(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 748 204**

(51) Int. CI.:

C12Q 1/6809 (2008.01)
C12Q 1/6858 (2008.01)
C12Q 1/6869 (2008.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(86) Fecha de presentación y número de la solicitud internacional: **20.02.2014 PCT/US2014/017416**

(87) Fecha y número de publicación internacional: **28.08.2014 WO14130685**

(96) Fecha de presentación y número de la solicitud europea: **20.02.2014 E 14754622 (0)**

(97) Fecha y número de publicación de la concesión europea: **18.09.2019 EP 2959020**

(54) Título: **Procedimiento para seleccionar clonotipos raros**

(30) Prioridad:

22.02.2013 US 201361768269 P
15.03.2013 US 201313834794

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
13.03.2020

(73) Titular/es:

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

(72) Inventor/es:

PEPIN, FRANCOIS;
FAHAM, MALEK y
MOORHEAD, MARTIN

(74) Agente/Representante:

UNGRÍA LÓPEZ, Javier

DESCRIPCIÓN

Procedimiento para seleccionar clonotipos raros

Antecedentes de la invención

La secuenciación de ADN a gran escala en aplicaciones diagnósticas y pronósticas se ha expandido rápidamente ya que su velocidad ha aumentado y su coste por base ha disminuido, por ejemplo, Ding y col., Nature, 481(7382): 506-510 (2012); Chiu y col., Brit. Med. J., 342: c7401 (2011); Ku y col., Annals of Neurology, 71(1): 5-14 (2012); y similares. En particular, perfiles de ácidos nucleicos que codifican moléculas inmunitarias, tales como receptores de linfocitos T o linfocitos B, o sus componentes, contienen una rica información sobre el estado de salud o enfermedad de un organismo, de modo que se ha propuesto el uso de tales perfiles como indicadores diagnósticos o pronósticas para una amplia variedad de afecciones, por ejemplo, Faham y Willis, publicación de la patente de Estados Unidos 2010/0151471; Freeman y col., Genome Research, 19: 1817-1824 (2009); Boyd y col, Sci. Transl. Med., 1(12): 12ra23 (2009); He y col, Oncotarget (March 8, 2011).

Los pacientes tratados para muchos cánceres a menudo retienen una enfermedad mínima residual (EMR) relacionada con el cáncer. Es decir, aunque un paciente pueda tener mediante medidas clínicas una remisión completa de la enfermedad en respuesta al tratamiento, puede permanecer una pequeña fracción de las células cancerígenas que, por una razón u otra, haya escapado de su destrucción. El tipo y tamaño de esta población residual es un factor pronóstico importante para el tratamiento continuado del paciente, por ejemplo, Campana, Hematol. Oncol. Clin. North Am., 23(5): 1083-1098 (2009); Buccisano y col., Blood, 119(2): 332-341 (2012). En consecuencia, se han desarrollado varias técnicas para evaluar esta población, incluidas técnicas basados en citometría de flujo, hibridación *in situ*, citogenética, ampliación de marcadores de ácidos nucleicos y similares, por ejemplo, Buccisano y col., Current Opinion in Oncology, 21: 582-588 (2009); van Dongen y col., Leukemia, 17(12): 2257-2317 (2003); y similares. La amplificación de los ácidos nucleicos que codifican segmentos de receptores inmunes recombinados (es decir, clonotipos) han sido particularmente útiles para evaluar la EMR en leucemias y linfomas, puesto que tales clonotipos tienen normalmente secuencias únicas que pueden servir como marcadores moleculares para sus células cancerosas asociadas. Sin embargo, no todos los clonotipos codifican segmentos receptores muy diversos, tales como segmentos V(D)J. Los clonotipos codifican, no con poca frecuencia, segmentos receptores de menor diversidad, tales como segmentos DJ o segmentos que forman motivos recurrentes, preferentemente representados por posibles razones de desarrollo o funcionales y, por lo tanto, relativamente comunes entre individuos diferentes, por ejemplo, Gauss y col., Mol. Cell. Biol., 16(1): 258-269 (1996); Murugan y col., Proc. Natl. Acad. Sci., 109(40): 16161-16166 (2012); Venturi y col., J. Immunol., 186: 4285-4294 (2011); Robins, J. Immunol., 189(6): 3221-3230 (2012). En cualquier circunstancia, tales clonotipos pueden ser subóptimos, o incluso fallar, como marcadores para evaluar la EMR.

A la vista de lo anterior, sería muy ventajoso si hubiera un procedimiento disponible para evaluar los clonotipos en busca de rarezas o singularidades para determinar si es probable que proporcionen una medida precisa de enfermedad residual mínima, con baja probabilidad de resultados falsos positivos.

Sumario de la invención

La presente invención se refiere a un procedimiento de selección de clonotipos raros para controlar la enfermedad residual mínima de un cáncer con el fin de minimizar los resultados falsos positivos. La invención se define por el conjunto de reivindicaciones adjuntas.

En el presente documento se desvela un procedimiento para seleccionar uno o más clonotipos específicos del paciente correlacionados con un trastorno proliferativo linfoide para controlar una enfermedad residual mínima del mismo, comprendiendo el procedimiento las etapas siguientes: (a) obtener una muestra de diagnóstico del paciente que comprende linfocitos T y/o linfocitos B; (b) amplificar moléculas de ácido nucleico de los linfocitos T y/o linfocitos B de la muestra, comprendiendo las moléculas de ácido nucleico secuencias de ADN recombinadas de genes del receptor de linfocitos T o genes de inmunoglobulina; (c) secuenciar las moléculas amplificadas de ácido nucleico para formar un perfil de clonotipo; (d) seleccionar un conjunto de uno o más clonotipos candidatos en función de su frecuencia en el perfil de clonotipo; (e) comparar el uno o más clonotipos candidatos del conjunto a los clonotipos de una base de datos de clonotipos que contiene clonotipos de al menos un individuo que no sea el paciente para determinar una presencia, ausencia y/o nivel en la base de datos de clonotipos de cada clonotipo candidato y eliminar del conjunto cualquier clonotipo candidato que tenga una frecuencia en la base de datos de clonotipo mayor que una frecuencia predeterminada; y (f) seleccionar como uno o más clonotipos específicos del paciente para monitorizar la enfermedad residual mínima los clonotipos candidatos que permanecen en el conjunto.

Además, se desvela un procedimiento para seleccionar uno o más clonotipos específicos del paciente correlacionados con una neoplasia linfoide para controlar una enfermedad residual mínima de la misma, en el que el procedimiento comprende las etapas de: (a) obtener una muestra del paciente que comprende linfocitos T y/o linfocitos B; (b) amplificar moléculas de ácido nucleico de los linfocitos T y/o linfocitos B de la muestra, comprendiendo las moléculas de ácido nucleico secuencias de ADN recombinadas de genes del receptor de

linfocitos T o genes de inmunoglobulina; (c) secuenciar las moléculas amplificadas de ácido nucleico para formar un perfil de clonotipo; (d) comparar los clonotipos del perfil de clonotipos con los clonotipos de una base de datos de clonotipos que contiene clonotipos de al menos un individuo que no sea el paciente para determinar una presencia, ausencia y/o nivel en la base de datos de clonotipos de cada clonotipo del perfil de clonotipo; y (e) seleccionar e uno o más clonotipos específicos del paciente para monitorizar la enfermedad residual mínima que están correlacionados con la neoplasia linfoide y que están ausentes de la base de datos de clonotipos o en un nivel en la base de datos de clonotipos por debajo de una frecuencia predeterminada.

Además, se desvela un procedimiento para determinar si una muestra de tejido que comprende linfocitos T y/o linfocitos B de un primer individuo contamina una muestra de tejido que comprende linfocitos T y/o linfocitos B de un segundo individuo, en el que el procedimiento comprende las etapas de: (a) generar un primer perfil de clonotipo a partir de ácido nucleico de la muestra de tejido del primer individuo; (b) generar un segundo perfil de clonotipo a partir de ácido nucleico de la muestra de tejido del segundo individuo; (c) comparar clonotipos del primer y segundo perfil de clonotipos con clonotipos de una base de datos de clonotipos que contiene clonotipos de al menos un individuo que no sea el primer y el segundo individuo para determinar una presencia, ausencia y/o nivel en la base de datos de clonotipo de cada clonotipo de los perfiles de clonotipo primero y segundo que están ausentes de la base de datos de clonotipo o A un nivel en la base de datos de clonotipo por debajo de un umbral predeterminado; y (d) clasificar la muestra de tejido del primer individuo como contaminada por ácidos nucleicos del segundo individuo siempre que cualquiera de dichos clonotipos determinados esté presente tanto en el primer como en el segundo perfil de clonotipo.

Breve descripción de los dibujos

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

Las figuras 1A-1C muestran un esquema de PCR de dos etapas para amplificar y secuenciar genes de IgH o TCRβ.
Las figuras 2A-2B ilustran diferentes realizaciones para determinar un clonotipo basado en lecturas de secuencia de un amplicón producido por el procedimiento ilustrado en las figuras 1A-1C.
La figura 3A ilustra un esquema de PCR para generar tres moldes de secuenciación a partir de una cadena de IgH en una sola reacción. Las figuras 3B-3C ilustran un esquema de PCR para generar tres moldes de secuenciación a partir de una cadena de IgH en tres reacciones separadas después de las cuales los amplicones resultantes se combinan para una PCR secundaria para añadir sitios de unión de los cebadores P5 y P7. La figura 3D ilustra las ubicaciones de las lecturas de secuencia generadas para una cadena de IgH.
Las figuras 4A-4B ilustran un modelo de frecuencia de clonotipos que puede emplearse para resumir información en una base de datos de clonotipos útil para seleccionar clonotipos raros.

Descripción detallada de la invención

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

En el presente documento se desvelan procedimientos para seleccionar un clonotipo raro y/o un grupo raro de clonotipos relacionados, que se correlacionan con un trastorno proliferativo linfoide o mieloide y que se pueden usar para controlar el estado del trastorno con una probabilidad mínima de que se haga una determinación falsamente positiva de la recurrencia de la enfermedad. Dichos clonotipos raros son particularmente útiles para controlar la enfermedad residual mínima de un cáncer después del tratamiento, en el que el resultado de tal control es un factor clave en la determinación de si continuar, no continuar o, de otro modo, modificar el tratamiento. En muchos neoplasias linfoides y mieloides malignos, una muestra de tejido de diagnóstico (o también denominada en el presente documento "muestra de diagnóstico"), tal como una muestra de sangre periférica o una muestra de médula ósea, se obtiene antes del tratamiento a partir de la cual se genera del perfil de clonotipo (un "perfil de clonotipo diagnóstico"). Uno o más clonotipos correlacionados con la enfermedad (es decir, "clonotipos correlacionados") se identifican en el perfil de clonotipo, generalmente como los clonotipos que tienen las frecuencias más altas. Por ejemplo, se puede determinar una frecuencia de diagnóstico predeterminada y cada clonotipo que tenga una frecuencia mayor que la frecuencia de diagnóstico predeterminada se selecciona como miembro de un conjunto de

clonotipos candidatos. En algunas realizaciones, uno o más clonotipos correlacionados utilizados para controlar la EMR se toman del conjunto de clonotipos candidatos que son lo suficientemente raros como para minimizar las determinaciones falsas positivas de EMR. El número de clonotipos candidatos en un conjunto puede variar según el tipo de cáncer, la etapa del cáncer, el historial de tratamiento y similares. En algunas realizaciones, un conjunto incluye hasta 10 de los clonotipos de mayor frecuencia de un perfil de diagnóstico de clonotipos. En otras realizaciones, un conjunto incluye hasta 5 de los clonotipos de frecuencia más alta de un perfil de diagnóstico de clonotipos. Después del tratamiento y, preferentemente después de conseguir una remisión completa del cáncer, la presencia, ausencia o frecuencia de tales clonotipos de correlación se evalúa periódicamente para determinar si la remisión permanece o si la neoplasia vuelve o reincide, basándose en la presencia de, o un aumento en la frecuencia de, los clonotipos correlacionados (o clonotipos relacionados) en un perfil de clonotipo posterior al tratamiento. Es decir, después del tratamiento, se evalúa la enfermedad mínima residual basándose en la presencia, ausencia o frecuencia de los clonotipos correlacionados. Normalmente, si la frecuencia o el nivel de dicho clonotipo correlacionado aumenta para igualar o exceder un valor predeterminado, por ejemplo, 0,01 por ciento o 0,1 por ciento o 1 por ciento (que puede depender de la enfermedad, la etapa de la enfermedad, el estado del paciente o similar), se puede tomar una decisión de tratamiento, tal como, implementar un tratamiento adicional de acuerdo con un protocolo anterior, modificar un protocolo de tratamiento, tal como aumentar las dosis, añadir diferentes modalidades de tratamiento, tal como el uso de un medicamento diferente o similar. Como se ha mencionado anteriormente, cuando tales clonotipos de correlación son comunes o se corresponden con un segmento de receptor reordenado que no tiene suficiente diversidad (de modo que las células no cancerosas pueden compartir el clonotipo), la aparición de tales clonotipos en un perfil de clonotipo postratamiento puede provocar una indicación de falso positivo de recaída. Además se desvela un procedimiento para evaluar la rareza de los clonotipos correlacionados y, por lo tanto, un procedimiento para seleccionar los clonotipos correlacionados que minimiza la probabilidad de determinaciones falsas positivas de recaída. En este aspecto, después de que se genera un perfil de clonotipo diagnóstico, se busca en una base de datos de clonotipos la presencia de clonotipos idénticos y clonotipos relacionados por clan, dado que el concepto se describe más completamente a continuación. De esta manera, independientemente de qué clonotipos correlacionados se seleccionen para el control, incluso antes de que comiencen las pruebas posteriores al tratamiento (ya sea por perfil de clonotipo o por PCR directa de los clonotipos seleccionados), se puede estimar una probabilidad de que los clonotipos seleccionados puedan dar lugar a una indicación de falso positivo.

Los procedimientos desvelados en el presente documento son aplicables para controlar cualquier enfermedad proliferativa en la que un ácido nucleico reordenado que codifica un receptor inmunitario o porción del mismo puede usarse como marcador de células implicadas en la enfermedad. Por lo tanto, el procedimiento es aplicable tanto a los procedimientos basados en la secuencia que implican la generación de perfiles de clonotipo en cada medición posterior al tratamiento como a los procedimientos de seguimiento de un solo clon que implican una amplificación por PCR de un solo clonotipo. En un aspecto, los procedimientos desvelados en el presente documento son aplicables a trastornos proliferativos linfoides y mieloides. En otro aspecto, los procedimientos son aplicables a linfomas y leucemias. En otro aspecto, los procedimientos son aplicables para control la EMR en el linfoma folicular, leucemia linfocítica crónica (LLC), leucemia linfocítica aguda (LLA), leucemia mielógena crónica (LMC), leucemia mielógena aguda (LMA), linfomas de Hodgkin y de no Hodgkin, mieloma múltiple (MM), gammapatía monoclonal de significancia indeterminada (MGUS), linfoma de células del manto (LCM), linfoma difuso de linfocitos B grandes (LDCBG), síndromes mielodisplásicos (SMD), linfoma de linfocitos T o similares. En una realización particular, un procedimiento desvelado en el presente documento es particularmente adecuado para controlar la EMR en la LLA, la MM o el LDCBG.

En algunas realizaciones, particularmente para trastornos proliferativos linfoides y mieloides, Se genera una pluralidad de perfiles de clonotipo a partir de una muestra de diagnóstico para que se puedan evaluar múltiples tipos de clonotipos. Es decir, el ADN recombinado o las secuencias de ácido nucleico (también denominadas en el presente documento "ADN o ácidos nucleicos reordenados somáticamente" o "ADN o ácido nucleico reordenado") de linfocitos T y/o linfocitos B pueden incluir secuencias de ácido nucleico que codifican cadenas o porciones completas de los siguientes: un reordenamiento de VDJ de IgH, un reordenamiento de DJ de IgH, un reordenamiento de VJ de IgK, un reordenamiento de VJ de IgL, un reordenamiento de VDJ de TCR β, un reordenamiento de DJ de TCR β, un reordenamiento de VJ de TCR α, un reordenamiento de VJ de TCR γ, un reordenamiento de VDJ de TCR δ, o un reordenamiento de VD de TCR δ. Si se sabe a través de otras pruebas que el cáncer de un paciente procede de los linfocitos B, los clonotipos analizados en una muestra de diagnóstico pueden limitarse a secuencias de ácido nucleico que codifican cadenas completas o porciones de lo siguiente: un reordenamiento de VDJ de IgH, un reordenamiento de DJ de IgH, un reordenamiento de VJ de IgK Y/o un reordenamiento de VJ de IgL. Asimismo, si se sabe a través de otras pruebas que el cáncer de un paciente deriva de linfocitos T, los clonotipos analizados en una muestra de diagnóstico pueden limitarse a secuencias de ácido nucleico que codifican cadenas completas o porciones de lo siguiente: un reordenamiento de VDJ de TCR β, un reordenamiento de DJ de TCR β, un reordenamiento de VJ de TCR α, un reordenamiento de VJ de TCR γ, un reordenamiento de VDJ de TCR δ y/o un reordenamiento de VD de TCR δ. En algunas realizaciones, los perfiles de clonotipos generados a partir de una muestra de diagnóstico en el procedimiento descrito en el presente documento se basan en clonotipos que codifican cadenas completas o porciones de lo siguiente: un segmento de IgH que contiene VDJ, un segmento de IgH que contiene DJ, un segmento de IgK que contiene VJ, un segmento de TCR β que contiene VDJ, un segmento de TCR γ y/o un segmento de TCR δ. En algunas realizaciones, las muestras de diagnóstico pueden comprender tejidos de

médula ósea, sangre periférica o ganglios linfáticos u otro tejido linfoide.

