en su implementación, por ejemplo, (i) el número y longitudes de lecturas de secuencias por matriz y (ii) el coste y duración de una operación de secuenciación. En una realización, las lecturas de secuencia están en el intervalo de desde 20 a 400 nucleótidos; en otra realización, las lecturas de secuencia están en un intervalo de desde 30 a 200 nucleótidos; en otra realización más, las lecturas de secuencia están en el intervalo de desde 30 a 120 nucleótidos. En una realización se generan 1 a 4 lecturas de secuencia para la determinación de la secuencia de cada clonotipo; en otra realización se generan 2 a 4 lecturas de secuencia para la determinación de cada clonotipo; y en otra realización se generan 2 a 3 lecturas de secuencia para la determinación de cada clonotipo. En las realizaciones anteriores, el número que se da es exclusivo de las lecturas de secuencia utilizadas para identificar las muestras de diferentes individuos. Las longitudes de distintas lecturas de secuencia que se utilizan en las realizaciones descritas posteriormente también pueden variar basándose en la información que se desea capturar por la lectura; por ejemplo, la localización de partida y la longitud de una lectura de secuencia se puede diseñar para proporciona la longitud de una región NDN, así como su secuencia de nucleótidos; por lo tanto, se seleccionan las lecturas de secuencia que abarcan la región NDN completa. En otros aspectos, son suficientes una o más lecturas de secuencia que en combinación (pero no separadamente) engloba una región D y/o NDN.

En otro aspecto de la invención, las secuencias de clonotipos se determinan en parte por lecturas de alineamiento de secuencias con una o más secuencias de referencia de la región V y una o más secuencias de referencia de la región J, y en parte por la determinación de bases sin alineamiento con las secuencias de referencia, tal como en la región NDN altamente variable. Se puede aplicar una variedad de algoritmos de alineamiento a las lecturas de secuencia y secuencias de referencia. Por ejemplo, una directriz para la sección de los métodos de alineamiento está disponible en Batzoglou, Briefings in Bioinformatics, 6: 6-22 (2005). En un aspecto, cualquiera de las lecturas V o lecturas C (como se ha mencionado anteriormente) se alinean con las secuencias de referencia de V y J, se emplea un algoritmo de árbol de búsqueda, por ejemplo, como describe en general Gusfield (citado anteriormente) y Cormen et al, Introduction to Algorithms, Tercera Edición (The MIT Press, 2009).

En otro aspecto, un extremo de al menos una lectura directa y un extremo de una lectura inversa se solapan en una región de solapamiento (por ejemplo, 308 en la Fig. 3B), de manera que las bases de las lecturas están una relación complementariamente inversa entre ellas. Por lo tanto, por ejemplo, si una lectura directa de la región solapada es “5’-acgttgc”, entonces una lectura inversa en una relación complementaria inversa es “5’-gcaacgt" en la misma región solapada. En un aspecto, se determinan las bases de dicha región de solapamiento, al menos en parte, a partir de dicha relación de complementariedad inversa. Es decir, una probabilidad de identificación de una base (o una valoración de calidad relacionada) en una región solapada posible aumenta si se conserva, o es consistente con, una relación complementaria inversa entre las dos lecturas de secuencia. En un aspecto, los clonotipos de cadenas TCRβ e IgH (que se ilustra en la Fig. 3B) se determinan mediante al menos una lectura de secuencia comenzando en su región J y extendiéndose en la dirección de su región V asociada (a la que se hace referencia en el presente documento como una “lectura C” (304)) y al menos una lectura de secuencia que comienza en su región V y se extiende en la dirección de su región J asociada (a la que se hace referencia en el presente documento como una “lectura V” (306)). La región solapada (308) puede englobar o no la región NDN (315) como se muestra en la Fig. 3B. La región solapada (398) puede estar completamente en la región J, completamente en la región NDN, completamente en la región V, o puede englobar el límite entre una región J y una región NDN o el límite entre una región V y una región NDN, o ambos de dichos límites (como se ilustra en la Fig. 3B). Normalmente, dichas lecturas de secuencias se generan extendiendo los cebadores de secuencia, por ejemplo (302 y (310) en la Fig. 3B, con una polimerasa en una reacción secuenciación-por-síntesis, por ejemplo, Metzger, Nature Reviews Genetics, 11: 31-46 (2010); Fuller et al, Nature Biotechnology, 27: 1013-1023 (2009). Los sitios de unión para los cebadores (302) y (310) están predeterminados, de manera que pueden proporcionar un punto de partida por punto de anclaje para el alineamiento inicial y el análisis de las lecturas de secuencia. En una realización, una lectura C se posiciona de manera que englobe la región D y/o NDN de la cadena TCRβ o IgH e incluye una parte de la región adyacente, por ejemplo, como se ilustra en las Fig. 3B y 3C. En un aspecto el solapamiento de la lectura V y la lectura C en la región se utiliza para alinear las lecturas entre ellas. En otras realizaciones, dicho alineamiento de las lecturas de secuencia no es necesario, por ejemplo, en las cadenas TCRβ, de manera que solo una lectura V puede ser lo suficientemente larga para identificar la región V particular de un clonotipo. Este último aspecto se ilustra en la Fig. 3C. La lectura de secuencia (330) se utiliza para identificar una región V, con o sin solapamiento con otra lectura de secuencia, y otra lectura de secuencia (332) atraviesa la región NDN y se utiliza para determinar la secuencia de la misma. La parte (334) de la lectura de secuencia (332) que se extiende en la región V se utiliza para asociar la información de lectura de secuencia (332) con la de la lectura de secuencia (330) para determinar un clonotipo. Para

algunos métodos de secuenciación, dichas estrategias base-por-base como el método de secuenciación Solexa, el tiempo de ejecución de la secuenciación y los costes se reducen minimizando el número de ciclos de secuenciación en un análisis. Opcionalmente, como se ilustra en la Fig. 3B, el amplicón (300) se produce con un marcador de muestra (313) para distinguir entre los clonotipos que se originan de diferentes muestras biológicas, por ejemplo, de diferentes pacientes. El marcador de muestra (312) puede identificarse hibridando un cebador a una región de unión al cebador (316) y extendiéndolo (314) para producir una lectura de secuencia a través del marcador (312), del cual se decodifica el marcador de muestra (312).

La cadena de IgH es más desafiante de analizar que la cadena TCRβ debido a al menos dos factores: i) la presencia de mutaciones somáticas hace que el mapeo o alineamiento sea más difícil, y ii) la región NDN es mayor de manera que a menudo no es posible mapear una parte del segmento V de la lectura C. En un aspecto de la invención, este problema se supera utilizando una pluralidad de conjuntos de cebadores para generar lecturas V, que se localizan en diferentes localizaciones a lo largo de la región V, preferentemente de manera que los sitios de unión al cebador no se solapen y se espacien por separado, y con al menos un sitio de unión al cebador adyacente a la región NDN, por ejemplo, en una realización de desde 5 a 50 bases de la unión V-NDN, o en otra realización desde 10 a 50 bases de la unión V-NDN. La redundancia de una pluralidad de conjuntos de cebadores minimiza el riesgo de fallar en la detección de un clonotipo debido a un fallo de uno o dos cebadores que tienen los sitios de unión afectados por mutaciones somáticas. Además, la presencia de al menos un sitio de unión al cebado adyacente a la región NDN lo hace más probable que una lectura V se solape con la lectura C y por lo tanto extienda eficazmente la longitud de la lectura C. Esto permite la generación de una secuencia continua que abarba todos los tamaños de regiones NDN y que también pueden mapear sustancialmente las regiones V y J completas a ambos lados de la región NDN. Las realizaciones para llevar a cabo dicho esquema se ilustran en las Figs. 4A y 4D. En la Fig. 4A, se secuencia una muestra que comprende cadenas IgH (400) generando una pluralidad de amplicones para cada cadena amplificando las cadenas con un conjunto único de cebadores de la región J (401) y una pluralidad (se muestran tres) de conjuntos de cebadores (404, 406, 408) de la región V (402) para producir una pluralidad de amplicones anidados (por ejemplo, 410, 412, 416) que comprenden todos la misma región NDN y que tienen diferentes longitudes que engloban partes sucesivamente mayores (411, 413, 415) de la región V (402). Los miembros de un conjunto anidado se pueden agrupar en conjunto después de la secuenciación anotando la identidad (o identidad sustancial) de sus regiones NDN, J y/o C respectivas, permitiendo de esta manera la reconstrucción de un segmento V(D)J más largo que sería el caso por otra parte de una plataforma de secuenciación con una longitud y/o calidad de lectura limitadas. En una realización, la pluralidad de conjuntos de cebadores puede ser un número en el intervalo de 2 a 5. En otra realización la pluralidad es de 2-4; y en otra realización más, la pluralidad es 3. Las concentraciones y posiciones de los cebadores en una pluralidad pueden variar ampliamente. Las concentraciones de los cebadores de la región V puede ser la misma o no. En una realización, el cebador más cercano a la región NDN tiene una concentración mayor que los otros cebadores de la pluralidad, por ejemplo, para asegurar que los amplicones que contienen la región NDN estén representados en el amplicón resultante. En una realización particular en la que se emplea una pluralidad de tres cebadores, se utiliza una relación de concentración de 60:20:20. Se pueden emplear uno o más cebadores (por ejemplo, 435 y 437 en la Fig. 4B) adyacentes a la región NDN (444) para generar una o más lecturas de secuencias (por ejemplo, 434 y 436) que se solapa con la lectura de secuencia (442) generada por el cebador de la región J (432), mejorando de esta manera la calidad de bases identificadas en la región solapada (440). Las lecturas de secuencia de la pluralidad de cebadores pueden solaparse o no con el sitio de unión al cebador adyacente corriente abajo y/o la lectura de secuencia adyacente corriente abajo. En una realización, las lecturas de secuencia proximales a la región NDN (por ejemplo, 436 y 438) se puede utilizar para identificar la región V particular asociada con el clonotipo. Dicha pluralidad de cebadores reduce la probabilidad de una amplificación incompleta o fallida en el caso que uno de los sitios de unión al cebador esté hipermutado durante el desarrollo de la inmunoglobulina. También aumenta la probabilidad de la diversidad introducida por la hipermutación de la región V será capturado en una secuencia de clonotipo. Se puede llevar a cabo una PCR secundaria para preparar los amplicones anidados para la secuenciación, por ejemplo, amplificando con los cebadores P5 (401) y P7 (404, 406, 408) como se ilustra para producir amplicones (420, 422, y 424), que se puede distribuir como moléculas únicas en una superficie sólida, en la que se amplifican adicionalmente por una PCR de puente, o una técnica similar.