En algunas realizaciones, el aspecto anterior se puede implementar con las siguientes etapas: (a) obtener una muestra de diagnóstico del paciente que comprende linfocitos T y/o linfocitos B; (b) amplificar moléculas de ácido nucleico de los linfocitos T y/o linfocitos B de la muestra, comprendiendo las moléculas de ácido nucleico secuencias de ADN recombinadas de genes del receptor de linfocitos T o genes de inmunoglobulina; (c) secuenciar las moléculas amplificadas de ácido nucleico para formar un perfil de clonotipo; (d) seleccionar un conjunto de uno o más clonotipos candidatos en función de su frecuencia en el perfil de clonotipo; (e) comparar el uno o más clonotipos candidatos del conjunto a los clonotipos de una base de datos de clonotipos que contiene clonotipos de al menos un individuo que no sea el paciente para determinar una presencia, ausencia y/o nivel en la base de datos de clonotipos de cada clonotipo candidato y eliminar del conjunto cualquier clonotipo candidato que tenga una frecuencia en la base de datos de clonotipo mayor que una frecuencia predeterminada; y (f) seleccionar como uno o más clonotipos específicos del paciente para monitorizar la enfermedad residual mínima los clonotipos candidatos que permanecen en el conjunto. Como se describe más completamente a continuación, la etapa de comparar los clonotipos candidatos de un conjunto obtenido del perfil de clonotipo (es decir, "clonotipos medidos") con los de la base de datos de clonotipos se lleva a cabo buscando en la base de datos de clonotipos, clonotipos coincidentes o clonotipos que de otro modo están relacionados, por ejemplo, por cercanía de secuencia o por evoluciones clonales. En algunas realizaciones, la etapa de la comparación puede implementarse requiriendo una coincidencia exacta entre los clonotipos medidos y los clonotipos de la base de datos para contar como coincidencia; en otras realizaciones, la etapa de la comparación puede implementarse requiriendo la coincidencia de solo un subconjunto de nucleótidos entre los clonotipos medidos y los clonotipos de la base de datos para contar como coincidencia. Por ejemplo, en algunas realizaciones, un clonotipo de base de datos puede contarse como coincidente con un clonotipo medido si existe una identidad superior al 90 por ciento entre los dos. En algunas realizaciones, un clonotipo de la base de datos se puede contar como un clonotipo medido si está relacionado por estar en el mismo clan que el clonotipo de medida. En algunas realizaciones, un clonotipo puede ser útil como clonotipo para controlar la EMR o contaminación, a pesar de que está presente en una base de datos de clonotipos a un nivel bajo, umbral o frecuencia predeterminado. En dichas realizaciones, se puede emplear un clonotipo para controlar la EMR o detectar contaminación si está presente a una frecuencia predeterminada o por debajo de ella, que puede depender del tamaño de las entradas de clonotipo en la base de datos o del número de las mismas. En una realización, una frecuencia predeterminada es 10^{-7}; en otra realización, una frecuencia predeterminada es 10^{-8}; y en otra realización; una frecuencia predeterminada es 10^{-9}.

En algunas realizaciones, un clonotipo candidato puede eliminarse de un conjunto siempre que una diferencia de secuencia entre él y un miembro de una base de datos de clonotipo esté por debajo de algún valor mínimo (denominado en el presente documento "diferencia predeterminada"). Las diferencias de secuencia incluyen diferencias de sustituciones, inserciones y deleciones de bases que pueden medirse mediante una medida de distancia de secuencia, tal como una distancia de Hamming. Las diferencias de secuencia también incluyen diferencias debido a reordenamientos somáticos, tales como, reemplazos de VH. En algunas realizaciones, un clonotipo candidato puede eliminarse de un conjunto si su secuencia está dentro de alguna medida de distancia de secuencia predeterminada de una secuencia de clonotipo en la base de datos. En una realización, tal medida de la distancia de secuencia es una distancia de Hamming.

En el caso de que ningún clonotipo candidato permanezca en un conjunto después de la aplicación de las etapas del procedimiento desvelado, las opciones para la frecuencia de diagnóstico predeterminada para seleccionar los clonotipos candidatos o para la frecuencia predeterminada para rechazar los clonotipos candidatos de un conjunto pueden alterarse para obtener clonotipos correlacionados para controlar la EMR que tienen una mayor probabilidad de resultados falsos positivos. En el caso de que una pluralidad de clonotipos candidatos permanezca en un conjunto después de la aplicación de las etapas de la invención, uno o más de los miembros restantes pueden usarse para controlar la EMR, por ejemplo, como lo enseñaron Faham y Willis, por ejemplo, las patentes de Estados Unidos 8.236.503 u 8.628.927. En algunas realizaciones, se puede usar un clonotipo candidato único que tenga la mayor diferencia con respecto a los clonotipos de la base de datos para controlar la EMR. En algunas realizaciones, la frecuencia predeterminada y la frecuencia de diagnóstico predeterminada se seleccionan de modo que al menos un clonotipo candidato permanezca en el conjunto después de la aplicación de las etapas del procedimiento desvelado.

Con respecto a contar dos clonotipos como "coincidentes" según la pertenencia al clan, en algunas realizaciones, se puede considerar que dos clonotipos están en el mismo clan si se cumple uno o más de los siguientes criterios: (a) los clonotipos son al menos noventa por ciento idénticos entre sí; (b) los clonotipos comprenden secuencias recombinadas de la cadena pesada de inmunoglobulina que están relacionadas por un reemplazo de VH; (c) los clonotipos comprenden secuencias recombinadas de la cadena pesada de inmunoglobulina que están relacionadas por hipermutación; y (d) los clonotipos están relacionados porque tienen una región V y una región J idénticamente mutadas pero tienen una región NDN diferente.

Además se desvela un procedimiento para seleccionar clonotipos raros y/o grupos raros de clonotipos o clanes relacionados, que pueden usarse para detectar y/o cuantificar la contaminación cruzada de las muestras, particularmente entre las muestras en las que las secuencias de nucleótidos que codifican los receptores inmunes

reordenados se amplifican mediante técnicas como la PCR. En situaciones, tal como un laboratorio clínico, donde las muestras de diferentes pacientes se procesan generando perfiles de clonotipo como se describe en el presente documento, existe una probabilidad significativa de contaminación cruzada de muestras, que puede implicar el registro de clonotipos derivados de una primera muestra como miembros de un perfil de clonotipo de una segunda muestra. Dicha contaminación potencial puede detectarse rápidamente comparando clonotipos raros de un perfil de clonotipo de la primera muestra con los de un perfil de clonotipo de la segunda muestra. Si hay una o más coincidencias, la contaminación cruzada entre las muestras es altamente probable.

En algunas realizaciones, el aspecto anterior se puede implementar con las siguientes etapas: (a) generar un primer perfil de clonotipo a partir de ácido nucleico de la muestra de tejido del primer individuo; (b) generar un segundo perfil de clonotipo a partir de ácido nucleico de la muestra de tejido del segundo individuo; (c) comparar clonotipos del primer y segundo perfil de clonotipos con clonotipos de una base de datos de clonotipos que contiene clonotipos de al menos un individuo que no sea el primer y el segundo individuo para determinar una presencia, ausencia y/o nivel en la base de datos de clonotipo de cada clonotipo de los perfiles de clonotipo primero y segundo que están ausentes de la base de datos de clonotipo o A un nivel en la base de datos de clonotipo por debajo de un umbral predeterminado; y (d) clasificar la muestra de tejido del primer individuo como contaminada por ácidos nucleicos del segundo individuo siempre que cualquiera de dichos clonotipos determinados esté presente tanto en el primer como en el segundo perfil de clonotipo.

En algunas realizaciones, un procedimiento para generar perfiles de clonotipo de secuencias de ADN recombinadas en linfocitos T y/o linfocitos B comprende las siguientes etapas: (a) obtener una muestra de un sujeto que comprende linfocitos T y/o linfocitos B; (b) aislar espacialmente moléculas individuales de secuencias de ADN recombinadas de dichas células en un sustrato sólido; (c) secuenciar dichas moléculas individuales aisladas espacialmente de secuencias de ADN recombinadas para proporcionar al menos 1000 lecturas de secuencia que tengan una tasa de error; (d) lecturas de secuencias coalescentes en diferentes secuencias de ADN recombinadas de la muestra siempre que las lecturas de secuencia sean distintas con una confianza de al menos 99,9 por ciento; y (e) determinar los niveles de las diferentes secuencias de ADN recombinadas de dicha muestra para generar dicho perfil de secuencias de ADN recombinadas.

En algunas realizaciones, una base de datos de clonotipos puede ser cualquier base de datos de clonotipos que contengan clonotipos de al menos un individuo que no sea el paciente. Normalmente, la base de datos de clonotipos es una colección de clonotipos de perfiles de clonotipos de una pluralidad de individuos. Por ejemplo, dicha colección puede incluir perfiles de clonotipos de al menos 10^4 clonotipos, cada uno de al menos 10 individuos distintos del paciente; o dicha colección puede incluir perfiles de clonotipos de al menos 10^4 clonotipos cada uno de al menos 100 individuos. En otra realización, una base de datos de clonotipos comprende perfiles de clonotipos que tienen cada uno al menos 10^4 clonotipos de una población de individuos, para que la base de datos contenga al menos 10^8 clonotipos o 10^9 clonotipos o 10^{10} clonotipos. Las bases de datos de clonotipos anteriores pueden comprender clonotipos que codifican segmentos específicos de moléculas de receptores inmunes particulares. Como se describe más completamente a continuación, en un aspecto, los clonotipos obtenidos en un procedimiento desvelado en el presente documento codifican segmentos de moléculas de receptores inmunes que son idénticos o sustancialmente idénticos a los codificados por los clonotipos de la base de datos de clonotipos buscados en el procedimiento, donde "sustancialmente idéntico" significa que existe una superposición suficiente para que los clonotipos medidos puedan buscarse efectivamente contra los clonotipos de la base de datos. En una realización, los clonotipos obtenidos en el procedimiento y los clonotipos de la base de datos codifican una región V(D)J de una molécula de IgH o una porción de la misma; en otra realización, los clonotipos obtenidos en el procedimiento y los clonotipos de la base de datos codifican una región DJ de una molécula de IgH o una porción de la misma; en otra realización, los clonotipos obtenidos en el procedimiento y los clonotipos de la base de datos codifican una molécula de IgK o una porción de la misma; en otra realización, los clonotipos obtenidos en el procedimiento y los clonotipos de la base de datos codifican una molécula de TCRβ o una porción de la misma; en otra realización, los clonotipos obtenidos en el procedimiento y los clonotipos de la base de datos codifican una molécula de TCRγ o una porción de la misma; en otra realización, los clonotipos obtenidos en el procedimiento y los clonotipos de la base de datos codifican una molécula de TCRδ o una porción de la misma. En otras realizaciones, los clonotipos medidos y los clonotipos de la base de datos codifican segmentos de los siguientes receptores inmunes: un reordenamiento de VDJ de IgH, un reordenamiento de DJ de IgH, un reordenamiento de VJ de IgK, un reordenamiento de VJ de IgL, un reordenamiento de VDJ de TCR β, un reordenamiento de DJ de TCR β, un reordenamiento de VJ de TCR α, un reordenamiento de VJ de TCR γ, un reordenamiento de VDJ de TCR δ y un reordenamiento de VD de TCR δ. En algunas realizaciones, los clonotipos medidos y los clonotipos de bases de datos tienen longitudes en el rango de 25 a 400 nucleótidos. En algunas realizaciones, los clonotipos medidos y los clonotipos de bases de datos tienen las mismas longitudes.

En algunas realizaciones, los individuos que contribuyen a una base de datos de clonotipos pueden ser de una población o grupo relacionado. Por ejemplo, las bases de datos de un clonotipo pueden ser específicas para individuos que tienen una enfermedad particular, por ejemplo, un trastorno proliferativo linfoide o mieloide, neoplasia linfoide, o una afección constitutiva, por ejemplo, MGUS. En algunas realizaciones, las personas que contribuyen a una base de datos de clonotipos pueden ser de un grupo étnico o racial relacionado, tales como, africanos, japoneses, caucásicos o similares. En una realización, una base de datos de clonotipos comprende perfiles de

clonotipos que tienen cada uno al menos 10^4 clonotipos obtenidos de una pluralidad de pacientes que padecen neoplasias de linfocitos B. En otra realización, una base de datos de clonotipos comprende perfiles de clonotipos que tienen cada uno al menos 10^4 clonotipos obtenidos de una pluralidad de pacientes que padecen linfomas de linfocitos B. En otra realización, una base de datos de clonotipos comprende perfiles de clonotipos que tienen cada uno al menos 10^4 clonotipos obtenidos de una pluralidad de pacientes que sufren linfomas difusos de células B grandes. En otra realización, una base de datos de clonotipos comprende perfiles de clonotipos que tienen cada uno al menos 10^4 clonotipos obtenidos de una pluralidad de pacientes que padecen mieloma múltiple.

En otra realización, una base de datos de clonotipos comprende perfiles de clonotipos que tienen al menos 10^4 clonotipos obtenidos de una pluralidad de pacientes que padecen neoplasias de linfocitos T.

Una amplia variedad de procedimientos de búsqueda o algoritmos pueden usarse para llevar a cabo la etapa de comparación de clonotipos medidos con clonotipos de bases de datos. Muchos algoritmos de búsqueda y alineación de secuencia convencionales están disponibles públicamente y se han descrito en las siguientes referencias: Mount, Bioinformatics Sequence and Genome Analysis, Segunda edición (Cold Spring Harbor Press, 2004); Batzoglou, Briefings in Bioinformatics, 6: 6-22 (2005); Altschul y col., J. Mol. Biol., 215(3): 403-410 (1990); Needleman y Wunsch, J. Mol. Biol., 48: 443-453 (1970); Smith y Waterman, Advances in Applied Mathematics, 2: 482-489 (1981); y similares. Como se describe más completamente a continuación, algunas etapas de comparación pueden requerir procedimientos de búsqueda más especializados, que pueden implementarlos los expertos en la técnica en el campo de la bioinformática; por ejemplo, para buscar clonotipos relacionados por reordenamientos, tales como reemplazos de VH, o similares, puede requerir modificaciones obvias de los procedimientos o algoritmos de búsqueda disponibles. Por ejemplo, puede que no tenga sentido realizar una alineación convencional entre un clonotipo basado en VDJ y aquellos en una base de datos para determinar la presencia de reordenamientos de VH. Un experto en la materia entendería que, e ese caso, se podría realizar una alineación usando la secuencia del clonotipo menos las secuencias de la región V.

En un aspecto, las bases de datos de clonotipos no solo buscan clonotipos idénticos a los clonotipos medidos, sino también clonotipos que están relacionados, por ejemplo, por ser miembros del mismo clan, o por tener una relación filogénica. Por lo tanto, en algunas realizaciones, una búsqueda en una base de datos de clonotipos recuperará cualquier clonotipo de base de datos que sea miembro del mismo clan que el clonotipo medido. Dicha recuperación indica la presencia de un miembro del clan que puede tener o no una secuencia idéntica al clonotipo medido, pero que satisface uno o más criterios de relación para determinar la pertenencia al clan. Los criterios de ejemplo para definir un clan pueden incluir uno o más de los siguientes: (a) los clonotipos son al menos noventa por ciento idénticos entre sí, (b) los clonotipos codifican segmentos de IgH y son idénticos, excepto por diferentes mutaciones de hipermutación somática, (c) los clonotipos están relacionados por un reemplazo de VH, (d) los clonotipos tienen regiones V idénticas y regiones J idénticas que incluyen mutaciones idénticas en cada región, pero tienen diferentes regiones NDN, (e) los clonotipos tienen secuencias idénticas, excepto por una o más inserciones y/o deleciones de 1-10 bases. En algunas realizaciones, en el ejemplo anterior (e), los clonotipos pueden ser miembros del mismo clan si tienen secuencias idénticas, excepto para una o más inserciones y/o deleciones de 1-5 bases o de 1-3 bases.

En otro aspecto, la información de una o más bases de datos de clonotipos puede usarse para determinar los parámetros de un modelo de desarrollo de clonotipos, que luego se puede usar para proporcionar rápidamente frecuencias o probabilidades de ocurrencia, de clonotipos basados en su estructura, es decir, los tipos y/o tamaños de sus segmentos de genes constituyentes y sus modificaciones mediante el proceso de reordenamiento. Por ejemplo, Murugan y col., (citado anteriormente) proporciona un ejemplo de dicho modelo para un conjunto limitado de clonotipos. Otro ejemplo se ilustra con la ayuda de la figura 4A para el caso de los clonotipos que codifican regiones VDJ de moléculas de IgH o TCRβ. El examen de las bases de datos de clonotipos ha indicado que no todos los clonotipos que pueden ser generados por procesos naturales son igualmente probables. El proceso para generar diversidad de una región VDJ se ilustra en la Figura 4A. Aproximadamente, el proceso implica mecanismos para (i) añadir nucleótidos "P" a los extremos de los segmentos génicos D y J, (ii) delecionar nucleótidos de segmentos de genes J o D en la unión DJ y de segmentos génicos D o V en la unión DV e (iii) insertar aleatoriamente de 1-40 nucleótidos en cada una de tales uniones, por ejemplo, Janeway y col., Immunobiology, Sexta edición (Garland Science, Nueva York, 2005). Por lo tanto, todos los posibles resultados de este proceso pueden enumerarse y sus respectivas frecuencias pueden determinarse a partir de bases de datos de clonotipos. Como alternativa, se pueden determinar las frecuencias de los segmentos componentes, tal como la frecuencia de una región J (o una región J con un truncamiento dado), una región V (o una región V con un truncamiento dado), una región NDN, etc. Tales modelos de frecuencia pueden usarse en lugar de una búsqueda en una base de datos de clonotipos para determinar la rareza de un clonotipo medido en particular. Este concepto se describe en el diagrama de flujo de la Figura 4B. Se crea un clonotipo (por ejemplo, un VDJ recombinante) mediante la operación de uno o más de los procedimientos anteriores. Dado que los procedimientos son probabilísticos y funcionan en varios casos (por ejemplo, adiciones de nucleótidos frente a deleciones de nucleótidos), casi todos los clonotipos pueden producirse por múltiples vías a través de las operaciones anteriores (por ejemplo, añadir 9 nucleótidos y luego eliminar 6 nucleótidos (para una adición neta de 3) puede ser equivalente a añadir 4 y luego delecionar 1). Sin embargo, la probabilidad de añadir 9 seguido de delecionar 6 puede tener una probabilidad muy diferente a la de añadir 4, seguido de delecionar 1. Se puede usar una base de datos de clonotipos para estimar las probabilidades asociadas con las diversas operaciones que resultan en clonotipos dados. Es decir, se puede usar una base de

datos de clonotipos para anotar cada característica de clonotipo con una probabilidad de que ocurra (452). Como se observa en (450) de la Figura 4B, para un clonotipo basado en un reordenamiento de VDJ, tales características pueden incluir la selección del segmento J, la selección del segmento D, la selección del segmento V, P nucleótidos presentes, N nucleótidos añadidos, Δ (número de nucleótidos delecionados) y similares. Una vez que se obtiene dicho modelo, la rareza (o estimación de la frecuencia) de un clonotipo dado en función de las características de su componente puede determinarse multiplicando las probabilidades del componente (454).

Volviendo a la Figura 4A, una molécula de IgH o TCRβ comprende una región VDJ ensamblada a partir de la región J (400), la región D (402) y la región V (403); sin embargo, durante el ensamblaje 0, se delecionan varios nucleótidos en los extremos de los segmentos génicos en la unión JD y la unión DV. Si las líneas (408) y (410) indican las posiciones de las uniones JD y DV sin deleciones, los intervalos (414) y (412) indicados por flechas, ilustra el intervalo de posibles deleciones de los extremos de los segmentos génicos. En algunos casos, la región D (402) puede delecionarse por completo. La diversidad se genera aún más mediante un mecanismo que inserta de 0 a aproximadamente 20 nucleótidos en cada unión. Así, después de las deleciones, los segmentos oligonucleotídicos (424) y (426) se insertan entre los extremos (posiblemente truncados) de los segmentos génicos. El posicionamiento de la inserción depende de lo que se delecionó, pero ocurre dentro de los intervalos (420) y (422). Como resultado de estos mecanismos, se genera una gran cantidad de clonotipos posibles (428) que pueden incluir o no una región D reconocible (430). Las probabilidades de cualquiera de estos reordenamientos pueden estimarse examinando sus ocurrencias en una o más bases de datos de clonotipos. Por lo tanto, la región J (434) de un tipo particular y que tiene una longitud particular puede tener una frecuencia esperada de ocurrencia; asimismo, la región NDN (432) de una longitud particular y la presencia de ausencia de un segmento de gen D identificable puede tener una frecuencia esperada de ocurrencia; y así sucesivamente para la región V (436). En una realización de este aspecto, la rareza de un clonotipo medido puede estimarse rápidamente multiplicando las frecuencias esperadas (o probabilidades) de sus regiones constituyentes.

Muestras

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

Las muestras usadas en los procedimientos de la invención pueden provenir de diversos tejidos, que incluyen, por ejemplo, tejido tumoral, sangre y plasma sanguíneo, fluido linfático, líquido cefalorraquídeo que rodea el cerebro y la médula espinal, fluido sinovial que rodea las articulaciones ósea y similares. En una realización, la muestra es una muestra de sangre. La muestra de sangre puede ser de aproximadamente 0,1, 0,2, 0,3, 0,4, 0,5, 0,6, 0,7, 0,8, 0,9, 1,0, 1,5, 2,0, 2,5, 3,0, 3,5, 4,0, 4,5, o 5,0 ml. La muestra puede ser una biopsia tumoral. La biopsia puede ser, de, por ejemplo, un tumor del cerebro, hígado, pulmón, corazón, colon, riñón o médula ósea. Cualquier técnica de biopsia usada por un experto en la técnica puede usarse para aislar una muestra del sujeto. Por ejemplo, una biopsia puede ser una biopsia abierta, en la que se usa anestesia general. La biopsia puede ser una biopsia cerrada, en la que se realiza un corte más pequeño que en una biopsia abierta. La biopsia puede ser una biopsia central o por incisión, en la que se retira parte del tejido. La biopsia puede ser una biopsia por escisión, en la que se intenta retirar una lesión completa. La biopsia puede ser una biopsia de aspiración por aguja fina, en la que se retira una muestra de tejido o fluido con una aguja.

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

La muestra se puede obtener del material corporal que deja un sujeto. Tal material desechado puede incluir desechos humanos. El material desechado también podría incluir células de piel desprendidas, sangre, dientes o cabello.

En el caso de trastornos linfoproliferativos, tales como los linfomas, puede obtenerse una muestra de calibración o diagnóstico de tejidos linfoides, de lesiones causadas por el trastorno, por ejemplo, lesiones metastásicas, o de tejidos afectados indirectamente por el trastorno. Para las neoplasias linfoides, existe una guía ampliamente disponible y kits disponibles comercialmente para inmunofenotipar y enriquecer linfocitos relacionados con enfermedades, por ejemplo, "U.S.-Canadian consensus recommendations on the immunophenotypic analysis of haematologic neoplasia by flow cytometry," Cytometry, 30: 214-263 (1997); paneles de anticuerpos MultiMix™ para inmunofenotipaje de leucemia y linfoma mediante citometría de flujo (Dako, Dinamarca); y similares. Los tejidos linfoides incluyen ganglios linfáticos, bazo, amígdalas, adenoides, timo y similares.

La muestra puede incluir ácido nucleico, por ejemplo, ADN (por ejemplo, ADN genómico) o ARN (por ejemplo, ARN mensajero). El ácido nucleico puede ser ADN o ARN sin células, por ejemplo, extraído del sistema circulatorio, Vlassov y col., Curr. Mol. Med., 10: 142-165 (2010); Swarup y col., FEBS Lett., 581: 795-799 (2007). En los procedimientos desvelados en el presente documento, la cantidad de ARN o ADN de un sujeto que puede analizarse incluye, por ejemplo, tan bajo como una sola célula en algunas aplicaciones (por ejemplo, una prueba de calibración) y hasta 10 millones de células o más que se traducen en un intervalo de ADN de 6 pg-60 ug y ARN de aproximadamente 1 pg-10 ug.

En un aspecto, una muestra de linfocitos para generar un perfil de clonotipo es suficientemente grande para que sustancialmente cada linfocito T o linfocito B con un clonotipo distinto se represente en este. En una realización, se toma una muestra que contiene una probabilidad del noventa y nueve por ciento cada clonotipo de una población presente a una frecuencia de 0,001 por ciento o superior. En otra realización, se toma una muestra que contiene una probabilidad del noventa y nueve por ciento cada clonotipo de una población presente a una frecuencia de 0,0001 por ciento o superior. En una realización, una muestra de linfocitos B o linfocitos T incluye al menos medio millón de células y, en otra realización, tal muestra incluye al menos un millón de células.

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

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

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

El ADN y ARN analizados en los procedimientos desvelados en el presente documento pueden corresponderse con secuencias que codifican inmunoglobulinas de cadena pesada (IgH) con regiones constantes (α, δ, ε, γ, o µ) o inmunoglobulinas de cadena ligera (IgK o IgL) con regiones constantes λ o k. Cada anticuerpo tiene dos cadenas ligeras idénticas y dos cadenas pesadas idénticas. Cada cadena está compuesta de una región constante (C) y una variable. Para cada cadena pesada, la región variable está compuesta de una variable (V), diversidad (D) y segmentos de unión (J). Varias secuencias distintas que codifican cada tipo de estos segmentos están presentes en el genoma. Un evento de recombinación VDJ específico se produce durante el desarrollo de un linfocito B, haciendo que esa célula genere una cadena pesada específica. La diversidad en la cadena ligera se genera de un modo

similar excepto en que no hay región D de modo que solo es una recombinación VJ. La mutación somática a menudo se produce cerca del sitio de recombinación, provocando la adición o supresión de varios nucleótidos, aumentando adicionalmente la diversidad de cadenas pesadas y ligeras generadas por los linfocitos B. La posible diversidad de los anticuerpos generados por un linfocito B es entonces el producto de las distintas cadenas pesadas y ligeras. Las regiones variables de las cadenas pesadas y ligeras contribuyen a formar el sitio o región de reconocimiento de antígeno (o unión). Añadida a esta diversidad hay un proceso de hipermutación somática que puede producirse después de que se haya montado una respuesta específica frente a algún epítopo.

Como se ha mencionado anteriormente, los cebadores pueden seleccionarse para generar amplicones de subconjuntos de ácidos nucleicos recombinados extraídos de linfocitos. En el presente documento se puede hacer referencia a tales subpoblaciones como "regiones somáticamente reordenadas". Las regiones somáticamente reordenadas pueden comprender ácidos nucleicos de linfocitos en desarrollo o completamente desarrollados, donde los linfocitos en desarrollo son células en la que la reorganización de genes inmunitarios no se ha completado para formar moléculas que tengan regiones V(D)J completas. Las regiones somáticamente reordenadas incompletas de ejemplo incluyen moléculas de IgH incompletas (tales como, moléculas que contienen solo regiones D-J), moléculas de TCRδ incompletas (tales como moléculas que contienen solo regiones D-J) e IgK inactiva (por ejemplo, que comprende regiones Kde-V).

El muestreo adecuado de las células es un aspecto importante de la interpretación de los datos del repertorio, como se describe más adelante en las definiciones de "clonotipo" y "repertorio". Por ejemplo, comenzar con 1.000 células crea una frecuencia mínima a la que el ensayo es sensible independientemente de cuántas lecturas de secuencia se obtengan. Por lo tanto, un aspecto es el desarrollo de procedimientos para cuantificar el número de moléculas de receptores inmunes de entrada. Esto se ha implementado para secuencias de TCRβ e IgH. En cualquier caso, se usa el mismo conjunto de cebadores que son capaces de amplificar todas las secuencias diferentes. Para obtener un número absoluto de copias, se realiza una PCR en tiempo real con el múltiplex de cebadores junto con un estándar con un número conocido de copias de receptores inmunes. Esta medición de PCR en tiempo real se puede hacer a partir de la reacción de amplificación que posteriormente se secuenciará o se puede hacer en una parte alícuota separada de la misma muestra. En el caso del ADN, el número absoluto de moléculas de receptores inmunes reordenados se puede convertir fácilmente en un número de células (dentro de 2 veces ya que algunos linfocitos tendrán 2 copias reordenadas del receptor inmunitario específico evaluado y otros tendrán una). En el caso del ADNc, el número total medido de moléculas reordenadas en la muestra en 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 procedimiento se puede combinar con un procedimiento para determinar la cantidad total de ARN para definir el número de moléculas de receptores inmunes reordenados en una cantidad unitaria (por ejemplo, 1 µg) de ARN suponiendo una eficiencia específica de la síntesis de ADNc. Si se mide la cantidad total de ADNc, no es necesario considerar la eficiencia de la síntesis de ADNc. Si también se conoce el número de células, se pueden calcular las copias del receptor inmunitario reordenadas por célula. Si no se conoce el número de células, se puede estimar a partir del ARN total, ya que las células de tipo específico generalmente generan una cantidad comparable de ARN. Por lo tanto, a partir de las copias de las moléculas de receptores inmunes reordenadas por 1 µg, se puede estimar el número de estas moléculas por célula.

Una desventaja de realizar una PCR en tiempo real separada de la reacción que se procesaría para la secuenciación es que puede haber efectos inhibitorios que son diferentes en la PCR en tiempo real de la otra reacción como enzimas diferentes, puede utilizarse ADN de entrada y otras condiciones. El procesamiento de los productos de la PCR en tiempo real para la secuencia mejoraría este problema. Sin embargo, un número bajo de copias usando PCR en tiempo real puede deberse a un número bajo de copias o a efectos inhibitorios u otras condiciones subóptimas en la reacción.

Otro enfoque que se puede utilizar es añadir una cantidad conocida de moléculas reordenadas del receptor inmune único con una secuencia conocida, es decir, cantidades conocidas de uno o más patrones internos, al ADNc o ADN genómico de una muestra de cantidad desconocida. Al contar el número relativo de moléculas que se obtienen para la secuencia añadida conocida en comparación con el resto de las secuencias de la misma muestra, se puede estimar el número de moléculas de receptores inmunes reordenadas en la muestra inicial de ADNc. (Tales técnicas para el recuento molecular son bien conocidas, por ejemplo, Brenner y col., patente de Estados Unidos 7.537.897). Los datos de la secuenciación de la secuencia única añadida se pueden usar para distinguir las diferentes posibilidades si también se usa una calibración de PCR en tiempo real. El bajo número de copias del receptor inmune reordenado en el ADN (o ADNc) crearía una alta relación entre el número de moléculas para la secuencia enriquecida en comparación con el resto de las secuencias de muestra. Por otra parte, si el número bajo de copias medido por PCR en tiempo real se debe a la ineficiencia en la reacción, la relación no sería alta.

Amplificación de poblaciones de ácidos nucleicos

Los amplicones de poblaciones objetivo de ácidos nucleicos pueden generarse mediante diversas técnicas de amplificación. En un aspecto de la invención, se utiliza PCR multiplex para amplificar los miembros de una mezcla de ácidos nucleicos, particularmente mezclas que comprenden moléculas inmunes recombinadas, tales como receptores de linfocitos T, o porciones de los mismos. La guía para llevar a cabo PCR multiplex de tales moléculas

inmunes se encuentra en las siguientes referencias: Morley, patente de Estados Unidos 5.296.351; Gorski, patente de Estados Unidos 5.837.447; Dau, patente de Estados Unidos 6.087.096; Von Dongen y col., publicación de la patente de Estados Unidos 2006/0234234; publicación de patente europea EP 1544308B1; y similares.

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

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

En un aspecto, las amplificaciones de multiplexación se llevan a cabo de modo que las cantidades relativas de secuencias en una población de partida son sustancialmente las mismas que las de en la población amplificada o amplicón. Es decir, las amplificaciones de multiplexación se llevan a cabo con un sesgado de amplificación mínimo entre secuencias miembro de una población de muestra. En una realización, tales cantidades relativas son sustancialmente las mismas si cada cantidad relativa de un amplicón se encuentra dentro de cinco veces su valor en la muestra de partida. En otra realización, tales cantidades relativas son sustancialmente las mismas si cada cantidad relativa de un amplicón se encuentra dentro de dos veces su valor en la muestra de partida. Como se describe más completamente a continuación, el sesgo de amplificación en PCR puede detectarse y corregirse usando técnicas convencionales de modo que un conjunto de cebadores de PCR puede seleccionarse para un repertorio predeterminado que proporciona amplificación no sesgada de cualquier muestra.

Con respecto a muchos repertorios basados en secuencias de TCR o BCR, una amplificación multiplex utiliza opcionalmente todos los segmentos V. La reacción se optimiza para intentar obtener una 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 "interaccionar", de modo que se amplifican los moldes que no coinciden perfectamente con él. Las condiciones se optimizan para que cada molde pueda amplificarse de manera similar, independientemente de qué cebador lo amplifique. En otras palabras, si hay dos moldes, luego, después de una amplificación de 1.000 veces, ambos moldes pueden amplificarse aproximadamente 1.000 veces y no importa que para uno de los moldes la mitad de los productos amplificados llevaran un cebador diferente debido a la interacción. En el análisis posterior de los datos de secuenciación, la secuencia del cebador se elimina del análisis y, por lo tanto, no importa qué cebador se use en la amplificación, siempre que los moldes se amplifiquen por igual.

En una realización, el sesgo de amplificación se puede evitar llevando a cabo una amplificación de dos etapas (como se describe en Faham y Willis, citado anteriormente) en el que se implementa un pequeño número de ciclos de amplificación en una primera, o primaria, etapa usando cebadores que tienen colas no complementarias con las secuencias diana. Las colas incluyen sitios de unión de cebador que se añaden a los extremos de las secuencias del amplicón primario de modo que tales sitios se usan en una segunda etapa de amplificación que usa solo un cebador único directo y un cebador único indirecto, eliminando, de este modo, una causa primaria del sesgo de amplificación. En algunas realizaciones, la PCR primaria tendrá un número suficientemente pequeño de ciclos (por ejemplo, 5-10) para minimizar la amplificación diferencial por los diferentes cebadores. Se lleva a cabo la amplificación secundaria con un par de cebadores, que minimiza la amplificación diferencial. En algunas realizaciones, un pequeño porcentaje, por ejemplo, uno por ciento, de la PCR primaria se lleva directamente a la PCR secundaria. En algunas realizaciones, un total de treinta y cinco ciclos (equivalentes a -28 ciclos sin la etapa de dilución de 100 veces) asignados entre una primera etapa y una segunda etapa generalmente son suficientes para mostrar una amplificación robusta independientemente de si los ciclos se dividen de la siguiente manera: 1 ciclo primario y 34 secundarios; o 25 primarios y 10 secundarios.