La identificación de bases en las regiones NDN (particularmente en las cadenas de IgH) se pueden mejorar utilizando la estructura de codón de las regiones J y V flanqueantes, como se ilustra en la Fig. 4E (como se utiliza en el presente documento, “estructura de codón” significa los codones de la fase de lectura natural de los segmentos de las transcripciones o genes de TCR o BCR fuera de las regiones NDN, por ejemplo, la región V, la región J o similares). El amplicón (450), que es una vista aumentada del amplicón de la Fig. 4B, se muestra junto con las posiciones relativas de la lectura C (442) y la lectura V (434) adyacente encima y las estructuras de codón (452 y 454) de la región V (430) y la región J (446) , respectivamente, debajo. De acuerdo con este aspecto de la invención, después de que se identifican las estructuras de codón (452 y 454) de la región V (430) y región J (446), respectivamente, debajo. De acuerdo con este aspecto de la invención, después de las estructuras (452 y 454) se identifican por alineamiento convencional de las secuencias de referencia V y J, las bases de la región NDN (456) se definen (o identifican) una base cada vez se mueve de la región J (446) hacia la región V (430) y en la dirección opuesta desde la región V (430) hacia la región J (446) utilizando las lecturas de secuencia (454) y (442). En condiciones biológicas normales, solamente las secuencia de TCR o IgH recombinadas que tienen codones en fase a partir de la región V a través de la región NDN y a la región J se expresan como proteínas. Es decir, de las variantes generadas somáticamente las únicas que se expresan son cuyas fases de codón de región J y región V

están en fase entre ellas y se mantienen en fase a través de la región NDN. (Aquí las fases correctas de las regiones V y J se determinan a partir de las secuencias de referencia). Si se identifica una secuencia fuera de fase basándose en una o más identificaciones de base de baja calidad, el clonotipo correspondiente se marca para reevaluación o como una anomalía relacionada con una enfermedad potencial. Si la secuencia identificada está en fase y se basa en identificaciones de base de alta calidad, entonces hay mayor confianza en que el clonotipo correspondiente se ha identificado correctamente. En consecuencia, en un aspecto, la invención incluye un método de determinación de clonotipos basados en V(D)J a partir de lecturas de secuencia bidireccionales que comprende las etapas de: (a) generar al menos una lectura de secuencia de la región J y la lectura de secuencia de región V están solapadas en una región de solapamiento, y la región J y la región V tienen, cada una, una estructura de codón; (b) determinar si la estructura de codón de la región J extendida en la región NDN está en fase con la estructura de codón de la región V extendida hacia la región NDN. En una realización adicional, la etapa de generación incluye la generación de al menos una lectura de que comienza en la región V y se extiende a través de la región NDN hacia la región J, de manera que la lectura de secuencia de la región J y la lectura de secuencia de la región V están solapadas en una región de solapamiento.

Hipermutaciones somáticas. En una realización, os clonotipos basados en IgH que se han sometido a hipermutación somática se determinan de la siguiente manera. Una mutación somática se define como una base secuenciada que es diferente de la base correspondiente de una secuencia de referencia (del segmento relevante, habitualmente V, J o C) y que está presente en un número estadísticamente significativo de lecturas. En una realización, las lecturas C se pueden utilizar para encontrar mutaciones somáticas con respecto al segmento J mapeado y de igual mente las lecturas V para el segmento V. Solamente se utilizan trozos de las lecturas C y V se mapean directamente a segmentos J o V o que están dentro de la extensión del clonotipo hasta el límite de NDN. De esta manera, la región NDN se evita y la misma 'información de secuencia' no se utiliza para el hallazgo de mutaciones que se utilizó previamente para la determinación del clonotipo (para evitar erróneamente la clasificación como mutaciones nucleótidos que realmente solo son diferentes regiones NDN recombinadas). Para cada tipo de segmento, el segmento mapeado (alelo principal) se utiliza como un armazón y todas las lecturas se considera que tienen mapeado este alelo durante la fase de lectura de mapeo. Cada posición de las secuencias de referencia en las que al menos una lectura se ha mapeado se analiza en cuanto a mutaciones somáticas. En una realización, los criterios para aceptar un base de no referencia como una mutación válida incluyen los siguientes: 1= al menos las lecturas N con la mutación de la base determinada, 2) al menos un fracción determinada de lecturas N/M (donde M es el número total de lecturas mapeadas en esta posición de bases) y 3) un corte estadístico basado en la distribución binómica, la valoración de la media de Q de las lecturas N en la mutación de la base así como el número de lecturas (M-N) con una base no mutada. Preferentemente, los parámetros anteriores se seleccionan de manera que la tasa de falsos descubrimientos por clonotipo sea menor de 1 en 1000, y más preferentemente, menos de 1 en 10000.

Análisis del repertorio de TCRβ

En este ejemplo, se analizan cadenas TCRβ. El análisis incluye la amplificación, secuenciación, y análisis de las secuencias de TCRβ. Un cebador es complementario a una secuencia común Cβ1 y Cβ2, y hay 34 cebadores V capaces de amplificar los 48 segmentos V. Cβ1 o Cβ2 se diferencias entre ellos en la posición 10 y 14 de la unión J/C. El cebador para los extremos Cβ1 y Cβ2 en la posición de 16 pb y no tiene preferencia para Cβ1 o Cβ2. Los 34 cebadores V están modificados a partir de un conjunto original de cebadores desvelados en Van Dongen et al, Publicación de Patente de EE. UU. 2006/0234234. Los cebadores modificados se desvelan en Faham et al, Publicación de Patente de EE. UU. 2010/0151471.

Se utilizó el Analizador de Genoma Illumina para secuenciar el amplicón producido por los cebadores anteriores. Se llevó a cabo una amplificación en dos etapas sobre las transcripciones de ARN (200), como se ilustra en las Fig. 2A-2B, empleando la primera etapa los cebadores anteriores y una segunda etapa para añadir cebadores comunes para la amplificación de puente y la secuenciación. Como se muestra en la FIG. 2A, se lleva a cabo una PCR primaria utilizando en un lado un cebador (202) de 20 pb cuyo extremo 3' es de 16 bases a partir de la unión J/C (204) y que es perfectamente complementario con Cβ1 (203) y los dos alelos de Cβ2. En la región V (206) de las transcripciones de ARN (200), se proporciona el conjunto de cebadores (212) que contiene las secuencias de cebador complementarias con las secuencias de la región V diferentes (34 en una realización). Los cebadores del conjunto (212) también contienen una cola no complementaria (214) que produce el amplicón (216) que tiene el sitio de unión (218) del cebador específico para los cebadores P7 (220). Después de la una PCR múltiple convencional, se forma el amplicón (216) que contiene la parte altamente diversa de la región J(D)V (206, 208, y 210) de las transcripciones de ARNm y los sitios de unión al cebador común (203 y 218) para una amplificación secundara para añadir un marcador de muestra (221) y los cebadores (220 y 222) para la formación de agrupamientos por una PCR de puente. En la PCR secundaria, en el mismo lado de la matriz, se utiliza un cebador (222 en la Fig. 2B y al que se hace referencia en el presente documento como "C10-17-P5") que tiene en su extremo 3' la secuencia de las 10 bases más cercanas a la unión J/C, seguido por las 17 pb con la secuencia de las posiciones 15-31 desde la unión J/C, seguido por la secuencia P5 (224), que tiene un papel en la formación de agrupamiento por la PCR de puente en la secuenciación Solexa. (Cuando el cebador C10-17-P5 (222) se hibrida a la matriz generada por la primera PCR, un bucle de 4 pb (posición 11-14) se crea en la matriz, ya que el cebador se hibrida con la secuencia de las 10 bases más cercanas a la unión J/C y las bases en las posiciones 15-31 de la unión J/C. El bucle de las posiciones 11-14 elimina la amplificación diferencial de matrices que albergan Cβ1 o Cβ2. La secuenciación se hace entonces

con un cebador complementario a la secuencia de las 10 bases más cercanas a la unión J/C y las bases en las posiciones 15-31 desde la unión J/C (este cebador se llama C'). El cebador C10-17-P5 puede purificarse por HPLC con el fin de asegurar que todo el material amplificado tiene los extremos intactos que se pueden utilizar eficazmente en la formación del agrupamiento).

En la FIG. 2A, la longitud de la protuberancia en los cebadores V (212) tiene preferentemente 14 pb. La PCR primaria se auxilia con una protuberancia más corta (214). De manera alternativa, en beneficio de la PCR secundaria, la protuberancia en el cebador V se utiliza en la PCR primaria siempre que sea posible debido a que la PCR secundaria se sensibiliza a partir de esta secuencia. Se investigó un mínimo tamaño de la protuberancia (214) que sustenta una PCR secundaria eficaz. Se produjeron dos series de cebadores V (para dos segmentos V diferentes) con tamaños de protuberancias de 10 a 30 con etapas de 2 pb. Utilizando las secuencias sintéticas apropiadas, se llevó a cabo la primera PCR con cada uno de los cebadores de la serie y se llevó a cabo la electroforesis para demostrar que todos se habían amplificado.