En resumen, El esquema de Faham y Willis (citado anteriormente) para amplificar los ácidos nucleicos (ARN) codificantes de IgH o TCRβ se ilustra en las figuras 1A-1C. Los esquemas de amplificación similares son fácilmente para otros segmentos de receptores inmunes, por ejemplo, Van Dongen y col., Leukemia, 17: 2257-2317 (2003), tales como, reordenamientos de IgH incompletos, IgK, Kde, IgL, TCRγ, TCRδ, Bcl1-IgH, Bcl2-IgH, y similares. Los ácidos nucleicos (1200) se extraen de linfocitos en una muestra y se combinan en una PCR con un cebador (1202) específico para la región C (1203) y cebadores (1212) específicos para las diversas regiones V (1206) de los genes de inmunoglobulina o TCR. Los cebadores (1212) tienen cada uno una cola idéntica (1214) que proporciona un sitio

de unión del cebador para una segunda etapa de amplificación. Como se ha mencionado anteriormente, el cebador (1202) se coloca adyacente a la unión (1204) entre la región C (1203) y la región J (1210). En la PCR, se genera un amplicón (1216) que contiene una parte de la región de codificación C (1203), región de codificación J (1210), región de codificación D (1208) y una parte de la región de codificación V (1206). El amplicón (1216) se amplifica aún más en una segunda etapa usando el cebador P5 (1222) y el cebador P7 (1220), cada uno de los cuales tiene colas (1225 y 1221/1223, respectivamente) diseñadas para su uso en un secuenciador de ADN Illumina. La cola (1221/1223) del cebador P7 (1220) incorpora opcionalmente la cola (1221) para marcar muestras separadas en el proceso de secuenciación. La amplificación de la segunda etapa produce un amplicón (1230) que puede usarse en un secuenciador de ADN Illumina.

Generación de lecturas de secuencia

Cualquier técnica de alto rendimiento para secuenciar ácidos nucleicos puede usarse en el procedimiento de la invención. Preferentemente, tal técnica tiene la capacidad de generar de un modo económico un volumen de datos de secuencia a partir de los cuales al menos 1.000 clonotipos pueden determinarse y, preferentemente, a partir de los cuales al menos 10.000 o 1.000.000 clonotipos pueden determinarse. Las técnicas de secuenciación de ADN incluyen reacciones de secuenciación didesoxi clásica (procedimiento Sanger) usando terminadores o cebadores marcados y separación por gel en plancha o capilar, secuenciación mediante síntesis usando nucleótidos marcados inversamente terminados, pirosecuenciación, secuenciación 454, hibridación específica de alelos con respecto a una librería de sondas de oligonucleótidos marcadas, secuenciación mediante síntesis usando hibridación específica de alelos con respecto a una librería de clones marcados que está seguida de ligación, control en tiempo real de la incorporación de nucleótidos marcados durante una etapa de polimerización, secuenciación de polonia y secuenciación SOLiD. La secuenciación de las moléculas separadas ha demostrado más recientemente mediante reacciones de extensión secuencial o única usando polimerasas o ligasas así como mediante hibridaciones diferenciales únicas o secuenciales con bibliotecas de sondas. Estas reacciones se han llevado a cabo sobre muchas secuencias clonales en paralelo incluyendo demostraciones en aplicaciones comerciales actuales de sobre 100 millones de secuencias en paralelo. Estos enfoques de secuenciación pueden, de este modo, usarse para estudiar el repertorio de receptor de linfocitos T (TCR) y/o receptor de linfocitos B (RLB). En un aspecto, se emplean procedimientos de alto rendimiento de secuenciación que comprenden una etapa de aislamiento espacial de moléculas individuales sobre una superficie sólida donde se secuencian en paralelo. Tales superficies sólidas pueden incluir superficies no porosas (tales como en secuenciación Solexa, por ejemplo, Bentley y col., Nature,456: 53-59 (2008) o secuenciación genómica completa, por ejemplo, Dr-manac y col., Science, 327: 78-81 (2010)), matrices de pocillos, que pueden incluir modelos unidos a perlas o partículas (tales como con 454, por ejemplo, Margulies y col., Nature, 437: 376-380 (2005) o secuenciación de Ion Torrent, publicación de patente de Estados Unidos 2010/0137143 o 2010/0304982), membranas micromecanizadas (como con secuenciación SMRT, por ejemplo, Eid y col., Science, 323: 133-138 (2009)) o matrices de perlas (como con secuenciación SOLiD o secuenciación de polonia, por ejemplo, Kim y col., Science, 316: 1481-1414 (2007)). En otro aspecto, tales procedimientos comprenden la amplificación de las moléculas aisladas o bien antes o bien después de que se aíslen espacialmente sobre una superficie sólida. Antes de la amplificación puede comprender la amplificación a base de emulsión, tal como PCR de emulsión o amplificación de círculo rodante. De particular interés es la secuenciación basada en Solexa donde las moléculas de molde individuales están aisladas espacialmente en una superficie sólida, después de lo cual se amplifican en paralelo mediante PCR en puente para formar poblaciones clonales separadas, o clústeres y, a continuación, se secuencian, como se describe en Bentley y col., (citado anteriormente) y en las instrucciones del fabricante (por ejemplo, Kit de preparación de muestras TruSeq™ y hoja de datos, Illumina, Inc., San Diego, CA, 2010); y adicionalmente en las siguientes referencias: Las patentes de Estados Unidos 6.090.592; 6,300,070; 7,115,400; y documento EP0972081B1. En una realización, moléculas individuales dispuestas y amplificadas sobre una superficie sólida forma clústeres en una densidad de al menos 10^5 clústeres por cm^2; o en una densidad de al menos $5x10^5$ por cm^2; o en una densidad de al menos 10^6 clústeres por cm^2. En una realización, se emplean químicas de secuenciación que tienen tasas de error relativamente altas. En dichas realizaciones, las puntuaciones de calidad promedio producidos por tales químicas son funciones monotónicamente decrecientes de las longitudes de lectura de secuencia. En una realización, dicha disminución corresponde al 0,5 por ciento de las lecturas de secuencia que tienen al menos un error en las posiciones 1-75; 1 por ciento de las lecturas de secuencia tienen al menos un error en las posiciones 76-100; y el 2 por ciento de las lecturas de secuencia tienen al menos un error en las posiciones 101-125.

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

realización, al menos una de las lecturas de secuencia directas y al menos una de las lecturas de secuencia indirectas tienen una región de solapamiento de modo que las bases de tal región de solapamiento se determinan mediante una relación complementaria inversa entre tales lecturas de secuencia. En aún otra realización, cada una de las regiones somáticamente reordenadas comprende una región V y una región J y la etapa de secuenciación incluye adicionalmente la determinación de una secuencia de cada una de las moléculas de ácido nucleico individuales a partir de una o más de sus lecturas de secuencia directas y al menos una lectura de secuencia indirecta que parte de una posición en una región J y se extiende en la dirección de su región V asociada. En otra realización, las moléculas individuales comprenden ácidos nucleicos seleccionados entre el grupo que consiste en moléculas de IgH completas, moléculas de IgH incompletas, moléculas completas de IgK, moléculas de IgK inactivas, moléculas de TCRβ, moléculas de TCRγ, moléculas de TCRδ completas y moléculas de TCRδ incompletas. En otra realización, la etapa de secuenciación comprende generar las lecturas de secuencia que tienen puntuaciones de calidad monotónicamente en disminución. Además de la última realización, las puntuaciones de calidad que disminuyen monotónicamente son tales que las lecturas de secuencia tienen tasas de error no mejores que las siguientes: el 0,2 por ciento de las lecturas de secuencia contienen al menos un error en las posiciones base de 1 a 50, del 0,2 al 1,0 por ciento de las lecturas de secuencia contienen al menos un error en las posiciones 51-75, del 0,5 al 1,5 por ciento de las lecturas de secuencia contienen al menos un error en las posiciones 76-100. En otra realización, el procedimiento anterior comprende las siguientes etapas: (a) obtener una muestra de ácido nucleico de linfocitos T y/o linfocitos B del individuo; (b) aislar espacialmente moléculas individuales derivadas de tal muestra de ácido nucleico, comprendiendo las moléculas individuales conjuntos anidados de modelos generados a partir de un ácido nucleico en la muestra y cada una conteniendo una región somáticamente reordenada o una porción de la misma, cada conjunto anidado siendo capaz de producir una pluralidad de lecturas de secuencia cada una extendiéndose en la misma dirección y cada una empezando desde una posición distinta sobre el ácido nucleico desde la cual el conjunto anidado se generó; (c) secuenciar dichas moléculas individuales espacialmente aisladas; y (d) determinar las abundancias de distintas secuencias de las moléculas de ácido nucleico a partir de la muestra de ácido nucleico para generar un perfil de clonotipo. En una realización, la etapa de secuenciación incluye la producción de una pluralidad de lecturas de secuencia para cada uno de los conjuntos anidados. En otra realización, cada una de las regiones somáticamente reordenadas comprende una región V y una región J, y cada una de la pluralidad de lecturas de secuencia se inicia desde una posición distinta en la región V y se extiende en la dirección de su región asociada J.

En un aspecto, para cada muestra de un individuo, la técnica de secuenciación usada en los procedimientos de la invención genera secuencias de al menos 1.000 clonotipos por ronda; en otro aspecto, tal técnica genera secuencias de al menos 10.000 clonotipos por tirada; en otro aspecto, tal técnica genera secuencias de al menos 100.000 clonotipos por tirada; en otro aspecto, tal técnica genera secuencias de al menos 500.000 clonotipos por tirada; y en otro aspecto, tal técnica genera secuencias de al menos 1.000.000 clonotipos por tirada. En otro aspecto adicional, tal técnica genera secuencias de entre 100.000 a 1.000.000 clonotipos por tirada por muestra individual.

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

Generación de clonotipos a partir de datos de secuencia

La construcción de clonotipos a partir de datos de lectura de secuencia se desvela en Faham y Willis (citados anteriormente). En resumen, la construcción de clonotipos a partir de datos de lectura de secuencia depende en parte del procedimiento de secuenciación utilizado para generar dichos datos, ya que los diferentes procedimientos tienen diferentes longitudes de lectura y calidad de datos esperadas. En un enfoque, se utiliza un secuenciador Solexa para generar datos de lectura de secuencia para el análisis. En una realización, se obtiene una muestra que proporciona al menos 0,5-1,0x10^6 linfocitos para producir al menos 1 millón de moléculas del molde, que después de la amplificación opcional puede producir un millón o más poblaciones clonales correspondientes de moléculas de molde (o grupos). Para la mayoría de los enfoques de secuenciación de alto rendimiento, incluyendo el enfoque de Solexa, tal sobremuestreo a nivel de grupo es deseable para que cada secuencia molde se determine con un alto grado de redundancia para aumentar la precisión de la determinación de la secuencia. Para implementaciones basadas en Solexa, preferentemente, la secuencia de cada molde independiente se determina 10 veces o más. Para otros enfoques de secuenciación con diferentes longitudes de lectura y calidad de datos esperadas, se pueden usar diferentes niveles de redundancia para una precisión comparable de la determinación de la secuencia. Los expertos en la materia reconocen que los parámetros anteriores, por ejemplo, el tamaño de la muestra, redundancia y similares, son opciones de diseño relacionadas con aplicaciones particulares.

En un aspecto, los clonotipos de cadenas IgH o cadenas TCRβ (ilustradas en la Figura 2A) se determinan mediante al menos una lectura de secuencia que comienza en su región C y se extiende en la dirección de su región V asociada (denominada en el presente documento "lectura C" (2304)) y al menos una lectura de secuencia que comienza en su región V y que se extiende en la dirección de su región J asociada (denominada en el presente

documento "lectura V" (2306)). Dichas lecturas pueden o no tener una región de superposición (2308) y dicha superposición puede o no abarcar la región de NDN (2315) como se muestra en la Figura 2A. La región de superposición (2308) puede estar completamente en la región J, completamente en la región NDN, completamente en la región V, o puede abarcar un límite de la región J-región NDN o un límite de la región V-región NDN, o ambos límites (como se ilustra en la Figura 2A). Típicamente, tales lecturas de secuencia se generan extendiendo cebadores de secuencia, por ejemplo, (2302) y (2310) en la Figura 2A, con una polimerasa en una reacción de secuenciación por síntesis, por ejemplo, Metzger, Nature Reviews Genetics, 11: 31-46 (2010); Fuller y col., Nature Biotechnology, 27: 1013-1023(2009). Los sitios de unión para los cebadores (2302) y (2310) están predeterminados, para que puedan proporcionar un punto de partida o punto de anclaje para la alineación inicial y el análisis de las lecturas de secuencia. En una realización, una lectura C se coloca de modo que abarque la región D y/o NDN de la cadena de IgH e incluya una parte de la región V adyacente, por ejemplo, como se ilustra en las Figuras 2A y 2B. En un aspecto, la superposición de la lectura V y la lectura C en la región V se utiliza para alinear las lecturas entre sí. En otras realizaciones, dicha alineación de las lecturas de secuencia no es necesaria, para que una lectura de V solo sea lo suficientemente larga como para identificar la región V particular de un clonotipo. Este último aspecto se ilustra en la Figura 2B. La lectura de secuencia (2330) se utiliza para identificar una región V, con o sin superposición de otra secuencia de lectura, y otra secuencia de lectura (2332) atraviesa la región NDN y se utiliza para determinar la secuencia de la misma. La porción (2334) de la lectura de secuencia (2332) que se extiende hacia la región V se usa para asociar la información de secuencia de la lectura de secuencia (2332) con la de la lectura de secuencia (2330) para determinar un clonotipo. Para algunos procedimientos de secuenciación, tales como enfoques base por base como el procedimiento de secuenciación de Solexa, el tiempo de ejecución de la secuenciación y los costes de los reactivos se reducen minimizando el número de ciclos de secuenciación en un análisis. Opcionalmente, como se ilustra en la Figura 2A, el amplicón (2300) se produce con el marcador de la muestra (2312) para distinguir entre clonotipos que se originan de diferentes muestras biológicas, por ejemplo, diferentes pacientes. El marcador de la muestra (2312) puede identificarse hibridando un cebador a la región de unión del cebador (2316) y extendiéndola (2314) para producir una lectura de secuencia a través del marcador (2312), a partir de la cual se descodifica el marcador de la muestra (2312).

En un aspecto, las secuencias de los clonotipos pueden determinarse combinando información de una o más lecturas de secuencia, por ejemplo, a lo largo de las regiones V(D)J de las cadenas seleccionadas. En otro aspecto, las secuencias de los clonotipos se determinan combinando 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 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, primero se generan moldes separados amplificando las moléculas de muestra con cebadores seleccionados para las diferentes posiciones de las lecturas de secuencia. Este concepto se ilustra en la Figura 3A donde los cebadores (3404, 3406 y 3408) se emplean para generar amplicones (3410, 3412 y 3414, respectivamente) en una sola reacción. Dichas amplificaciones pueden llevarse a cabo en la misma reacción o en reacciones separadas. En un aspecto, cada vez que se emplea PCR, se utilizan reacciones de amplificación separadas para generar los moldes separados que, a su vez, se combinan y se utilizan para generar múltiples lecturas de secuencia a lo largo de la misma cadena. Este último enfoque es preferible para evitar la necesidad de equilibrar las concentraciones de cebadores (y/u otros parámetros de reacción) para garantizar la amplificación igual de los moldes múltiples (a veces denominados en el presente documento "amplificación equilibrada" o "amplificación sin sesgo"). La generación de moldes en reacciones separadas se ilustra en las Figuras 3B-3C. Allí, una muestra que contiene IgH (3400) se divide en tres porciones (3470, 3472 y 3474) que se añaden a PCR separadas usando cebadores de la región J (3401) y cebadores de la región V (3404, 3406 y 3408, respectivamente) para producir amplicones (3420, 3422 y 3424, respectivamente). Los últimos amplicones se combinan (3478) después en PCR secundaria (3480) usando cebadores P5 y P7 para preparar los moldes (3482) para PCR puente y secuenciación en un secuenciador Illumina GA, o instrumento similar.

Las lecturas de secuencia pueden tener una amplia variedad de longitudes, dependiendo en parte de la técnica de secuenciación empleada. Por ejemplo, para algunas técnicas, pueden surgir varias compensaciones en su implementación, por ejemplo, (i) el número y la duración de las lecturas de secuencia por molde y (ii) el coste y la duración de una operación de secuenciación. En una realización, las lecturas de secuencia están en el intervalo de 20 a 200 nucleótidos; en otra realización, las lecturas de secuencia están en un intervalo de 30 a 200 nucleótidos; en aún otra realización, las lecturas de secuencia están en el intervalo de 30 a 120 nucleótidos. En una realización, se generan de 1 a 4 lecturas de secuencia para determinar la secuencia de cada clonotipo; en otra realización, se generan de 2 a 4 lecturas de secuencia para determinar la secuencia de cada clonotipo; y, en otra realización, se generan de 2 a 3 lecturas de secuencia para determinar la secuencia de cada clonotipo. En algunas realizaciones, se utiliza una pluralidad de lecturas de secuencia para generar cada clonotipo; en algunas realizaciones, la pluralidad de lecturas de secuencia utilizadas para generar un clonotipo es de al menos 10; en otras realizaciones, la pluralidad de lecturas de secuencia utilizadas para generar un clonotipo es de al menos 20. En algunas realizaciones, la pluralidad de lecturas de secuencia utilizadas para generar un clonotipo es un número requerido para determinar que diferentes clonotipos son diferentes con un nivel de confianza de al menos 99 por ciento; en otras realizaciones, la pluralidad de lecturas de secuencia utilizadas para generar un clonotipo es un número requerido para determinar que los diferentes clonotipos son diferentes con un nivel de confianza de al menos 99,9 por ciento. Como se señala a continuación, los clonotipos pueden generarse a partir de lecturas de secuencia

mediante una etapa de combinación, o como a veces se hace referencia en el presente documento, una etapa de fusión, lecturas de secuencia de acuerdo con un procedimiento (como se ilustra en la divulgación siguiente) que tiene en cuenta las tasas de error de secuenciación y/o amplificación, las frecuencias de las lecturas de secuencia, las diferencias de secuencia y similares. En las realizaciones anteriores, los números dados son exclusivos de las lecturas de secuencia utilizadas para identificar muestras de diferentes individuos. Las longitudes de las diversas lecturas de secuencia utilizadas en las realizaciones descritas a continuación también pueden variar en función de la información que se pretende capturar mediante la lectura; por ejemplo, la ubicación y la longitud iniciales de una lectura de secuencia pueden diseñarse para proporcionar la longitud de una región NDN así como su secuencia de nucleótidos; por tanto, se seleccionan lecturas de secuencia que abarcan toda la región NDN. En otros aspectos, una o más lecturas de secuencia que, en combinación (pero no por separado), abarcan una región D y/o NDN son suficientes.

En otro aspecto, las secuencias de clonotipos se determinan en parte alineando las lecturas de secuencia 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 mediante la determinación de la base sin alineación con las secuencias de referencia, tal como en la región NDN altamente variable. Se pueden aplicar diversos algoritmos de alineación a las lecturas de secuencia y secuencias de referencia. Por ejemplo, se dispone de directrices para seleccionar los procedimientos de alineación en Batzoglou, Briefings in Bioinformatics, 6: 6-22(2005). En un aspecto, siempre que las lecturas V o C (como se ha mencionado anteriormente) estén alineadas con las secuencias de referencia de las regiones V y J, se emplea un algoritmo de búsqueda de árbol, por ejemplo, como se describe generalmente en Gusfield (citado anteriormente) y Cormen y col., Introduction to Algorithms, Tercera edición (The MIT Press, 2009).

La construcción de los clonotipos de IgH a partir de las lecturas de secuencia se caracteriza por al menos dos factores: i) la presencia de mutaciones somáticas que dificultan la alineación, y ii) la región NDN es más grande, por lo que a menudo no es posible mapear una porción del segmento V a la lectura C. En un aspecto, este problema se supera mediante el uso de una pluralidad de conjuntos de cebadores para generar lecturas de V, que se encuentran en diferentes lugares a lo largo de la región V, preferentemente de modo que los sitios de unión del cebador no se solapen y se separen, y con al menos un sitio de unión del cebador adyacente a la región NDN, por ejemplo, en una realización de 5 a 50 bases de la unión V-NDN, o en otra realización de 10 a 50 bases de la unión V-NDN. La redundancia de una pluralidad de conjuntos de cebadores minimiza el riesgo de no detectar un clonotipo debido a un fallo de uno o dos cebadores que tienen sitios de unión afectados por mutaciones somáticas. Además, la presencia de al menos un sitio de unión de cebador adyacente a la región NDN hace que sea más probable que una lectura de V se superponga con la lectura de C y, por lo tanto, extienda efectivamente la longitud de la lectura de C. Esto permite la generación de una secuencia continua que abarca todos los tamaños de regiones NDN y que también puede mapear sustancialmente todas las regiones V y J en ambos lados de la región NDN. Las realizaciones para llevar a cabo tal esquema se ilustran en las Figuras 3A y 3D. En la figura 3A, una muestra que comprende cadenas de IgH (3400) se secuencia mediante la generación de una pluralidad de amplicones para cada cadena al amplificar las cadenas con un único conjunto de cebadores de la región J (3401) y una pluralidad (se muestran tres) de conjuntos de cebadores de la región V (3402) (3404, 3406, 3408) para producir una pluralidad de amplicones anidados (por ejemplo, 3410, 3412, 3416) que comprenden todos la misma región NDN y tienen diferentes longitudes que abarcan porciones sucesivamente más grandes (3411, 3413, 3415) de la región V (3402). Los miembros de un conjunto anidado se pueden agrupar después de la secuencia observando la identificación (o identidad sustancial) de sus regiones NDN, J y/o C respectivas, permitiendo así la reconstrucción de un segmento V(D)J más largo de lo que sería el caso de otra manera para una plataforma de secuenciación con longitud de lectura limitada y/o calidad de secuencia. En una realización, la pluralidad de conjuntos de cebadores puede ser un número en el rango de 2 a 5. En otra realización, la pluralidad es 2-3; y aún otra realización, 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 pueden o no ser las mismas. En una realización, el cebador más cercano a la región NDN tiene una concentración más alta 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 donde se emplea una pluralidad de tres cebadores, Se utiliza una relación de concentración de 60:20:20. Uno o más cebadores (por ejemplo, 3435 y 3437 en la figura 3D) adyacente a la región NDN (3444) puede usarse para generar una o más lecturas de secuencia (por ejemplo, 3434 y 3436) que se superponen a la lectura de secuencia (3442) generada por el cebador de la región J (3432), mejorando así la calidad de las llamadas de base en la región de superposición (3440). Las lecturas de secuencia de la pluralidad de cebadores pueden superponerse o no al sitio de unión del cebador adyacente aguas abajo y/o la lectura de secuencia aguas abajo adyacente. En una realización, la secuencia lee proximal a la región NDN (por ejemplo, 3436 y 3438) pueden usarse para identificar la región V particular asociada con el clonotipo. Dicha pluralidad de cebadores reduce la probabilidad de amplificación incompleta o fallida en caso de que uno de los sitios de unión del cebador se hipermuta durante el desarrollo de inmunoglobulina. También aumenta la probabilidad de que la diversidad introducida por la hipermutación de la región V se capture en una secuencia de clonotipo. Se puede realizar una PCR secundaria para preparar los amplicones anidados para la secuenciación, por ejemplo, amplificando con los cebadores P5 (3401) y P7 (3404, 3406, 3408) como se ilustra para producir amplicones (3420, 3422 y 3424), que pueden distribuirse como moléculas individuales en una superficie sólida, donde se amplifican aún más por PCR de puente, o como técnica.

Hipermutaciones somáticas. En una realización, los clonotipos basados en IgH que han sufrido 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, generalmente V, J o C) y eso está presente en un número estadísticamente significativo de lecturas. En una realización, las lecturas de C se pueden usar para encontrar mutaciones somáticas con respecto al segmento J mapeado y, asimismo, las lecturas de V para el segmento V. Solo se utilizan partes de las lecturas C y V que se asignan directamente a segmentos J o V o que están dentro de la extensión del clonotipo hasta el límite de NDN. De este modo, se evita la región NDN y no se usa la misma "información de secuencia" para el hallazgo de mutaciones que se usó previamente para la determinación del clonotipo (para evitar clasificar erróneamente como mutaciones nucleótidos que en realidad son solo regiones NDN recombinadas diferentes). Para cada tipo de segmento, el segmento mapeado (alelo principal) se usa como un andamio y se consideran todas las lecturas que se han mapeado a este alelo durante la fase de mapeo de lectura. Cada posición de las secuencias de referencia donde se ha mapeado al menos una lectura se analiza para detectar mutaciones somáticas. En una realización, los criterios para aceptar una base sin referencia como una mutación válida incluyen los siguientes: 1) al menos N lecturas con la base de mutación dada, 2) al menos una fracción dada de lecturas N/M (donde M es el número total de lecturas mapeadas en esta posición base) y 3) un corte estadístico basado en la distribución binomial, la puntuación Q promedio de las N lecturas en la base de mutación, así como el número (M-N) de lecturas con una base sin mutación. Preferentemente, los parámetros anteriores se seleccionan de modo que la tasa de descubrimiento falso de mutaciones por clonotipo sea menor que 1 en 1000 y, más preferentemente, menos de 1 en 10000.

Se espera que el error de PCR se concentre en algunas bases que se mutaron en los primeros ciclos de PCR. Se espera que el error de secuencia se distribuya en muchas bases, aunque sea totalmente aleatorio, ya que es probable que el error tenga algunos sesgos sistemáticos. Se supone que algunas bases tendrán un error de secuencia a una velocidad mayor, es decir 5 % (5 veces el promedio). Dados estos supuestos, el error de secuenciación se convierte en el tipo de error dominante. Distinguir los errores de PCR de la aparición de clonotipos altamente relacionados desempeñará un papel en el análisis. Dada la importancia biológica para determinar que hay dos o más clonotipos altamente relacionados, se adopta un enfoque conservador para hacer tales llamadas. Se considera la detección de suficientes clonotipos menores para asegurarse con alta confianza (es decir, 99,9 %) de que hay más de un clonotipo. Por ejemplo, de clonotipos que están presentes en 100 copias/1.000.000, la variante menor se detecta 14 o más veces para que se designe como un clonotipo independiente. De forma similar, para los clonotipos presentes en 1.000 copias/1.000.000, la variante menor puede detectarse 74 o más veces para su diseño como un clonotipo independiente. Este algoritmo se puede mejorar utilizando la puntuación de calidad base que se obtiene con cada base secuenciada. Si la relación entre la puntuación de calidad y la tasa de error se valida anteriormente, en lugar de emplear la tasa de error conservadora del 5 % para todas las bases, la puntuación de calidad se puede usar para decidir la cantidad de lecturas que deben estar presentes para consultar un clonotipo independiente. La mediana de la puntuación de calidad de la base específica en todas las lecturas se puede utilizar, o más rigurosamente, la probabilidad de ser un error se puede calcular dada la puntuación de calidad de la base específica en cada lectura, y luego las probabilidades se pueden combinar (suponiendo independencia) para estimar el número probable de error de secuencia para esa base. Como resultado, existen diferentes umbrales para rechazar la hipótesis del error de secuenciación para diferentes bases con diferentes puntuaciones de calidad. Por ejemplo, para un clonotipo presente en 1.000 copias/1.000.000, la variante menor se designa como independiente cuando se detecta 22 y 74 veces si la probabilidad de error fue 0,01 y 0,05, respectivamente.

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

La nube de lecturas de secuencia que rodea cada clonotipo candidato puede modelarse usando la distribución

binomial y un modelo simple para la probabilidad de un error de base única. Este último modelo de error puede inferirse a partir del mapeo de segmentos V y J o a partir del algoritmo de hallazgo de clonotipo mismo, a través de autoconsistencia y convergencia. Se construye un modelo para la probabilidad a una secuencia 'nube' dada Y con recuento de lectura C2 y errores E (con respecto a la secuencia C) que es parte de una verdadera secuencia de clonotipo X con un recuento de lectura perfecto C1 con el modelo nulo que X es el único clonotipo verdadero en esta región de espacio de secuencia. Se toma la decisión de si fusionar o no la secuencia Y en el clonotipo X de acuerdo con los parámetros C1, C2 y E. Para cualquier C1 y E dado se precalcula un valor máx. C2 para decidir fusionar la secuencia Y. Los valores máx. para C2 se escogen de modo que la probabilidad de no realizar la fusión Y con la hipótesis nula de que Y es parte del clonotipo X es menos que algún valor P después de integrar sobre todas las secuencias posibles Y con error E en los alrededores de la secuencia X. El valor P controla el comportamiento del algoritmo y hace que la fusión sea más o menos permisiva.

Si una secuencia Y no se fusiona en el clonotipo X porque su recuento de lectura está por encima del umbral C2 para fusionarse en el clonotipo X, se convierte en un candidato para sembrar clonotipos separados. Un algoritmo que implementa dichos principios asegura que cualquier otra secuencia Y2, Y3, etc. que son las "más cercanas" a esta secuencia Y (que se han considerado independientes de X) no se agreguen en X. Este concepto de "cercanía" incluye tanto recuentos de errores con respecto a Y y X como el recuento de lectura absoluto de X e Y, es decir, se modela del mismo modo que el anterior modelo para la nube de secuencias de error alrededor del clonotipo X. De este modo, las secuencias "nube" pueden atribuirse adecuadamente a su clonotipo correcto si parecen estar "cerca" más de un clonotipo.

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

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

Clonotipos relacionados

Con frecuencia, los linfocitos producen clonotipos relacionados. Es decir, pueden existir o desarrollarse linfocitos múltiples que producen clonotipos cuyas secuencias son similares. Esto puede deberse a diversos mecanismos, tales como la hipermutación en el caso de las moléculas de IgH. Como otro ejemplo, en el cáncer, tal como las neoplasias linfoides, un solo progenitor de linfocitos puede dar lugar a muchos descendientes de linfocitos relacionados, poseyendo cada uno y/o expresando un TCR o BCR ligeramente diferente, y, por lo tanto, un clonotipo diferente, debido a mutaciones somáticas relacionadas con el cáncer, tales como sustituciones de bases, reordenamientos aberrantes, o similares. Un conjunto de tales clonotipos relacionados se denomina en el presente documento "clan". En algún caso, los clonotipos de un clan pueden surgir de la mutación de otro miembro del clan. Tal clonotipo de "descendencia" se puede denominar clonotipo filogénico. Los clonotipos dentro de un clan pueden identificarse mediante una o más medidas de relación con un clonotipo padre o entre sí. En una realización, os clonotipos pueden agruparse en el mismo clan por porcentaje de homología, tal como se explica más completamente a continuación. En otra realización, los clonotipos pueden asignarse a un clan mediante uso común de las regiones V, regiones J y/o regiones NDN. Por ejemplo, un clan puede definirse por clonotipos que tienen las mismas regiones J y ND, pero regiones V diferentes; o puede definirse por clonotipos que tienen las mismas regiones V y J (incluyendo mutaciones de sustituciones de bases idénticas) pero con regiones NDN diferentes; o puede definirse por un clonotipo que ha sufrido una o más inserciones y/o deleciones de 1-10 bases, o de 1-5 bases, o de 1-3 bases, para generar miembros del clan. En otra realización, los miembros de un clan se determinan de la siguiente manera.

Los clonotipos se asignan al mismo clan si cumplen los siguientes criterios: i) se mapean en los mismos segmentos de referencia V y J, produciéndose los mapeos en las mismas posiciones relativas en la secuencia del clonotipo, y ii) sus regiones NDN son sustancialmente idénticas. "Sustancial" en referencia a la pertenencia al clan significa que se

permiten algunas pequeñas diferencias en la región NDN porque pueden haberse producido mutaciones somáticas en esta región. Preferentemente, en una realización, para evitar denominar falsa a una mutación en la región NDN, si una sustitución de base se acepta como una mutación relacionada con el cáncer depende directamente del tamaño de la región NDN del clan. Por ejemplo, un procedimiento puede aceptar un clonotipo como un miembro del clan si tiene una diferencia de una base de la o las secuencias de NDN del clan como una mutación relacionada con el cáncer si la longitud de la o las secuencias de NDN del clan tiene m nucleótidos o más, por ejemplo, 9 nucleótidos o más, de lo contrario, no se acepta, o si tiene una diferencia de dos bases con respecto a la o las secuencias NDN del clan como mutaciones relacionadas con el cáncer si la longitud de la o las secuencias NDN del clan tiene n nucleótidos o más, por ejemplo, 20 nucleótidos o más, de lo contrario no se acepta, En otra realización, los miembros de un clan se determinan utilizando los siguientes criterios: (a) V lee mapas a la misma región V, (b) C lee mapas en la misma región J, (c) la región NDN sustancialmente idéntica (como se ha descrito anteriormente), y (d) la posición de la región NDN entre el límite V-NDN y el límite J-NDN es la misma (o equivalente, el número de adiciones de bases aguas abajo a D y el número de las adiciones de base aguas arriba a D son las mismas). Los clonotipos de una sola muestra pueden agruparse en clanes y los clanes de muestras sucesivas adquiridas en diferentes momentos pueden compararse entre sí. En particular, en un aspecto, los clanes que contienen clonotipos correlacionados con una enfermedad, tal como una neoplasia linfoide, se identifican a partir de los clonotipos de cada muestra y se comparan con los de la muestra inmediatamente anterior para determinar el estado de la enfermedad, tales como, remisión continua, recaída incipiente, evidencia de una evolución clonal adicional, o similar. Como se usa en el presente documento, "tamaño" en referencia a un clan significa el número de clonotipos en el clan.