Como se ilustra en la FIG. 2A, la PCR primaria utiliza 34 cebadores V (212) diferentes que se hibridan con la región (206) de las matrices de ARN (200) y que contienen una protuberancia común de 14 pb en la cola 5'. Las 14 pb son la secuencia parcial de uno de los cebadores de secuenciación Illumina (denominado el cebador Read 2). El cebador de amplificación (220) secundario del mismo lado incluye la secuencia de P7, un marcador (221) y la secuencia del cebador Read 2 (223) (este cebador se denomina Read 2 tagX P7). La secuencia P7 se utiliza para la formación del agrupamiento. El cebador Read 2 y su complemento se utilizan para la secuenciación del segmento V y el marcador respectivamente. Se creó un conjunto de 96 de estos cebadores con marcadores que se numeraron del 1 hasta el 96 (véase posteriormente). Estos cebadores se purificaron por HPLC con el fin de asegurar que todo el material amplificado tenía extremos intactos y se pueden utilizar eficazmente en la formación del agrupamiento.

Como se ha mencionado anteriormente, el cebador de segunda etapa, C-10-17-P5 (222, FIG. 2B) tiene una homología interrumpida con la matriz generada por la PCR de la primera etapa. La eficacia de la amplificación utilizando este cebador se había validado. Un cebador alternativo al C-10-17-P5, denominado CsegP5 tiene una homología perfecta con el cebador C de la primera etapa y una cola 5' que alberga la P5. La eficacia de la utilización de C-10-17-P5 y CsegP5 en la aplicación de las matrices de la PCR de primera etapa se comparó llevando a cabo una PCR en tiempo real. En varias repeticiones, se descubrió que la PCR que utiliza el cebador C-10-17-P5 tiene una pequeña diferencia o ninguna de eficacia en comparación con la PCR que utiliza el cebador CsegP5.

El amplicón (230) resultante de la amplificación de 2 etapas que se ilustra en las Fig. 2A-2C tiene la estructura que se utiliza normalmente con el secuenciador Illumina como se muestra en la FIG. 2C. se utilizan dos cebadores que se hibridan con la mayor parte de las moléculas, los cebadores de Illumina P5 y P, para la amplificación en fase sólida de la molécula (formación de agrupamiento). Se hacen tres lecturas de secuencia por molécula. La primera lectura de 100 pb se hace con el cebador C', que tiene una temperatura de fusión que es apropiada para el procedimiento de secuenciación Illumina. La segunda lectura tiene solo 6 pb de longitud y solamente con el fin de identificar el marcador de muestra. Se genera utilizando un cebador marcador proporcionado por el fabricante (Illumina). La lectura final es el cebador Read 2, también proporcionado por el fabricante (Illumina). Utilizando este cebador se genera una lectura de 100 pb en el segmento V comenzando con la secuencia del cebador V de la primera PCR.

Definiciones

A menos que se defina específicamente otra cosa en el presente documento, los términos y símbolos de química, bioquímica, genética y biología molecular de ácidos nucleicos que se utiliza en el presente documento siguen los de los tratados y textos convencionales del campo, por ejemplo, Kornberg y Baker, DNA Replication, Segunda Edición (W.H. Freeman, New York, 1992); Lehninger, Biochemistry, Segunda Edición (Worth Publishers, New York, 1975); Strachan y Read, Human Molecular Genetics, Segunda Edición (Wiley-Liss, New York, 1999); Abbas et al, Cellular and Molecular Immunology, 6ª edición (Saunders, 2007).

"Alineamiento" significa un método para comparar una secuencia de ensayo, tal como una lectura de secuencia, con una o más secuencias de referencia o cuya parte de una secuencia de referencia está basada estrechamente en alguna medición de la 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 los nucleótidos de las secuencias que se van a comparar.

"Amplicón" significa el producto de la reacción de amplificación de un nucleótido; es decir, una población clónica de polinucleótidos, que pueden ser de cadena sencilla o de cadena dobla, que se replican a partir de una o más secuencias de partida. La una o más secuencias de partida pueden ser una o más copias de la misma secuencia, o pueden ser una mezcla de diferentes secuencias. Preferentemente, los amplicones se forman por amplificación de una única secuencia de partida. Los amplicones se pueden producir mediante una variedad de reacciones de amplificación cuyos productos comprenden replicados de uno o más ácidos nucleicos de partida. O dianas. En un aspecto, las reacciones de amplificación que producen los amplicones son "dirigidos por la matriz" en las que el

emparejamiento de bases de los reactantes, sean nucleótidos u oligonucleótidos, tienen complementos en el polinucleótido matriz que se necesita para la creación de los productos de reacción. En un aspecto, las reacciones dirigidas por la matriz son extensiones del cebador con una polimerasa de ácido nucleico o ligadura de oligonucleótidos con una ligasa de ácido nucleico. Dichas reacciones incluyen, pero no se limita a, reacciones en cadena de polimerasa (PCR), reacciones de polimerasa lineal, amplificación basada en la secuencia de ácido nucleico (NASBA), amplificaciones de círculo rodante, y similares, desveladas en las siguientes referencias: Mullis et al, Patentes de EE. UU. 4.683.195; 4.965.188; 4.683.202; 4.800.159 (PCR); Gelfand et al, Patente de EE. UU. 5.210.015 (PCR en tiempo real con sondas “taqman”); Wittwer et al, Patente de EE. UU. 6.174.670; Kacian et al, Patente de EE. UU. 5.399.491 ("NASBA"); Lizardi, Patente de EE. UU. 5.854.033; Aono et al, Publicación de patente Japonesa JP 4-262799 (amplificación en círculo rodante); y similares. Los amplicones se pueden producir por PCR. Una reacción de amplificación puede ser una amplificación en “tiempo real” si hay disponible una detección química que permita que se mida un producto de reacción según progresa la reacción, por ejemplo “PCR en tiempo real” que se describe posteriormente, o “NASBA en tiempo real” como se describe en Leone et al, Nucleic Acids Research, 26: 2150-2155 (1998), y referencias similares. Como se utiliza en el presente documento, el término “amplificación” significa realizar una reacción de amplificación. Una “mezcla de reacción” significa una solución que contiene todos los reactivos necesarios para llevar a cabo una reacción, que puede incluir, pero no se limita a, agentes tampón para mantener el pH a un nivel seleccionado durante la reacción, sales, cofactores, neutralizantes, y similares.

“Clonotipo” significa una secuencia de nucleótidos recombinada de un linfocito que codifica un receptor inmunitario o una parte del mismo. Más particularmente, clonotipo significa una secuencia de nucleótidos recombinada de una célula T o células B que codifica un receptor de célula T (TCR) o un receptor de célula B (BCR), o una parte de los mismos. En distintas realizaciones, los clonotipos pueden codificar todos o una parte de un reordenamiento VDJ de IgH, un reordenamiento DJ de IgH, un reordenamiento de IgK, un reordenamiento VJ de IgL, un reordenamiento VDJ de TCRβ, un reordenamiento DJ de TCRβ, un reordenamiento VJ de TCRα, un reordenamiento VJ de TCRγ, un reordenamiento VDJ de TCRδ, un reordenamiento VD de TCRδ, un reordenamiento Kde-V, o similares. Los clonotipos también pueden codificar regiones de punto de ruptura de translocalización que implican genes del receptor inmunitario, tales como BcII-IgH o BcII.IgH. En un aspecto, los clonotipos tienen secuencias que son suficientemente largas para representar o reflejar la diversidad de las moléculas inmunitarias de las que derivan; en consecuencia, los clonotipos pueden variar de longitud ampliamente. En algunas realizaciones, los clonotipos tienen longitudes en el intervalo de desde 25 a 400 nucleótidos; en otras realizaciones, los clonotipos tienen longitudes en el intervalo de desde 25 a 200 nucleótidos.

“Perfil de clonotipos” significa un listado de distintos clonotipos y sus abundancias relativas que se derivan de una población de linfocitos. Normalmente, la población de linfocitos se obtiene a partir de una muestra de tejido. La expresión “perfil de clonotipos” se relaciona, pero más general que, el concepto inmunológico de “repertorio” inmunitario como se describe en referencias tales como las siguientes: Arstila et al, Science, 286: 958-961 (1999); Yassai et al, Immunogenetics, 61: 493-502 (2009); Kedzierska et al, Mol. Immunol., 45(3): 607-618 (2008); y similares. La expresión “perfil de clonotipos” incluye una amplia variedad de listas y abundancias de ácidos nucleicos que codifican un receptor inmunitario reordenado, que se pueden derivar de subconjuntos de linfocitos seleccionados (por ejemplo, linfocitos que infiltran tejidos, subconjuntos inmunofenotípicos, o similares), o que pueden codificar partes de receptores inmunitarios que tienen una diversidad reducida en comparación con los receptores inmunitarios completos. En algunas realizaciones los perfiles de clonotipos pueden comprender al menos 10^3 clonotipos distintos; en otras realizaciones, los perfiles de clonotipos pueden comprender al menos 10^4 clonotipos distintos; en otras realizaciones, los perfiles de clonotipos pueden comprender al menos 10^5 clonotipos distintos; en otras realizaciones, los perfiles de clonotipos pueden comprender al menos 10^6 clonotipos distintos. En dichas realizaciones, dichos perfiles de clonotipos pueden comprender adicionalmente abundancias o frecuencias relativas de cada uno de los distintos clonotipos. En un aspecto, un perfil de clonotipos es un conjunto de secuencias de nucleótidos recombinadas distintas (con sus abundancias) que codifican receptores de células T (TCR) o de receptores de células B (BCR), o fragmentos de los mismos, respectivamente, en una población de linfocitos de un individuo, en el que las secuencias de nucleótidos del conjunto tienen una correspondencia uno a uno con distintos linfocitos o sus sub-poblaciones clínicas para sustancialmente todos los linfocitos de la población. En un aspecto, los segmentos de ácidos nucleicos que definen los clonotipos se seleccionan de manera que su diversidad (es decir, el número de secuencias de ácido nucleico distintas en el conjunto) es lo suficientemente grande para que sustancialmente cada célula T o célula B o clones de las mismas en un individuo, alberguen una única secuencia de ácido nucleico de dicho repertorio. Es decir, preferentemente cada clon diferente de una muestra tiene un clonotipo diferente. En otros aspectos de la invención, la población de linfocitos que se corresponde con un repertorio puede ser de células B circulantes, o pueden ser células T circulantes, o pueden ser sub-poblaciones de cualquiera de las poblaciones anteriores, incluyendo, pero sin limitarse a células T CD4+, o células T CD8+, u otras sub-poblaciones definidas por marcadores de superficie celular, o similares. Dichas subpoblaciones se pueden adquirir tomando muestras de tejidos particulares, por ejemplo, medula ósea, o ganglios linfáticos, o similares, o clasificando o enriqueciendo células de una muestra (tal como de sangre periférica) basándose en uno o más marcadores de superficie, tamaño, morfología, o similares. En otros aspectos más, la población de linfocitos correspondientes a un repertorio se puede derivar de tejidos enfermos, tales como un tejido tumoral, un tejido infectado, o similares. En una realización, un perfil de clonotipos que comprende cadenas TCRβ humanas o fragmentos de las mismas comprende un número de secuencias de nucleótidos distintos en el intervalo de desde $0,1 \times 10^6$ a $1,8 \times 10^6$, o en el intervalo de desde $0,5 \times 10^6$ a $1,5 \times 10^6$, o en el intervalo de desde $0,8 \times 10^6$ a $1,2 \times 10^6$. En otra realización, un perfil de