Como se ha mencionado anteriormente, en un aspecto, los procedimientos desvelados en el presente documento controlan un nivel de un clan de clonotipos en lugar de un clonotipo individual. Esto se debe a los fenómenos de la evolución clonal, por ejemplo, Campbell y col., Proc. Natl. Acad. Sci., 105: 13081-13086 (2008); Gerlinger y col., Br. J. El cáncer, 103: 1139-1143 (2010). La secuencia de un clon que está presente en la muestra de diagnóstico puede no ser exactamente la misma que la de una muestra posterior, tal como una tomada en una recaída de la enfermedad. Por lo tanto, si se sigue la secuencia de clonotipo exacta que coincide con la secuencia de muestra de diagnóstico, la detección de una recaída podría fallar. Tal clon evolucionado se detecta e identifica fácilmente mediante secuenciación. Por ejemplo, muchos de los clones evolucionados emergen por reemplazo de la región V (llamado reemplazo de VH). Estos tipos de clones evolucionados se pierden con las técnicas de PCR en tiempo real, ya que los cebadores se dirigen al segmento V incorrecto. Sin embargo, dado que la unión D-J permanece intacta en el clon evolucionado, Se puede detectar e identificar mediante la secuenciación de moléculas individuales aisladas espacialmente. Adicionalmente, la presencia de estos clonotipos relacionados con una frecuencia apreciable en la muestra de diagnóstico aumenta la probabilidad de la relevancia del clonotipo. Del mismo modo, el desarrollo de hipermutaciones somáticas en la secuencia del receptor inmune puede interferir con la detección de la sonda de PCR en tiempo real, pero los algoritmos apropiados aplicados a la lectura de secuenciación (como se describe anteriormente) aún pueden reconocer un clonotipo como un clonotipo en evolución. Por ejemplo, las hipermutaciones somáticas en los segmentos V o J pueden reconocerse. Esto se realiza mediante el mapeo de los clonotipos a las secuencias V y J de la línea germinal más cercana. Las diferencias de las secuencias de la línea germinal se pueden atribuir a las hipermutaciones somáticas. Por lo tanto, los clonotipos que evolucionan a través de hipermutaciones somáticas en los segmentos V o J se pueden detectar e identificar fácilmente. Se pueden predecir hipermutaciones somáticas en la región NDN. Cuando el segmento D restante es lo suficientemente largo como para ser reconocido y mapeado, cualquier mutación somática en ella puede reconocerse fácilmente. Las hipermutaciones somáticas en las bases N + P (o en el segmento D que no es mapeable) no pueden reconocerse con certeza ya que estas secuencias pueden modificarse en células recién recombinadas que pueden no ser la progenie del clonotipo canceroso. Sin embargo, los algoritmos se construyen fácilmente para identificar cambios de base que tienen una alta probabilidad de ser debidos a una mutación somática. Por ejemplo, un clonotipo con los mismos segmentos V y J y una diferencia de 1 base en la región NDN del clon o los clones originales tiene una alta probabilidad de ser el resultado de la recombinación somática. Esta probabilidad se puede aumentar si hay otras hipermutaciones somáticas en los segmentos V y J porque esto identifica este clonotipo específico como uno que ha sido objeto de hipermutación somática. Por lo tanto, la probabilidad de que un clonotipo sea el resultado de la hipermutación somática de un clonotipo original se puede calcular utilizando varios parámetros: el número de diferencias en la región NDN, la longitud de la región NDN, así como la presencia de otras hipermutaciones somáticas en los segmentos V y/o J.

Los datos de evolución clonal pueden ser informativos. Por ejemplo, si el clon principal es un clon evolucionado (uno que estaba ausente previamente y, por lo tanto, no registrado anteriormente), esto es una indicación de que el tumor ha adquirido nuevos cambios genéticos con posibles ventajas selectivas. Esto no quiere decir que los cambios específicos en el receptor de las células inmunes son la causa de la ventaja selectiva, sino que pueden representar un marcador para ello. Los tumores cuyos clonotipos han evolucionado pueden estar potencialmente asociados con un pronóstico diferencial. En un aspecto, un clonotipo o clonotipos que se usan como biomarcador de una enfermedad específico del paciente, tal como una neoplasia linfoide, por ejemplo, una leucemia, incluye clonotipos no registrados previamente que son mutantes somáticos del clonotipo o clonotipos que se controlan. En otro aspecto, siempre que cualquier clonotipo no registrado anteriormente sea al menos el noventa por ciento homólogo a un clonotipo o grupo de clonotipos existente que sirva como biomarcadores específicos del paciente, dicho clonotipo homólogo se incluye con o en el grupo de clonotipos que se están controlando. Es decir, si se identifican

uno o más clonotipos específicos del paciente en una neoplasia linfoide y se utilizan para controlar periódicamente la enfermedad (por ejemplo, realizando mediciones en muestras de sangre adquiridas de forma menos invasiva) y si en el curso de una de tales mediciones se realiza una nueva (previamente no registrada) se detecta el clonotipo que es una mutación somática de un clonotipo del conjunto actual, se añade al conjunto de clonotipos específicos del paciente que se controlan para mediciones posteriores. En una realización, si dicho clonotipo previamente no registrado es al menos noventa por ciento homólogo con un miembro del conjunto actual, se añade al conjunto de biomarcadores de clonotipo específicos del paciente para la siguiente prueba realizada en el paciente; es decir, dicho clonotipo no registrado anteriormente se incluye en el clan del miembro del conjunto actual de clonotipos de los que deriva (basado en el análisis anterior de los datos del clonotipo). En otra realización, dicha inclusión se lleva a cabo si el clonotipo no registrado anteriormente es al menos noventa y cinco por ciento homólogo con un miembro del conjunto actual. En otra realización, dicha inclusión se lleva a cabo si el clonotipo previamente no registrado es al menos noventa y ocho por ciento homólogo con un miembro del conjunto actual.

También es posible que una célula evolucione a través de un proceso que reemplaza la región NDN pero conserva los segmentos V y J junto con sus mutaciones acumuladas. Dichas células pueden identificarse como clonotipos de cáncer no registrados previamente mediante la identificación del segmento V y J común, siempre que contengan un número suficiente de mutaciones para reducir la posibilidad de que estas mutaciones deriven de forma independiente. Una restricción adicional puede ser que la región NDN sea de tamaño similar al clon secuenciado previamente.

Definiciones

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

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

"Amplicón" se refiere al producto de una reacción de amplificación de polinucleótidos; es decir, una población clonal de polinucleótidos, que pueden ser monocatenarios o bicatenarios, que se replican a partir de una o más secuencias de partida. La una o más secuencias de partida puede ser una o más copias de la misma secuencia o pueden ser una mezcla de distintas secuencias. Preferentemente, los amplicones se forman mediante la amplificación de una única secuencia de partida. Los amplicones pueden producirse mediante diversas reacciones de amplificación cuyos productos comprenden réplicas de uno o más ácidos nucleicos de partida o diana. En un aspecto, las reacciones de amplificación que producen amplicones son "dirigidas por modelo" en ese emparejamiento de bases de reactivos, o bien nucleótidos bien oligonucleótidos, tienen complementos en un polinucleótido modelo que se requieren para la creación de productos de reacción. En un aspecto, las reacciones dirigidas por modelo son extensiones de cebador con una polimerasa de ácido nucleico o ligaduras de oligonucleótidos con una ligasa de ácido nucleico. Tales reacciones incluyen, pero sin limitación, reacciones en cadena de la polimerasa (PCR), reacciones de la polimerasa lineales, amplificación a base de secuencias de ácidos nucleicos (NASBA), amplificaciones de círculo rodante y similares, desveladas en las siguientes referencias: Mullis y col., las patentes de Estados Unidos 4.683.195; 4,965,188; 4,683,202; 4.800.159 (ADN); Gelfand y col., patente de los EE.UU. 5.210.015 (PCR en tiempo real con sondas "taqman"); Wittwer y col., la patente de Estados Unidos 6.174.670; Kacian y col., patente de los EE.UU. 5.399.491 ("NASBA"); Lizardi, la patente de Estados Unidos 5.854.033; Aono y col., publ. de patente japonesa, JP 4-262799 (amplificación de círculo rodante); y similares. En un aspecto, los amplicones se producen mediante PCR. Una reacción de amplificación puede ser una amplificación en "tiempo real" si hay disponible una química de detección que permite que un producto de reacción se mida según progresa la reacción de amplificación, por ejemplo, "PCR en tiempo real" que se describe a continuación o "NASBA en tiempo real" como se describe en Leone y col., Nucleic Acids Research, 26: 2150-2155 (1998), y referencias similares. Como se usa en el presente documento, el término "amplificar" se refiere a realizar una reacción de amplificación. Una "mezcla de reacción" se refiere a una solución que contiene todos los reactivos necesarios para llevar a cabo una reacción, que puede incluir, pero sin limitación, agentes tamponantes para mantener el pH en un nivel seleccionado durante una reacción, sales, cofactores, neutralizantes y similares.

"Clonalidad" tal como se utiliza en el presente documento se refiere al grano en el cual la distribución de las abundancias de clonotipo entre clonotipos de un repertorio se sesgan en un único o unos pocos clonotipos. Aproximadamente, la clonalidad es una medida inversa de la diversidad de clonotipo. Existen muchas medidas o estadísticas disponibles a partir de la ecología que describen las relaciones de abundancia de especies que

pueden usarse para medidas de clonalidad, por ejemplo, Capítulos 17 y 18, en Pielou, An Introduction to Mathematical Ecology, (Wiley-Interscience, 1969). En un aspecto, una medida de clonalidad es una función de un perfil de clonotipo (es decir, el número de clonotipos distintos detectados y sus abundancias), de modo que después de medir un perfil de clonotipo, se puede informatizar la clonalidad a partir de este para proporcionar un único número. Una medida de clonalidad es la medida de Simpson, que es simplemente la probabilidad de que dos clonotipos extraídos aleatoriamente sean los mismos. Otras medidas de clonalidad incluyen medidas a base de información y el índice de diversidad de McIntosh, que se desvela en Pielou (citado anteriormente).

"Clonotipo" se refiere a una secuencia de nucleótido recombinado de un linfocito que codifica un receptor inmunitario o una porción del mismo. Más particularmente, clonotipo significa una secuencia de nucleótidos recombinada de un linfocito T o linfocito B que codifica un receptor de linfocitos T (TCR) o un receptor de linfocitos B (BCR), o una porción de la misma. En diversas realizaciones, los clonotipos pueden codificas toda o una porción de un reordenamiento de VDJ de IgH, un reordenamiento de DJ de IgH, un reordenamiento de VJ de IgK, un reordenamiento de VJ de IgL, un reordenamiento de VDJ de TCR β, un reordenamiento de DJ de TCR β, un reordenamiento de VJ de TCR α, un reordenamiento de VJ de TCR γ, un reordenamiento de VDJ de TCR δ, un reordenamiento de VD de TCR δ, un reordenamiento de Kde-V o similares. Los clonotipos también pueden codificar regiones de puntos de rotura de translocación que implican genes de receptores inmunitarios, tales como Bcl1-IgH o Bcl1-IgH. En un aspecto, los clonotipos tiene secuencias que son lo suficientemente largas para representar o reflejar la diversidad de moléculas inmunitarias de las que derivan; por consiguiente, los clonotipos pueden variar ampliamente en longitud. En algunas realizaciones, los clonotipos longitudes en el intervalo de 25 a 400 nucleótidos; en otras realizaciones, los clonotipos tienen longitudes en el intervalo de 25 a 200 nucleótidos.

"Perfil de clonotipo" se refiere a una enumeración de distintos clonotipos y sus abundancias relativas que derivan de una población de linfocitos. Típicamente, la población de linfocitos se obtiene a partir de una muestra de tejido. El término "perfil de clonotipo" se refiere a, pero más general que, el concepto de inmunología de "repertorio" inmunitario tal como se describe en las referencias, tales como las siguientes: Arstila y col., Science, 286: 958-961 (1999); Yassai y col., Immunogenetics, 61: 493-502 (2009); Kedzierska y col., Mol. Immunol., 45(3): 607-618 (2008); y similares. La expresión "perfil de clonotipo" incluye una amplia variedad de listados y abundancias de ácidos nucleicos que codifican receptores inmunitarios reordenados, que pueden derivarse de subconjuntos seleccionados de linfocitos (por ejemplo, linfocitos de infiltración tisular, subconjuntos inmunofenotípicos, o similares) o que pueden codificar porciones de receptores inmunitarios que tienen diversidad reducida en comparación con receptores inmunitarios completos. En algunas realizaciones, los perfiles de clonotipo puede comprender al menos 10^3 de clonotipos distintos; en otras realizaciones, los perfiles de clonotipo puede comprender al menos 10^4 de clonotipos distintos; en otras realizaciones, los perfiles de clonotipo puede comprender al menos 10^5 de clonotipos distintos; en otras realizaciones, los perfiles de clonotipo puede comprender al menos 10^6 de clonotipos distintos. En dichas realizaciones, tales perfiles de clonotipo pueden comprender adicionalmente abundancias o frecuencias relativas de cada uno de los distintos clonotipos. En un aspecto, un perfil de clonotipo es un conjunto de distintas secuencias de nucleótidos recombinados (con sus abundancias) que codifican receptores de linfocitos T (TCR) o receptores de linfocitos B (RLB) o fragmentos de los mismos, respectivamente, en una población de linfocitos de un individuo, en donde las secuencias de nucleótidos del conjunto tienen una correspondencia uno-a-uno con distintos linfocitos o sus subpoblaciones clonales para sustancialmente todos los linfocitos de la población. En un aspecto, los segmentos de ácidos nucleicos que definen clonotipos se seleccionan de modo que su diversidad (es decir, el número de distintas secuencias de ácidos nucleicos en el conjunto) es lo suficientemente grande de modo que sustancialmente cada linfocito T o linfocito B o clon del mismo en un individuo porta una única secuencia de ácido nucleico de tal repertorio. Es decir, preferentemente cada clon distinto de una muestra tiene un clonotipo distinto. La población de linfocitos que se corresponden con un repertorio puede ser linfocitos B circulantes o pueden ser linfocitos T circulantes o pueden ser subpoblaciones de o bien las poblaciones anteriores, incluyendo, aunque no de forma limitativa, linfocitos T CD4+ o linfocitos T CD8+ o bien otras poblaciones definidas mediante marcadores de superficie celular o similares. Tales poblaciones pueden adquirirse tomando muestras de tejidos particulares, por ejemplo, médula ósea, ganglios linfáticos o similares o clasificando o enriqueciendo células de una muestra (tal como sangre periférica) basándose en uno o más marcadores de superficie celular, tamaño, morfología o similar. En aún otros aspectos, la población de linfocitos que se corresponde con un repertorio puede derivar de tejidos enfermos, tales como un tejido tumoral, un tejido infectado o similar. En una realización, un perfil de clonotipo que comprende cadenas β de TCR humano o fragmentos de las mismas comprende varias secuencias de nucleótidos distintas en el intervalo de $0,1 \times 10^6$ a $1,8 \times 10^6$, o en el intervalo de $0,5 \times 10^6$ a $1,5 \times 10^6$, o en el intervalo de $0,8 \times 10^6$ a $1,2 \times 10^6$. En otra realización, un perfil de clonotipo que comprende cadenas de IgH humanas o fragmentos de las mismas comprende varias secuencias de nucleótidos distintas en el rango de $0,1 \times 10^6$ a $1,8 \times 10^6$, o en el intervalo de $0,5 \times 10^6$ a $1,5 \times 10^6$, o en el intervalo de $0,8 \times 10^6$ a $1,2 \times 10^6$. En una realización particular, un perfil de clonotipo comprende un conjunto de secuencias de nucleótidos que codifican sustancialmente todos los segmentos de la región V(D)J de una cadena IgH. En un aspecto, "sustancialmente todo" tal como se utiliza en el presente documento se refiere a cada segmento que tiene una abundancia relativa del 0,001 por ciento o superior; o en otro aspecto, "sustancialmente todo" tal como se utiliza en el presente documento se refiere a cada segmento que tiene una abundancia del 0,0001 por ciento o superior. En otra realización particular, un perfil de clonotipo comprende un conjunto de secuencias de nucleótidos que codifica sustancialmente todos los segmentos de la región V(D)J de una cadena de TCR β. En otra realización, un perfil de clonotipo comprende un conjunto de secuencias de nucleótidos que tiene longitudes en el intervalo de 25-200 nucleótidos y que incluye segmentos de las regiones V, D y J de una cadena de TCR β. En otra realización, un

perfil de clonotipo comprende un conjunto de secuencias de nucleótidos que tiene longitudes en el intervalo de 25-200 nucleótidos y que incluye segmentos de las regiones V, D y J de una cadena de IgH. En otra realización, un perfil de clonotipo comprende un número de distintas secuencias de nucleótidos que es sustancialmente equivalente al número de linfocitos que expresan una cadena de IgH distinta. En otra realización, un perfil de clonotipo comprende un número de distintas secuencias de nucleótidos que es sustancialmente equivalente al número de linfocitos que expresa una cadena de TCR β distinta. En aún otra realización, "sustancialmente equivalente" se refiere a que con el noventa y nueve por ciento de probabilidad un perfil de clonotipo incluirá una secuencia de nucleótidos que codifique una IgH o TCR β o porción del mismo portado o expresado por cada linfocito de una población de un individuo a una frecuencia del 0,001 por ciento o superior. En aún otra realización, "sustancialmente equivalente" se refiere a que el con el noventa y nueve por ciento de probabilidad un repertorio de secuencias de nucleótidos incluirá una secuencia de nucleótidos que codifique una IgH o TCR β o porción del mismo portado o expresado por cada linfocito presente a una frecuencia del 0,0001 por ciento o superior. En algunas realizaciones, los perfiles de clonotipo derivan de muestras que comprenden de 10^5 a 10^7 linfocitos. Tales números de linfocitos pueden obtenerse a partir de muestras de sangre periférica de 1-10 ml.

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

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

"Fusionar" se refiere tratar dos clonotipos con diferencia de secuencia como el mismo determinando que tales diferencias se deben a un error experimental o de medición y no debido a diferencias biológicas genuinas. En un aspecto, una secuencia de un clonotipo de frecuencia superior se compara con la de un clonotipo de frecuencia inferior y si se cumples los criterios predeterminados entonces el número de clonotipos de frecuencia inferior se añade al del clonotipo de frecuencia superior y el clonotipo candidato de frecuencia inferior se descarta a continuación. Es decir, los recuentos de lectura de secuencia asociados con el clonotipo de frecuencia más baja se agregan a los del clonotipo de frecuencia más alta (y en algunas realizaciones, la secuencia asociada con el clonotipo de frecuencia más baja se elimina de una consideración adicional en la generación de un perfil de clonotipo).

"Trastorno proliferativo linfoide o mieloide" se refiere a cualquier trastorno proliferativo anormal en el que una o más secuencias de nucleótidos que codifican uno o más receptores inmunitarios reordenados pueden usarse como un marcador para controlar tal trastorno. "Neoplasia linfoide o mieloide" se refiere a una proliferación anormal de células de linfocitos o mieloides que pueden ser malignas o no malignas. Un cáncer linfoide es una neoplasia linfoide maligna. Un cáncer mieloide es una neoplasia mieloide maligna. Las neoplasias linfoides y mieloides son el resultado de, o están asociados con, trastornos linfoproliferativos o mieloproliferativos e incluyen, pero sin limitación, linfoma folicular, leucemia linfocítica crónica (LLC), leucemia linfocítica aguda (LLA), leucemia mielógena crónica (LMC), leucemia mielógena aguda (LMA), linfomas de Hodgkin y de no Hodgkin, mieloma múltiple (MM), gammapatía monoclonal de significancia indeterminada (MGUS), linfoma de células del manto (LCM), linfoma difuso de células B grandes (LDCBG), síndromes mielodisplásicos (SMD), linfoma de linfocitos T o similares, por ejemplo, Jaffe y col., Blood, 112: 4384-4399 (2008); Swerdlow y col., Clasificación de la OMS de tumores de tejidos hematopoyéticos y linfoides (4ª ed.) (IARC Press, 2008).

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

Sequence Analysis Package, Genetics Computer Group, Madison, WI). En otras palabras, por ejemplo, para obtener un polinucleótido que tenga una secuencia de nucleótidos de al menos el 95 por ciento de identidad con una secuencia de nucleótidos de referencia, hasta el cinco por ciento de nucleótidos en la secuencia de referencia pueden eliminarse o sustituirse con otro nucleótido, o un número de nucleótidos de hasta el cinco por ciento del número total de nucleótidos en la secuencia de referencia puede insertarse en la secuencia de referente.

"Reacción en cadena de la polimerasa", o "PCR", se refiere a una reacción para la amplificación *in vitro* de secuencias de ADN específicas mediante la extensión de cebador simultánea de cadenas complementarias de ADN. En otras palabras, la PCR es una reacción para realizar múltiples copias o réplicas de un ácido nucleico recombinado flanqueado por sitios de unión de cebador, comprendiendo tal reacción una o más repeticiones de las siguientes etapas: (i) desnaturalizar el ácido nucleico diana, (ii) hibridar los cebadores con los sitios de unión de cebador y (iii) extender los cebadores mediante una polimerasa de ácido nucleico en presencie de trifosfatos de nucleósidos. Normalmente, la reacción se somete a ciclos a través de distintas temperaturas optimizadas para cada etapa en un instrumento de ciclos térmico. Temperaturas particulares, duraciones en cada etapa y tasa de cambio entre etapas depende de muchos factores bien conocidos por los expertos en la técnica, por ejemplo, ejemplificados por las referencias: McPherson y col., editores, PCR: A Practical Approach and PCR2: A Practical Approach (IRL Press, Oxford, 1991 y 1995, respectivamente). Por ejemplo, en una PCR convencionales que usa ADN polimerasa Taq, puede desnaturalizarse un ácido nucleico diana bicatenario a una temperatura de >90 °C, los cebadores hibridan a una temperatura en el intervalo de 50-75 °C, y los cebadores extendieron a una temperatura en el intervalo de 72-78 °C. El término "PCR" abarca formas derivadas de la reacción, incluyendo, aunque no de forma limitativa, RT-PCR, PCR en tiempo real, PCR anidada, PCR cuantitativa, PCR multiplexada y similares. Los volúmenes de reacción varían desde unos cientos de nanolitros, por ejemplo, 200 nl, a unos pocos cientos de µl, por ejemplo, 200 µl. "PCR de transcripción inversa", o "RT-PCR", se refiere a una PCR que está precedida por una reacción de transcripción inversa que convierte una ARN diana en una ADN monocatenario complementario, que, a continuación, se amplifica, por ejemplo, Tecott y col., la patente de Estados Unidos 5.168.038. "PCR en tiempo real" se refiere a una PCR para cual la cantidad de producto de reacción, es decir, amplicón, se controla según procede la reacción. Existen muchas formas de PCR en tiempo real que difieren principalmente en las químicas de detección usadas para controlar el producto de reacción, por ejemplo, Gelfand y col., patente de los EE.UU. 5.210.015 ("taqman"); Wittwer y col., patentes de Estados Unidos 6.174.670 y 6.569.627 (tintes de intercalado); Tyagi y col., patente de Estados Unidos 5.925.517 (balizas moleculares). Las químicas de detección para PCR en tiempo real se revisan en Mackay y col., Nucleic Acids Research, 30: 1292-1305 (2002). "PCR anidada" se refiere a una PCR de dos etapas en donde el amplicón de un primer PCR se convierte en la muestra para la segunda PCR que usa un nuevo conjunto de cebadores, al menos uno del cual se une en un emplazamiento interior del primer amplicón. Como se usa en el presente documento, "cebadores iniciales" en referencia a una reacción de amplificación anidada se refiere a los cebadores usados para generar un primer amplicón y "cebadores secundarios" se refiere al uno o más cebadores usados para generar un segundo, o anidado, amplicón. "PCR multiplexada" se refiere a una PCR en donde múltiples secuencias diana (o una única secuencia diana y una o más secuencias de referencia) se llevan a cabo simultáneamente en la misma mezcla de reacción, por ejemplo, Bernard y col., Anal. Biochem., 273: 221-228 (1999)(PCR en tiempo real de dos colores). Normalmente, se emplean distintos conjuntos de cebadores para cada secuencia que se está amplificando. Típicamente, El número de secuencias diana en una PCR multiplex está en el intervalo de 2 a 50, o de 2 a 40, o de 2 a 30. "PCR cuantitativa" se refiere a una PCR designada para medir la abundancia de una o más secuencias diana específicas en una muestra o espécimen. La PCR cuantitativa incluye tanto la cuantificación absoluta como la cuantificación relativa de tales secuencias diana. Las mediciones cuantitativas se realizan usando unas o más secuencias de referencia o estándares internos que pueden someterse a ensayo por separado o juntos con una secuencia diana. La secuencia de referencia puede ser endógena o exógena con respecto a una muestra o espécimen y, en el último caso, puede comprender uno o más modelos de competidor. Secuencias de referencia endógenas típicas incluyen segmentos de transcripciones de los siguientes genes: β-actina, GAPDH, β_2-microglobulina, ARN ribosómico y similares. Las técnicas para la PCR cuantitativa son bien conocidos de aquellas personas normalmente expertas en la materia, como se ilustran en las siguientes referencias: Freeman y col., Biotechniques, 26: 112-126 (1999); Becker-Andre y col., Nucleic Acids Research, 17: 9437-9447 (1989); Zimmerman y col., Biotechniques, 21: 268-279 (1996); Diviacco y col., Gene, 122: 3013-3020 (1992); Becker-Andre y col., Nucleic Acids Research, 17: 9437-9446 (1989); y similares. "Cebador" se refiere a un oligonucleótido, o bien natural o bien sintético que es capaz, cuando se forma un híbrido con un modelo de polinucleótido, de actuar como punto de inicio de la síntesis de ácidos nucleico y de extenderse desde su extremo 3' a lo largo del modelo de modo que se forma un híbrido extendido. La extensión de un cebador se lleva a cabo normalmente con una polimerasa de ácido nucleico, tal como ADN o ARN polimerasa. La secuencia de nucleótidos añadida en el proceso de extensión se determina mediante la secuencia del polinucleótido modelo. Normalmente, los cebadores se extienden mediante ADN polimerasa. Los cebadores tienen normalmente una longitud en el intervalo de 14 a 40 nucleótidos o en el intervalo de 18 a 36 nucleótidos. Los cebadores se emplean en una variedad de reacciones de amplificación nucleica, por ejemplo, las reacciones de amplificación lineal que usan un único cebador o reacciones de cadena de la polimerasa, que emplean dos o más cebadores. Directrices para seleccionar las longitudes y secuencias de cebadores para aplicaciones particulares son bien conocidos por los expertos en la técnica, tal como se muestra en las siguientes referencias: Dieffenbach, editor, PCR Primer: A Laboratory Manual, 2ª Edición (Cold Spring Harbor Press, Nueva York, 2003). "Puntuación de calidad" se refiere a una medición de la probabilidad de que

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

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

REIVINDICACIONES

1. Un procedimiento para seleccionar uno o más clonotipos específicos del paciente correlacionados con una neoplasia linfoide para controlar una enfermedad residual mínima de la misma, comprendiendo el procedimiento las etapas de:

(a) amplificar moléculas de ácido nucleico de linfocitos T y/o linfocitos B de una muestra obtenida de un paciente, comprendiendo las moléculas de ácido nucleico secuencias de ADN recombinadas de genes del receptor de linfocitos T o genes de inmunoglobulina;
(b) secuenciar las moléculas amplificadas de ácido nucleico para formar un perfil de clonotipo;
(c) comparar clonotipos del perfil de clonotipo con los clonotipos de una base de datos de clonotipos, en el que la base de datos de clonotipos comprende clonotipos de perfiles de clonotipos de al menos 10^4 clonotipos de al menos 10 individuos distintos del paciente, para determinar una presencia, ausencia y/o nivel en la base de datos de clonotipos de cada clonotipo del perfil de clonotipo, en el que la etapa de comparación incluye contar un clonotipo como presente en la base de datos de clonotipos si un clonotipo en la base de datos de clonotipos es miembro del mismo clan que el clonotipo del perfil de clonotipo, y un clonotipo se considera miembro del mismo clan si

(i) el clonotipo de la base de datos de clonotipos es al menos noventa por ciento idéntico al clonotipo del perfil de clonotipo;
(ii) el clonotipo de la base de datos de clonotipos y el clonotipo del perfil de clonotipo comprenden secuencias recombinadas de la cadena pesada de inmunoglobulina y están relacionadas por un reemplazo de VH;
(iii) el clonotipo de la base de datos de clonotipos y el clonotipo del perfil de clonotipo tienen una región V y una región J mutadas idénticamente pero tienen una región NDN diferente;
(iv) el clonotipo de la base de datos de clonotipos y el clonotipo del perfil de clonotipo comprenden cada uno secuencias recombinadas de la cadena pesada de inmunoglobulina y están relacionados por hipermutación; o
(v) si los clonotipos han sufrido una o más inserciones y/o deleciones de 1-10 bases; y

(d) seleccionar uno o más clonotipos específicos del paciente para controlar la enfermedad residual mínima, cada uno de los cuales se correlaciona con la neoplasia linfoide y cada uno de ellos está:

i) ausente de la base de datos de clonotipos; o
ii) a un nivel en la base de datos de clonotipos por debajo de una frecuencia predeterminada.

2. El procedimiento de la reivindicación 1, en el que el uno o más clonotipos específicos del paciente son cada uno de 25 a 400 secuencias de nucleótidos que codifican un segmento de un receptor inmune o componente del receptor inmune seleccionado del grupo que consiste en un reordenamiento de VDJ de IgH, un reordenamiento de DJ de IgH, un reordenamiento de VJ de IgK, un reordenamiento de VJ de IgL, un reordenamiento de VDJ de TCR β, un reordenamiento de DJ de TCR β, un reordenamiento de VJ de TCR α, un reordenamiento de VJ de TCR γ, un reordenamiento de VDJ de TCR δ, o un reordenamiento de VD de TCR δ.

3. El procedimiento de la reivindicación 1, en el que la etapa de comparación incluye alinear secuencias de los clonotipos del perfil de clonotipo con secuencias de los clonotipos de la base de datos de clonotipos.

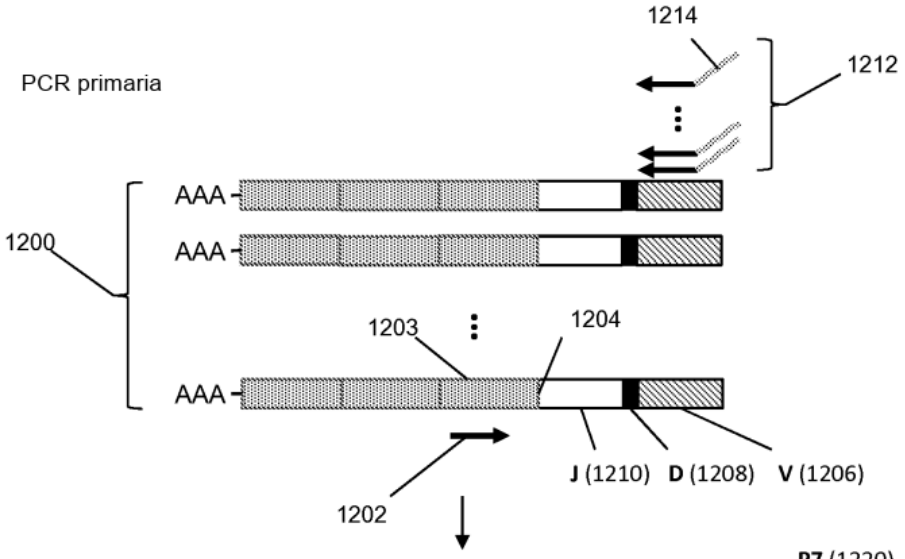


Fig. 1A

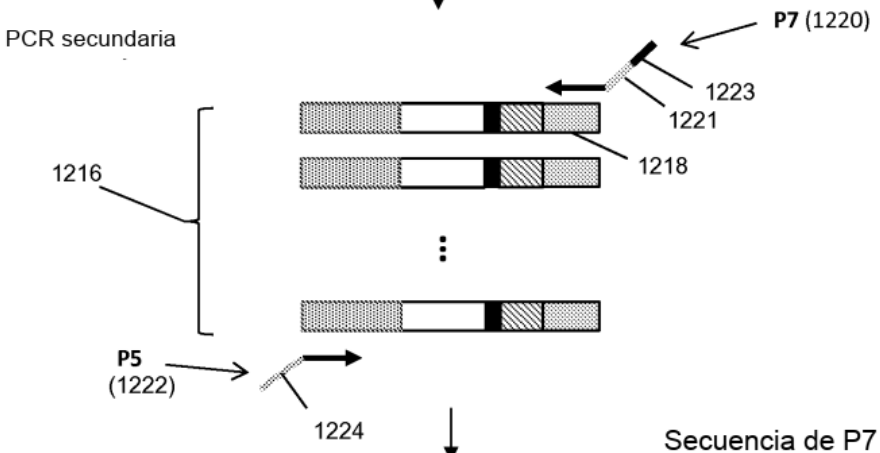


Fig. 1B

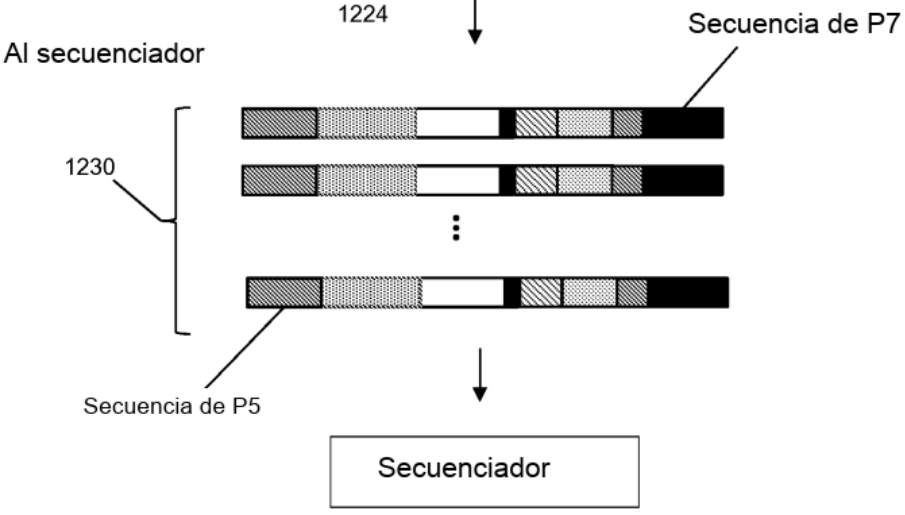


Fig. 1C

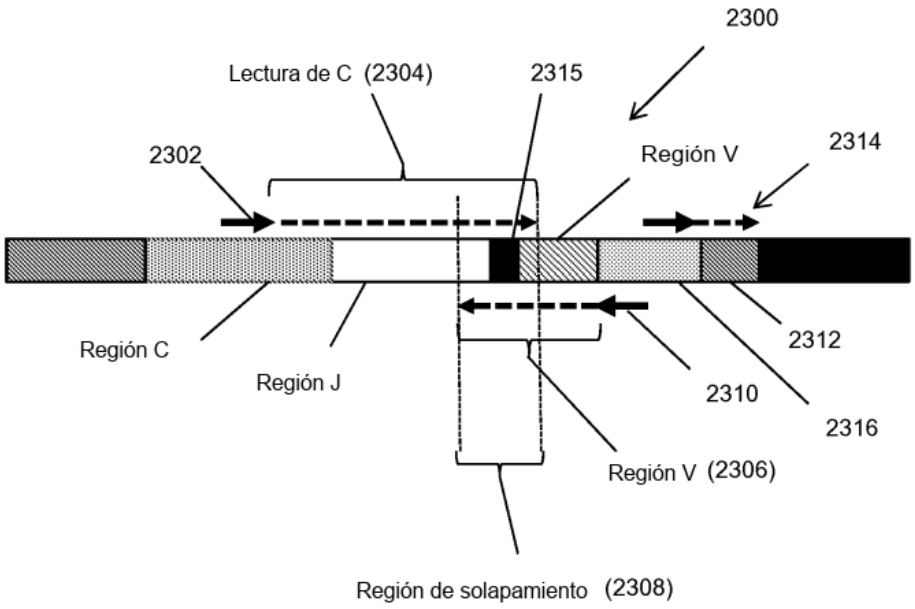

2300
Lectura de C (2304)
2315
2302
Región V
2314
Región C
Región J
2312
2310
2316
Región V (2306)
Región de solapamiento (2308)