clonotipos que comprende cadenas de IgH humana o fragmentos de las mismas comprende un número de secuencias de nucleótidos distintos en el intervalo de desde 0,1 x 10^6 a 1,8 x 10^6, o en el intervalo de desde 0,5 x 10^6 a 1,5 x 10^6, o en el intervalo de desde 0,8 x 10^6 a 1,2 x 10^6. En una realización particular, un perfil de clonotipos comprende un conjunto de secuencias de nucleótidos que codifica sustancialmente todos los segmentos de la región V(D)J de una cadena de IgH. En un aspecto, “sustancialmente todos" como se utiliza en el presente documento significa cada segmento que tiene una abundancia relativa del 0,001 por ciento o mayor; o en otro aspecto, “sustancialmente todos “ como se utiliza en el presente documento significa cada segmento que tiene una abundancia relativa del 0,0001 por ciento o más. En otra realización particular, un perfil de clonotipos 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 TCRβ. En otra realización, un perfil de clonotipos de la invención comprende un conjunto de secuencias de nucleótidos que tienen longitudes en el intervalo de desde 25-200 nucleótidos e incluyen los segmentos de las regiones V, D, y J de una cadena TCRβ. En otra realización, un perfil de clonotipos de la invención comprende un conjunto de secuencias de nucleótidos que tienen longitudes en el intervalo de desde 25-200 nucleótidos e incluyen los segmentos de las regiones V, D, y J de una cadena de IgH. En otra realización, un perfil de clonotipos de la invención comprende un número de secuencias de nucleótidos distintos que es sustancialmente equivalente al número de linfocitos que expresan una cadena de IgH distinta. En otra realización, un perfil de clonotipos de la invención comprende un número de distintas secuencias de nucleótidos que es sustancialmente equivalente al número de linfocitos que expresan una cadena de TCRβ distinta. En otra realización más, “sustancialmente equivalente” significa que un perfil de clonotipos incluirá con un noventa y nueve por ciento de probabilidad una secuencia de nucleótidos que codifique una IgH o TCRβ o parte de la misma albergada o expresada por cada linfocito de una población de un individuo con una frecuencia de 0,001 por ciento o mayor. En otra realización más, “sustancialmente equivalente” significa que un repertorio de secuencias de nucleótidos incluirá con un noventa y nueve por ciento de probabilidad una secuencia de nucleótidos que codifique una IgH o TCRβ o parte de la misma albergada o expresada por cada linfocito presente con una frecuencia de un 0,0001 por ciento o mayor. En algunas realizaciones, los perfiles de clonotipos se derivan de muestras que comprenden de 10^5 a 10^7 linfocitos. Dicha cantidad de linfocitos puede obtenerse de muestras de sangre periférica de 1-10 ml.

“Regiones determinantes de complementariedad” (CDR) significa regiones de una inmunoglobulina (es decir, un anticuerpo) o un receptor de célula T en las que la molécula se complementa con una conformación de antígeno, determinando de esta manera la especificidad de la molécula y el contacto con un antígeno específico. Los receptores de células T y las inmunoglobulinas tienen cada uno tres CDR: CDR1 y CDR2 que se encuentran en el domino variable (V), y la CDR3 que incluye algo del dominio V, todo el diverso (D) (solo las cadenas pesadas) y la unión (J), y algo del constante (C).

“Por ciento homólogo”, “por ciento idéntico”, o expresiones similares que se utilizan en referencia a la comparación de una secuencia de referencia y otra secuencia (“secuencia de comparación”) significa que en un alineamiento óptimo entre las dos secuencias, la secuencia de comparación es idéntica a la secuencia de referencia en un número posiciones de subunidades equivalente al porcentaje indicado, siendo las subunidades nucleótidos para las comparaciones de polinucleótidos o aminoácidos para las comparaciones de polipéptidos. Como se utiliza en el presente documento, un “alineamiento óptimo” de secuencias que se comparan es el que maximiza las coincidencias entre subunidades y minimiza el número de huecos empleado en la construcción del alineamiento. El porcentaje de identidades se puede determinar con implementaciones de algoritmos disponibles en el mercado, tales como los descritos por Needleman y Wunsch, J. Mol. Biol., 48: 443-453 (1970)(programa "GAP" del Paquete de Análisis de Secuencia Wisconsin, Genetics Computer Group, Madison, WI), o similares. Otros paquetes de software en la técnica para construir alineamientos y calcular el porcentaje de identidad u otras mediciones de similitud incluyen el programa “BestFit”, basado en el algoritmo de Smith y Waterman, Advances in Applied Mathematics, 2: 482-489 (1981) (Paquete de Análisis de Secuencia Wisconsin, Genetics Computer Group, Madison, WI). En otras palabras, por ejemplo, para obtener un polinucleótido que tiene una secuencia de nucleótidos al menos un 95 por ciento idéntica a una secuencia de nucleótidos de referencia, se pueden borrar o sustituir hasta un cinco por ciento de los nucleótidos en la secuencia de referencia con otros nucleótidos, o un número de nucleótidos hasta un 5 por ciento del número total de nucleótidos en la secuencia de referencia se pueden insertar en la secuencia de referencia.

“Reacción en cadena de la polimerasa”, o “PCR”, significa una reacción para la amplificación *in vitro* de secuencias de ADN específicas mediante la extensión simultánea de un cebador de cadenas complementarias de ADN. En otras palabras, la PCR es una reacción para fabricar múltiples copias o replicados de un ácido nucleico diana flanqueado por los sitios de unión al cebador, dicha reacción comprende una o más repeticiones de las siguientes etapas: (i) desnaturalización del ácido nucleico diana, (ii) hibridación de los cebadores a los sitios de unión al cebador, y (iii) extensión de los cebadores por una polimerasa de ácido nucleico en presencia de trifosfatos de nucleótido. Habitualmente, la reacción se cicla a través de diferentes temperaturas optimizadas para cada etapa en un instrumento ciclador térmico. Las temperaturas en particular, las duraciones de cada etapa, y las tasas de cambio entre las etapas dependen de muchos factores bien conocidos por los expertos en la técnica, por ejemplo, como se ejemplifica en las referencias: McPherson et al, editores, PCR: A Practical Approach and PCR2: A Practical Approach (IRL Press, Oxford, 1991 y 1995, respectivamente). Por ejemplo, en una PCR convencional utilizando un ADN polimerasa Taq, se puede desnaturalizar un ácido nucleico diana de doble cadena a una temperatura > 90 ºC, los cebadores se hibridan a una temperatura en el intervalo de 50-75 ºC, y los cebadores se extienden a una

temperatura en el intervalo de 72-78 ºC. El término “PCR” engloba formas derivadas de la reacción, incluyendo, pero sin limitarse a, RT-PCR, PCR en tiempo real, PCR anudada, PCR cuantitativa, PCR multiplexada, y similares. Los volúmenes de reacción varían desde uno cuantos cientos de nanolitros, por ejemplo, 200 nl, a unos pocos cientos de ml, por ejemplo 200 ml. “PCR de transcripción inversa” o “RT-PCR”, significa una PCR que está precedida por una reacción de transcripción inversa que convierte un ARN diana en un ADN complementario de cadena sencilla, que entonces se amplifica, por ejemplo, Tecott et al, Patente de EE. UU. 5.168.038. “PCR en tiempo real” significa una PCR para la que la cantidad del producto de reacción, es decir, el amplicón, se controla según progresa la reacción. Ha y muchas formas de PCR en tiempo real que se diferencian principalmente en las químicas de detección que se utiliza para controlar el producto de reacción, por ejemplo, Gelfand et al, Patente de EE. UU. 5.210.015 ("taqman"); Wittwer et al, Patentes de EE. UU. 6.174.670 y 6.569.627 (colorantes de intercalación); Tyagi et al, Patente de EE. UU. 5.925.517 (balizas moleculares). La química de detección para la PCR en tiempo real se ha revisado en Mackay et al, Nucleic Acids Research, 30: 1292-1305 (2002). “PCR anidada” significa una PCR en dos etapas en la que el amplicón de una primera PCR se convierte en la muestra para una segunda PCR utilizando un nuevo conjunto de cebadores, al menos uno de los cuales se une a una localización interior del primer amplicón. Como se utiliza en el presente documento, cebadores iniciales” en referencia a una reacción de amplificación anidada significa que los cebadores utilizados para generar un primer amplicón, y “cebadores secundarios” significa el uno o más cebadores utilizados para generar un segundo amplicón, o anidado.

“PCR multiplexada” significa una PCR en la que se llevan a cabo múltiples secuencias diana (o una única secuencia diana y una o más secuencias de referencia) simultáneamente en la misma mezcla de reacción, por ejemplo, Bernard et al, Anal. Biochem., 273: 221-228 (1999) (PCR en tiempo real de dos colores). Habitualmente, se emplean distintos conjuntos de cebadores para cada secuencia que se va a amplificar. Normalmente, el número de secuencias diana en una PCR múltiple está en el intervalo de desde 2 a 50, o desde 2 a 40, o desde 2 a 30. “PCR cuantitativa” significa una PCR diseñada para medir la abundancia de una o más secuencias diana en una muestra o espécimen. La PCR cuantitativa incluye tanto la cuantificación absoluta como la cuantificación relativa de dichas secuencias diana. Las mediciones cuantitativas se hacen utilizando una o más secuencias de referencia o referencias internas que se pueden ensayar por separado o en conjunto con la secuencia diana. La secuencia de referencia puede ser endógena o exógena a una muestra o espécimen, y en el último caso, puede comprender una o más matrices competidoras. Las secuencias de referencia endógenas típicas incluyen segmentos de transcripciones de los siguientes genes: β-actina, GAP-DH, β-microglobulina, ARN ribosómico, y similares. Las técnicas para la PCR cuantitativa son bien conocidas por los expertos habituados en la técnica, como se ejemplifica en las siguientes referencias: Freeman et al, Biotechniques, 26: 112-126 (1999); Becker-Andre et al, Nucleic Acids Research, 17: 9437-9447 (1989); Zimmerman et al, Biotechniques, 21: 268-279 (1996); Diviacco et al, Gene, 122: 3013-3020 (1992); Becker-Andre et al, Nucleic Acids Research, 17: 9437-9446 (1989); y similares.

“Cebador” dignifica un oligonucleótido, sea natural o sintético, que es capaz, al formar un dúplex con una matriz de polinucleótido de actuar como un punto de inicio de la síntesis de ácido nucleico y de extenderse a partir de su extremo 3’ a lo largo de la matriz, de manera que se forma un dúplex extendido. La extensión de un cebador se lleva a cabo habitualmente con una polimerasa de ácido nucleico, tal como una ADN o ARN polimerasa. La secuencia de nucleótidos añadida en el procedimiento de extensión se determina por la secuencia de la matriz de polinucleótido. Habitualmente, los cebadores se extienden mediante una ADN polimerasa. Los cebadores habitualmente tienen una longitud en el intervalo de desde 18 a 36 nucleótidos. Los cebadores se emplean en una variedad de reacciones de amplificación nucleica, por ejemplo, reacciones de amplificación lineal utilizando un cebador único, o reacciones en cadena de polimerasa, que emplean dos o más cebadores. Las directrices para la selección de las longitudes y secuencias de lo cebadores para las aplicaciones en particular son bien conocidas por los expertos habituados en la técnica, como se prueba por las siguientes referencias: Dieffenbach, editor, PCR Primer: A Laboratory Manual, 2ª Edición (Cold Spring Harbor Press, New York, 2003).

“Valoración de calidad” significa una medición de la probabilidad de que una asignación de base en una localización de secuencia particular sea correcta. Una variedad de métodos es bien conocida por los expertos habituados en la técnica para calcular las valoraciones de calidad para circunstancias particulares, tales como, para bases identificadas como resultado de las diferentes reacciones químicas de secuenciación, sistemas de detección, algoritmos de identificación de bases, y demás. En general, los valores de la valoración de calidad están relacionados monótonamente con las probabilidades de la identificación de bases correcta. Por ejemplo, una valoración de calidad, o Q, de 10 puede significar que hay un 90 por ciento de probabilidad que una base se identifique correctamente, una Q de 20 puede significar que hay un 99 por ciento de probabilidad de que la base se identifique correctamente, y así. Para algunas plataformas de secuenciación, particularmente las que utilizan reacciones químicas de secuenciación-por-síntesis, la media de valoraciones de calidad disminuye en función a la longitud de la lectura de secuencia, de manera que las valoraciones de calidad al principio de una lectura de secuencia es mayor que las del final de la lectura de secuencia, dicha disminución se debe a fenómenos tales como extensiones incompletas, extensiones que se llevan directamente, pérdida de matriz, pérdida de polimerasa, fallos de protección, fallos de desprotección, y similares.

“Lectura de secuencia” significa una secuencia de nucleótidos determinada a partir de una secuencia o corriente de datos generados por una técnica de secuenciación, cuya determinación se hace, por ejemplo, por medio de un software de identificación de bases asociado con la técnica, por ejemplo, el software de identificación de bases a partir de un proveedor comercial de una plataforma de secuenciación de ADN. Una lectura de secuencia incluye

habitualmente valores de calidad para cada nucleótido de la secuencia. Normalmente, las lecturas de secuencia se hacen extendiendo un cebador a lo largo de una matriz de ácido nucleico, por ejemplo, con una ADN polimerasa o una ADN ligasa. Los datos se generan registrando señales, tales como ópticas, químicas (por ejemplo, un cambio de pH), o señales eléctricas, asociadas con dicha extensión. Dichos datos iniciales se convierten en una lectura de secuencia.

REIVINDICACIONES

1. Un método de determinación de cadenas de receptor inmunitario emparejadas en una muestra, comprendiendo el método las etapas de:

(a) dividir una muestra que contiene linfocitos que expresan parejas de cadenas de receptor inmunitario en una pluralidad de subconjuntos;
(b) determinar las secuencias de nucleótidos de una primera cadena de cada pareja de cadenas de receptor inmunitario de linfocitos que tienen dichas parejas en una porción de la pluralidad de subconjuntos;
(c) determinar las secuencias de nucleótidos de una segunda cadena de cada pareja de cadenas de receptor inmunitario de linfocitos que tienen dichas parejas de la misma porción de la pluralidad de subconjuntos;
(d) identificar como cadenas de receptor inmunitario emparejadas las parejas de primeras cadenas y segundas cadenas (i) que, para cada subconjunto de la porción, o existen juntas o no existen y (ii) que existen juntas en al menos un subconjunto de la porción y no existen en al menos un subconjunto de la porción.

2. El método de la reivindicación 1 que comprende adicionalmente la etapa de repetir dichas etapas (a)-(d) para otra pluralidad de subconjuntos diferente de cualquier pluralidad anterior hasta que se obtenga un número deseado de dichos receptores inmunitarios emparejados.

3. El método de la reivindicación 1, donde dicha pluralidad es mayor de 100 y dicha porción de dicha pluralidad está en el intervalo de desde 10 a 100.

4. El método de la reivindicación 1, donde dichas cadenas de receptor inmunitario son cadenas α de receptor de célula T y cadenas β de receptor de células T.

5. El método de la reivindicación 1, donde dichas cadenas de receptor inmunitario son cadenas γ de receptor de célula T y cadenas δ de receptor de células T.

6. El método de la reivindicación 1, donde dichas cadenas de receptor inmunitario son regiones variables de cadena pesada del receptor de célula B y regiones variables de cadena ligera de receptor de célula B.

7. El método de la reivindicación 1, donde dicha muestra contiene una población de dichos linfocitos que expresan dichas parejas de cadenas de receptor inmunitario, y donde la población tiene un tamaño y cada linfocito diferente de la población tiene una frecuencia en la población.

8. El método de la reivindicación 7, donde dicha pluralidad de subconjuntos depende de dicho tamaño de dicha población y dicha frecuencia de dichos linfocitos cuyas cadenas de receptor inmunitario emparejadas se van a determinar.

9. El método de la reivindicación 1, donde un tamaño de dicha muestra y dicha pluralidad de subconjuntos se seleccionan de manera que dichos linfocitos de dicha muestra se distribuyen entre dichos subconjuntos de acuerdo con un modelo binómico.

10. El método de una cualquiera de las reivindicaciones 1 a 9, donde la muestra contiene células T.

11. El método de una cualquiera de las reivindicaciones 1 a 10, donde la muestra contiene células B.

12. El método de una cualquiera de las reivindicaciones 1 a 11, donde las cadenas de receptor inmunitario identificadas en la etapa (d) forman un perfil de clonotipos, y donde cada clonotipo del perfil de clonotipos es una pareja de secuencias de nucleótidos que codifica una pareja de cadenas de TCR que se expresan en la misma célula T o donde cada clonotipo del perfil de clonotipos es una pareja de secuencias de nucleótidos que codifica una cadena pesada y ligera de inmunoglobulinas que se expresa en la misma célula B.

13. El método de la reivindicación 12, donde el perfil de clonotipos comprende al menos 100 clonotipos, y donde cada secuencia de nucleótidos de cada clonotipo comprende una secuencia de desde 30 a 300 nucleótidos.

14. El método de la reivindicación 12, donde el perfil de clonotipos comprende al menos 500 clonotipos, y donde cada secuencia de nucleótidos de cada clonotipo comprende una secuencia de 30 a 300 nucleótidos.

15. El método de la reivindicación 12, donde el perfil de clonotipos comprende al menos 1000 clonotipos, y donde cada secuencia de nucleótidos de cada clonotipo comprende una secuencia de 30 a 300 nucleótidos.

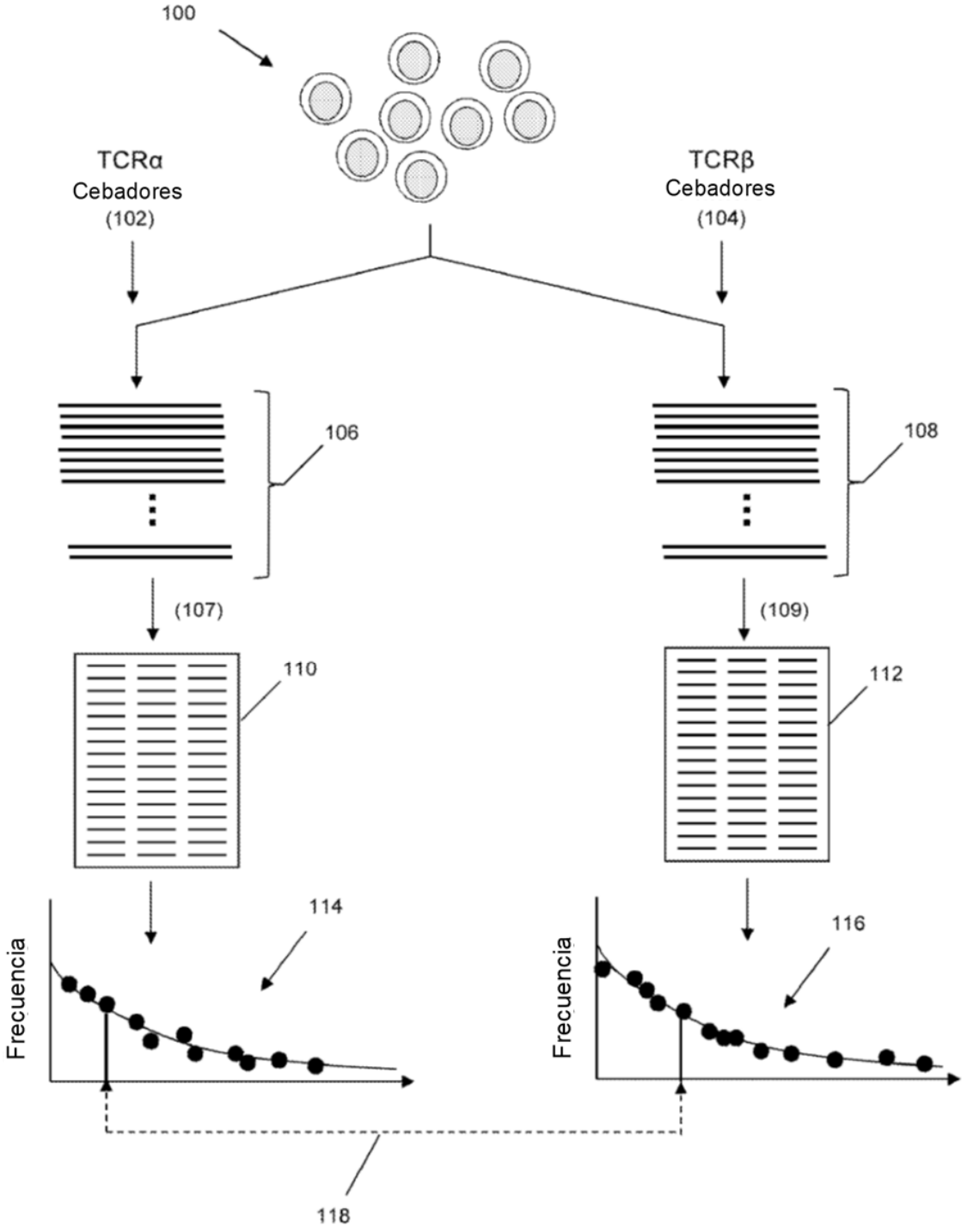
100
TCRα
Cebadores
(102)
TCRβ
Cebadores
(104)
106
108
(107)
(109)
110
112
114
116
Frecuencia
Frecuencia
118

Fig. 1A

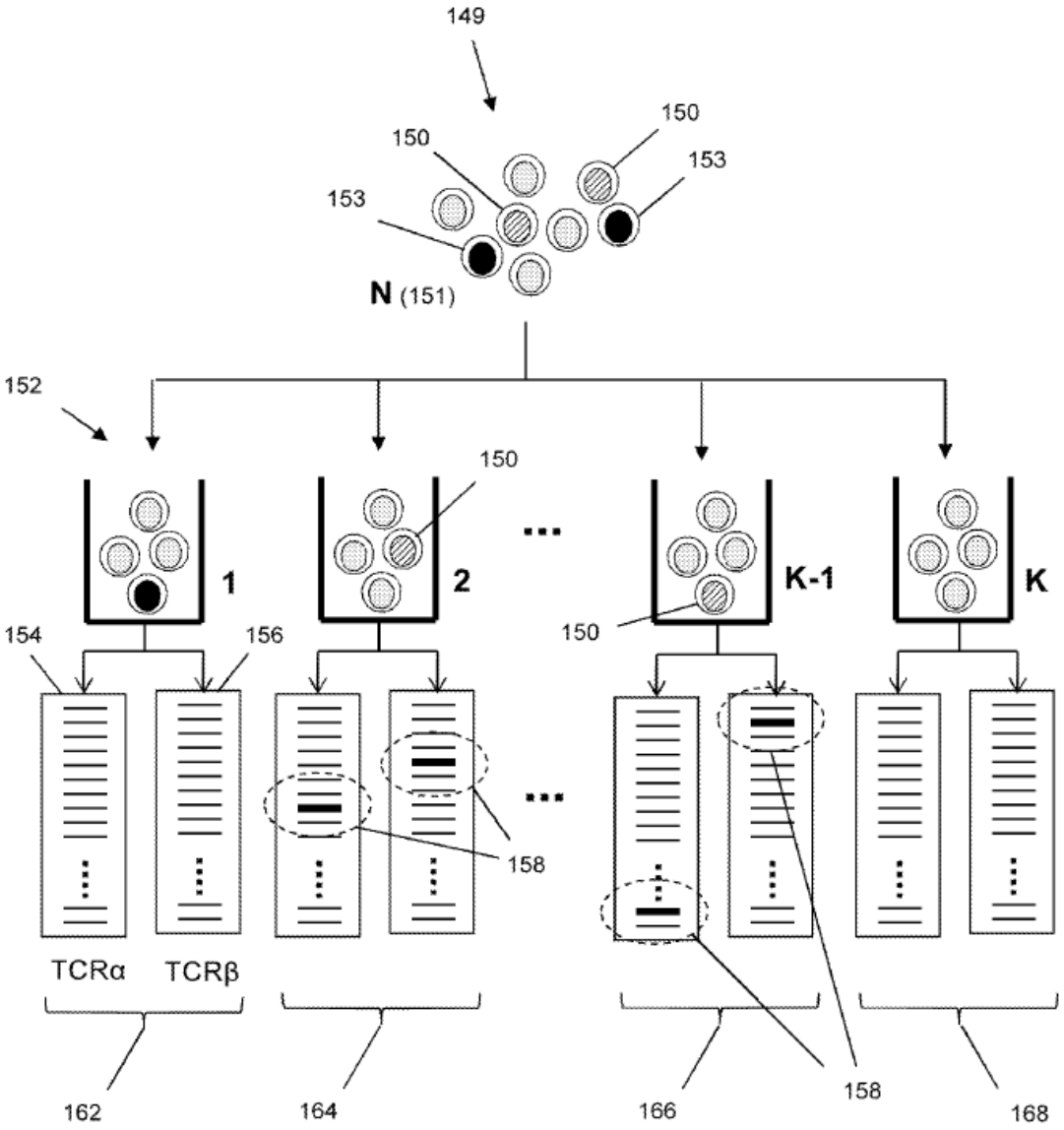
149
150
150
153
153
N (151)
152
150
1
2
K-1
K
154
156
150
158
158
TCRα
TCRβ
162
164
166
168

Fig. 1B

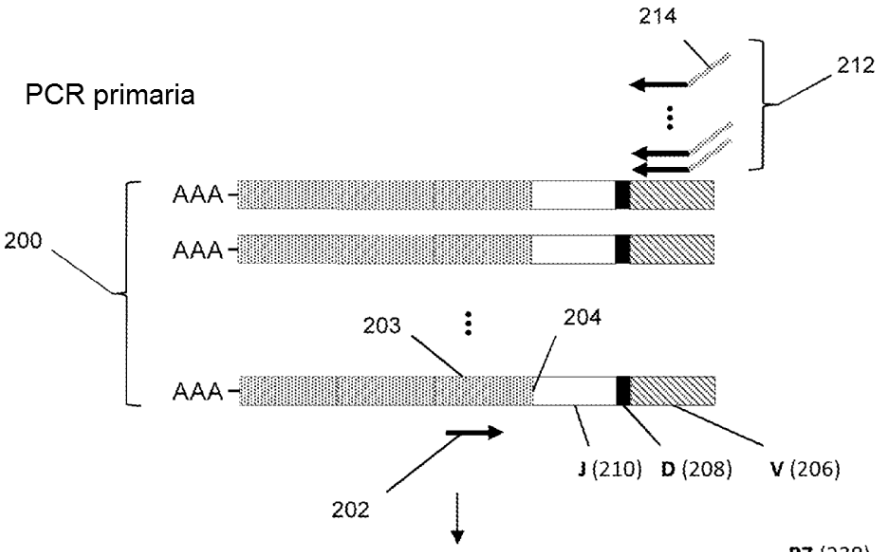

Fig. 2A

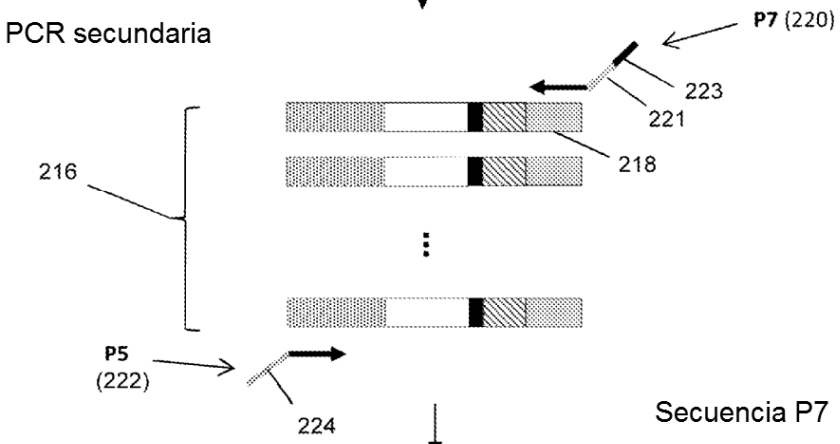

Fig. 2B

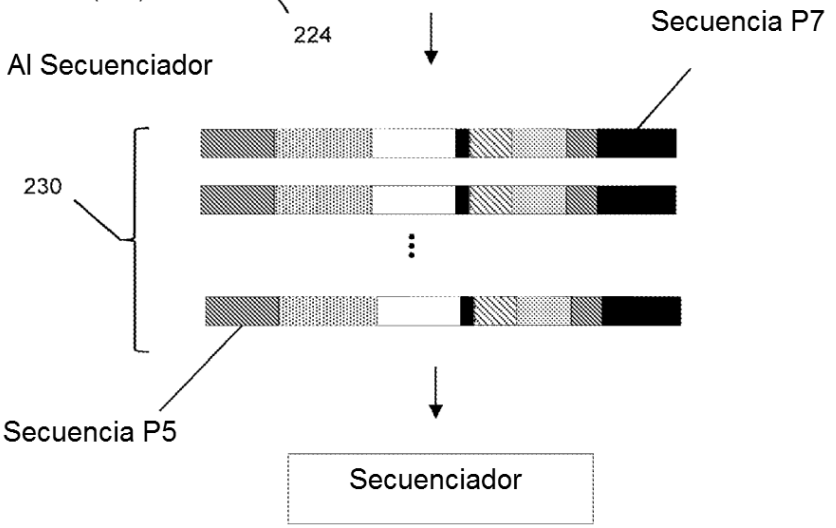

Fig. 2C

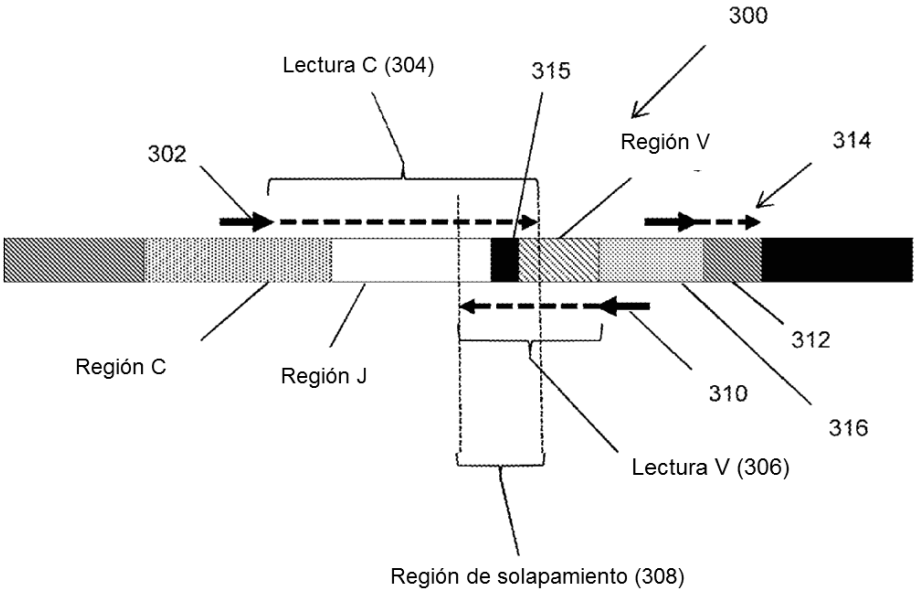
300
Lectura C (304)
315
Región V
314
302
Región C
Región J
312
310
316
Lectura V (306)
Región de solapamiento (308)

Fig. 3A

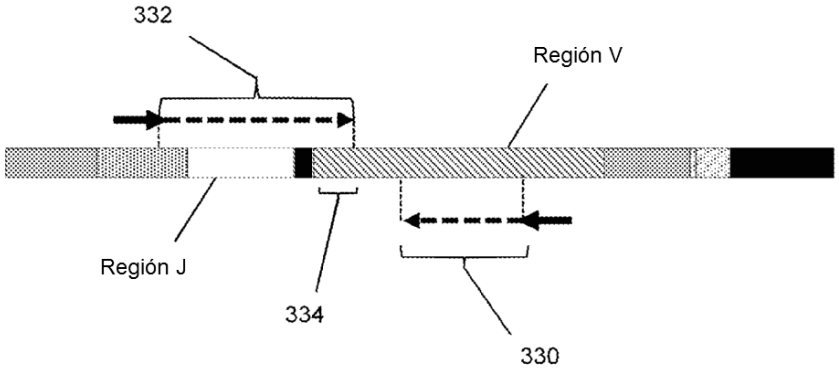
332
Región V
Región J
334
330

Fig. 3B

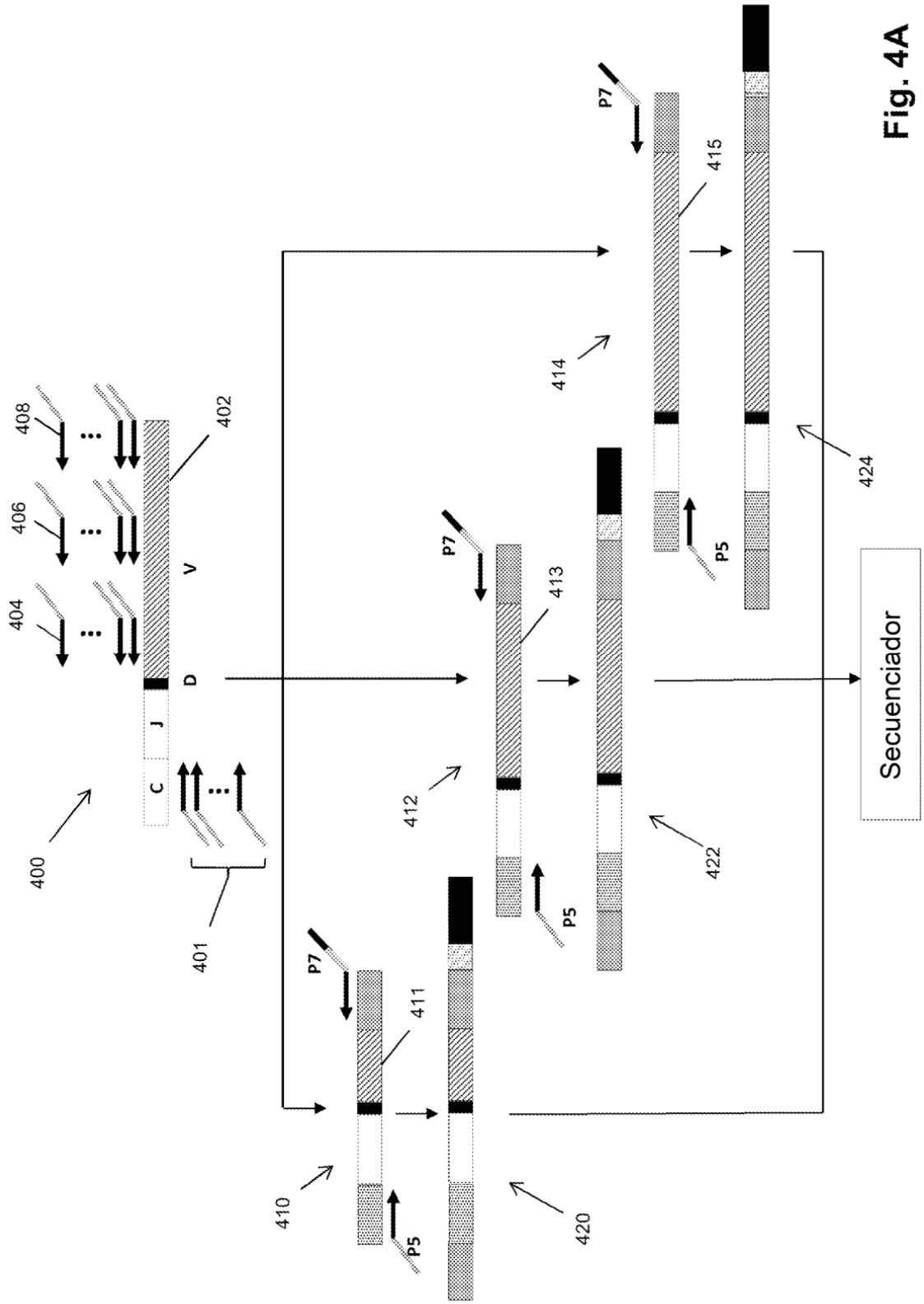
Secuenciador
P7
P5
V
D
J
C
400
401
402
404
406
408
410
411
412
413
414
415
420
422
424

Fig. 4A

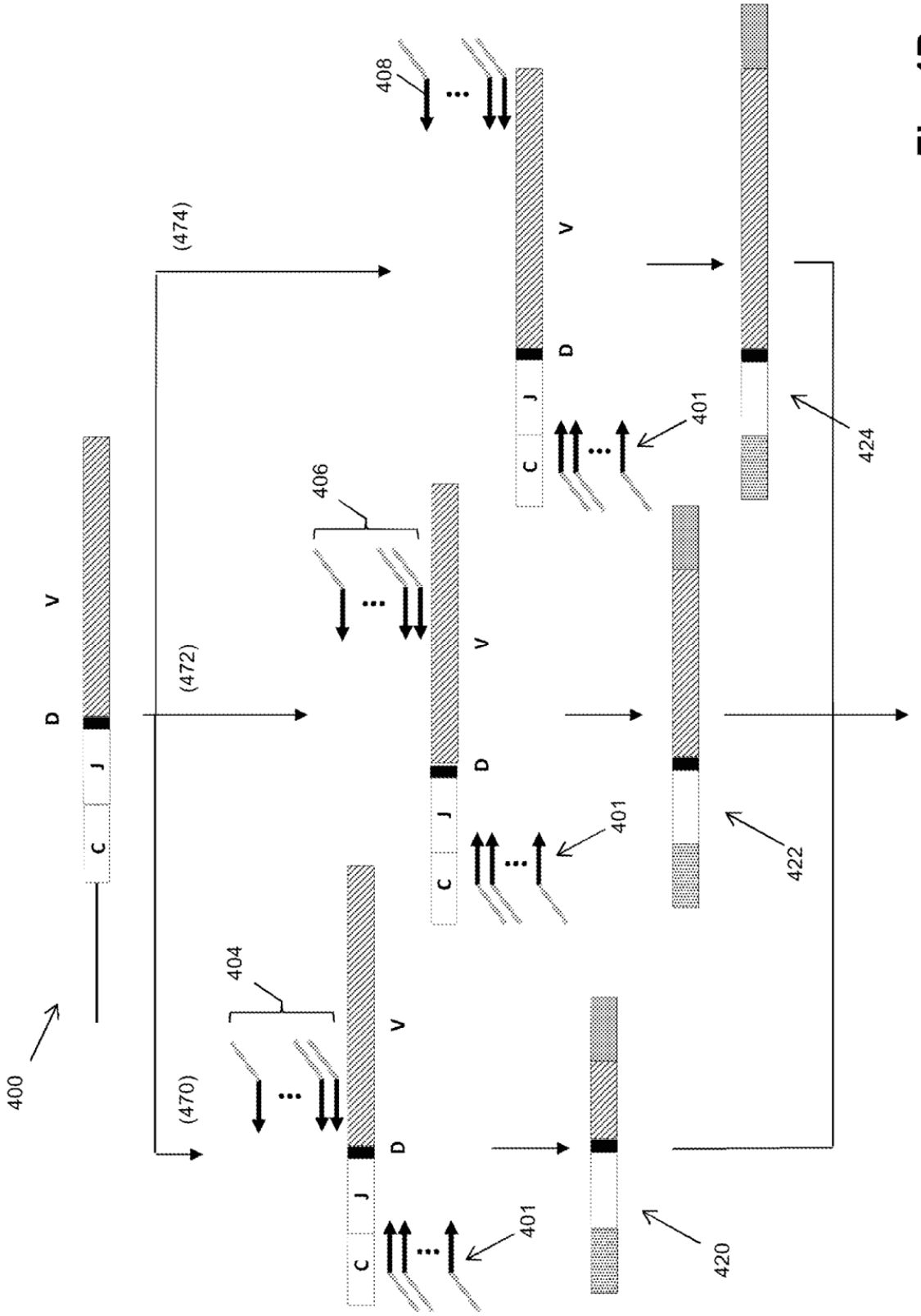

Fig. 4B

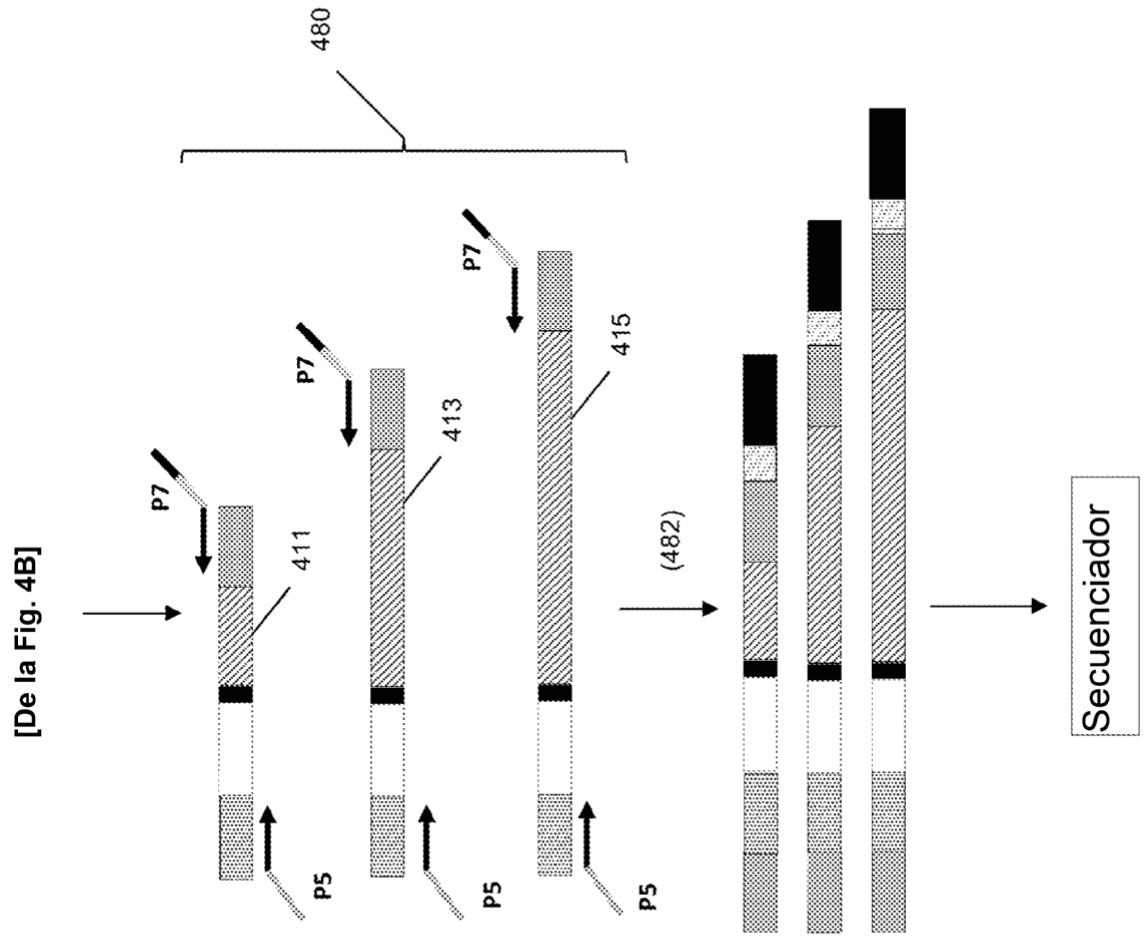

Fig. 4C

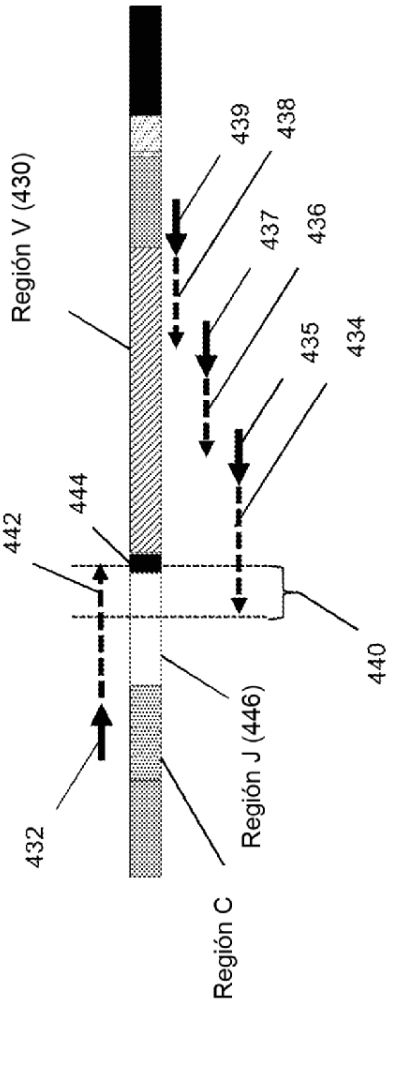
Región V (430)
439
438
437
436
435
434
442
444
440
Región J (446)
432
Región C

Fig. 4D

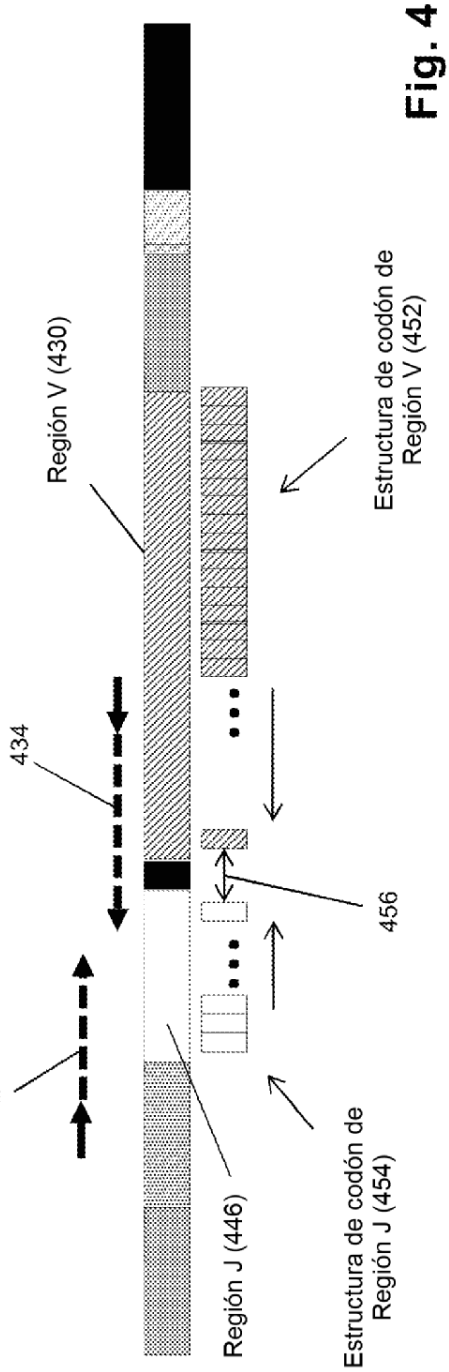
Región V (430)
Estructura de codón de Región V (452)
434
456
442
Región J (446)
Estructura de codón de Región J (454)
450

Fig. 4E