Fig. 2A

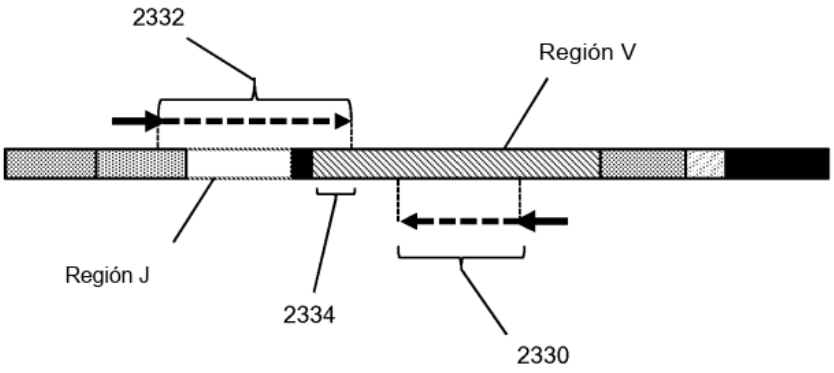

2332
Región V
Región J
2334
2330

Fig. 2B

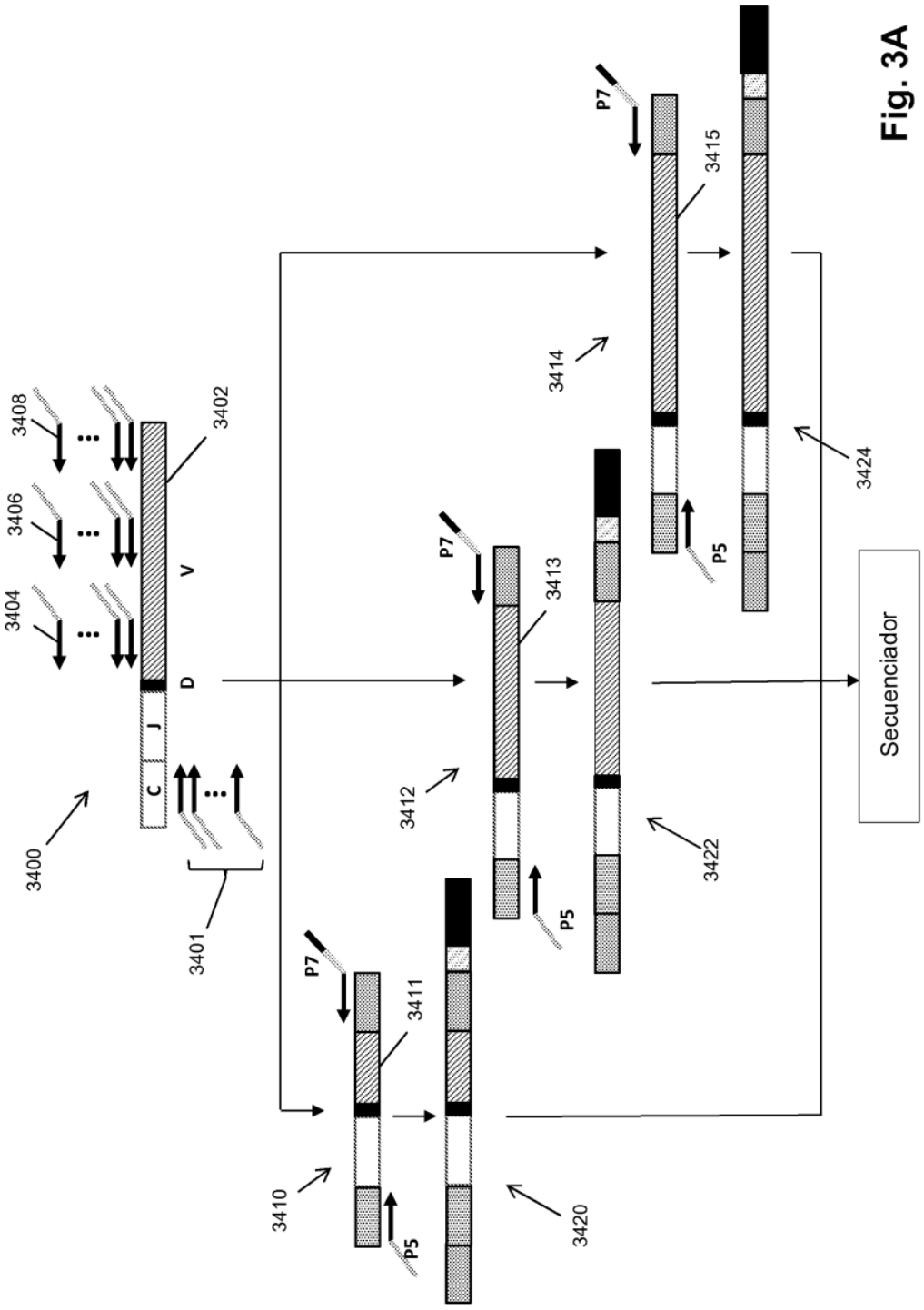

3408
3406
3404
3402
V
D
J
C
3400
3401
P7
3411
P5
3410
3420
3412
3413
3414
3415
3422
3424
Secuenciador

Fig. 3A

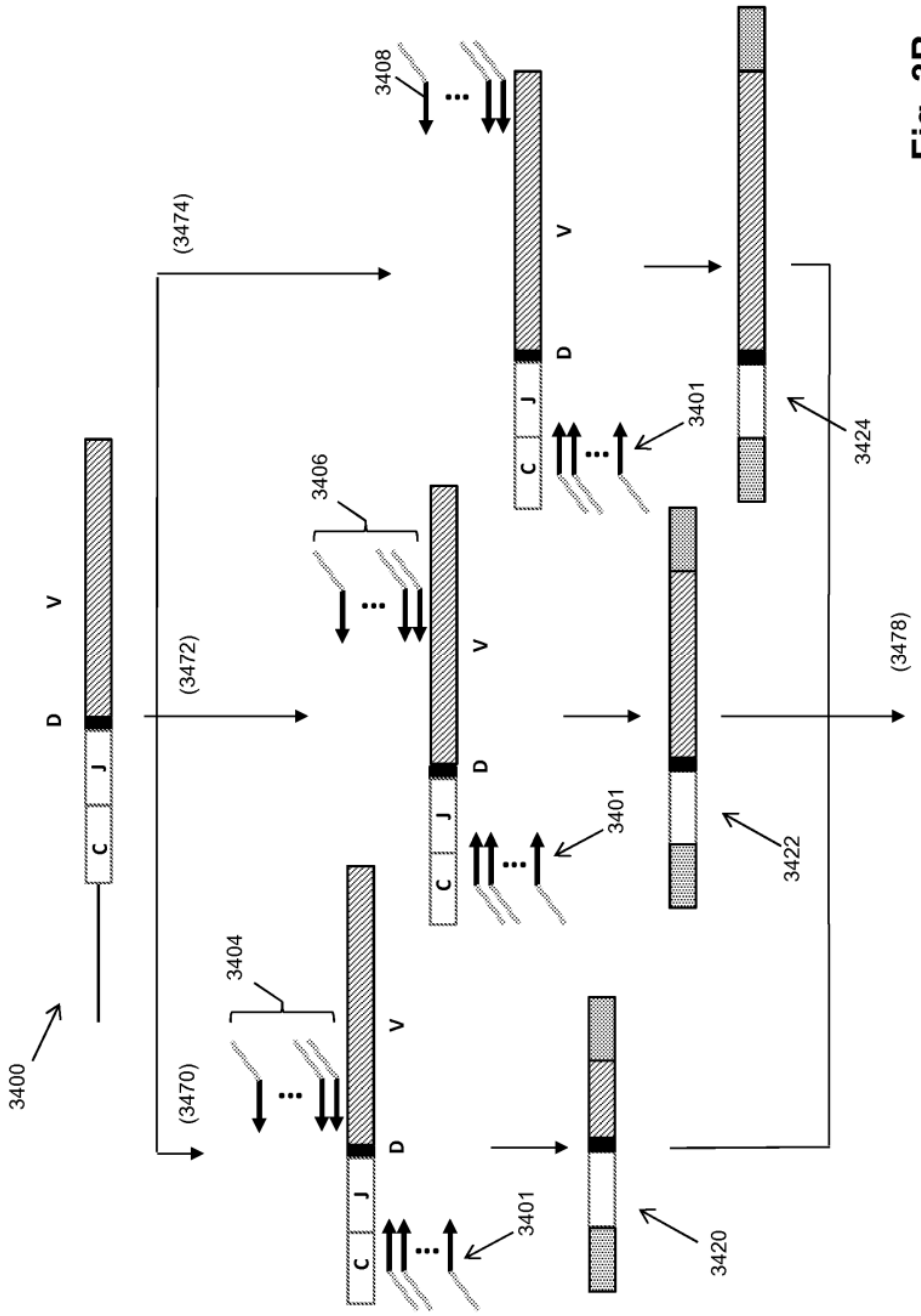

3400
V
D
J
C
(3470)
3404
3401
3420
(3472)
3406
3422
(3474)
3408
3424
(3478)

Fig. 3B

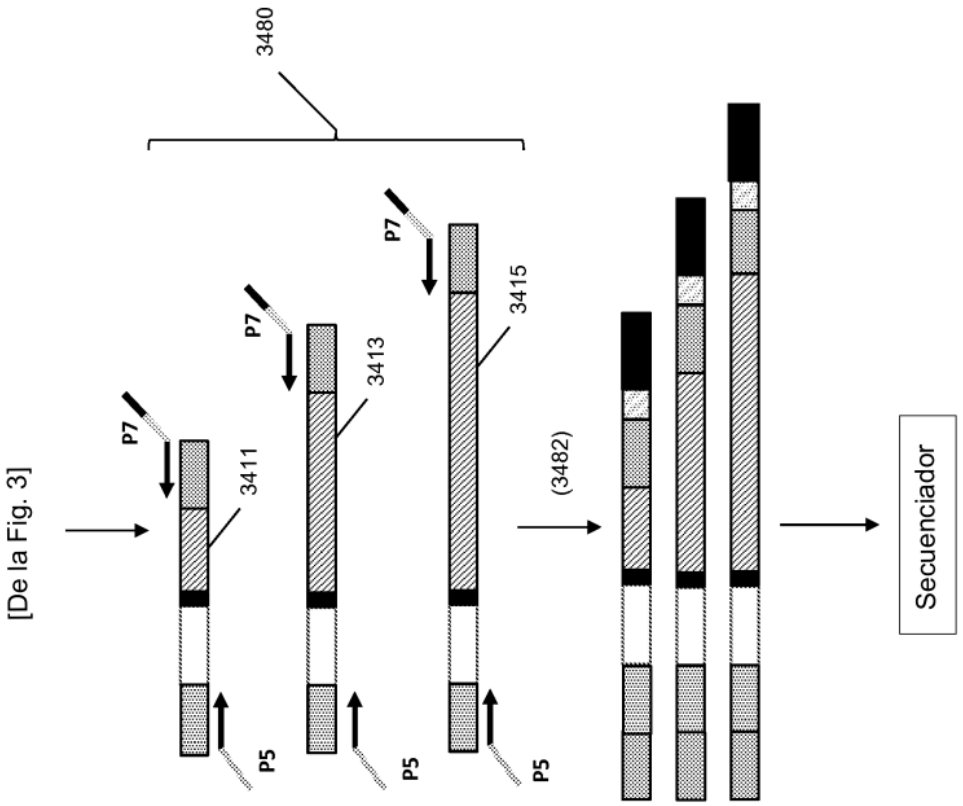
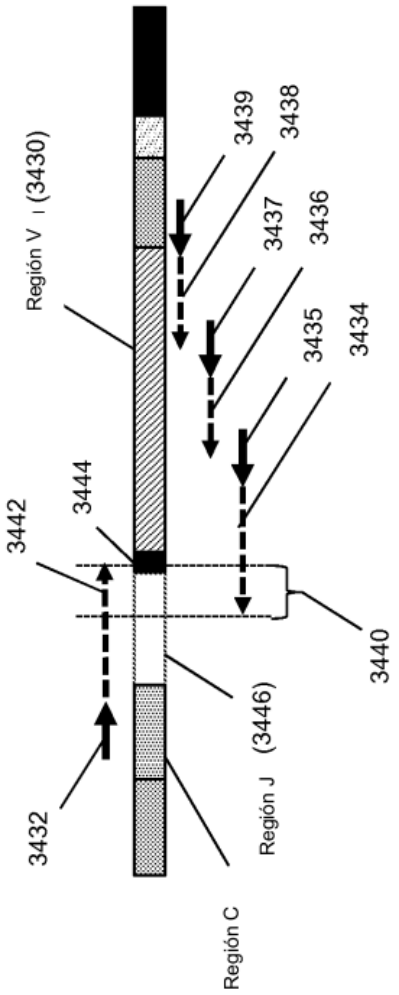

[De la Fig. 3]
P7
3411
P5
P7
3413
P5
P7
3415
P5
3480
(3482)
Secuenciador

Fig. 3C

Fig. 3D


Región V I (3430)
3439
3438
3437
3436
3435
3434
3444
3442
3440
3432
Región J (3446)
Región C

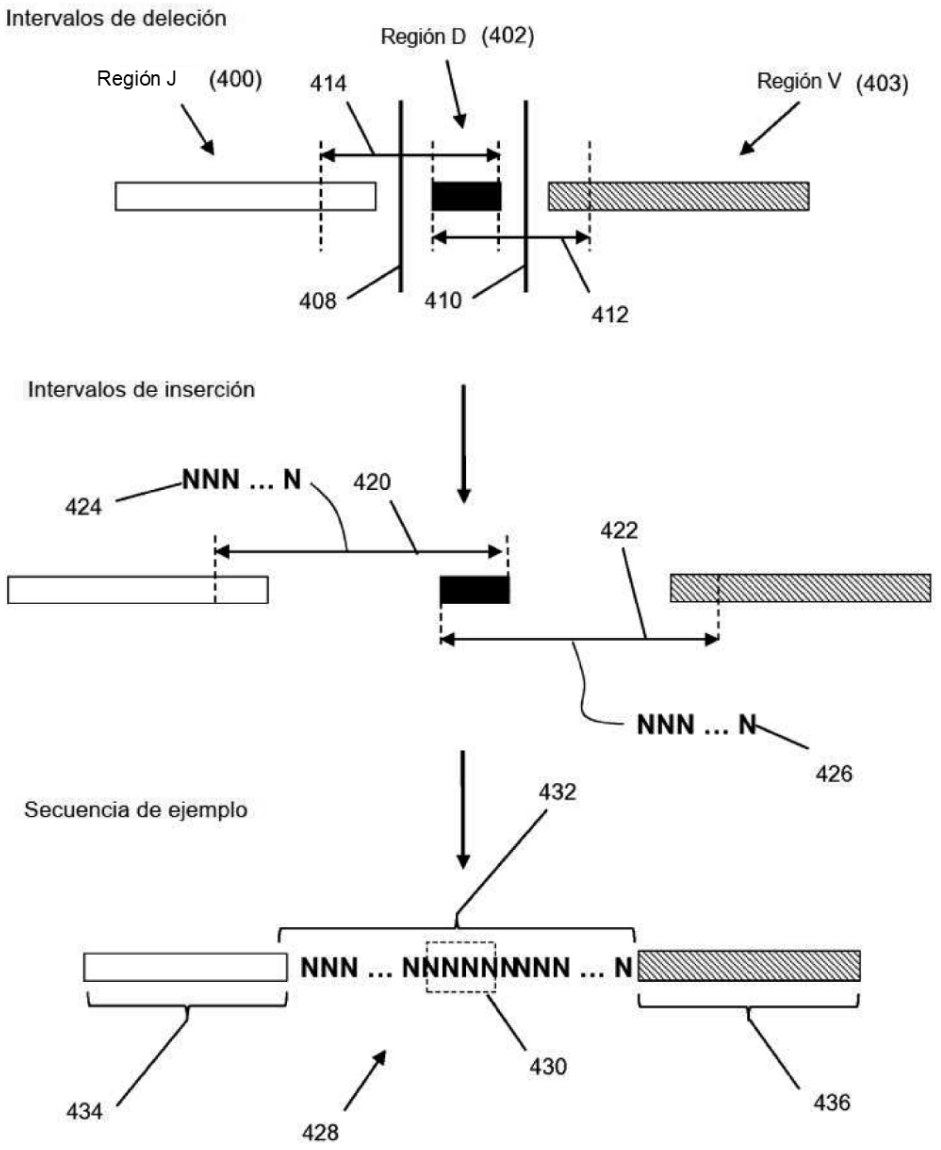

Intervalos de deleción
Región D (402)
Región J (400)
414
Región V (403)
408
410
412
Intervalos de inserción
NNN ... N
424
420
422
NNN ... N
426
Secuencia de ejemplo
432
NNN ... NNNNNNNNN ... N
430
434
428
436

Fig. 4A

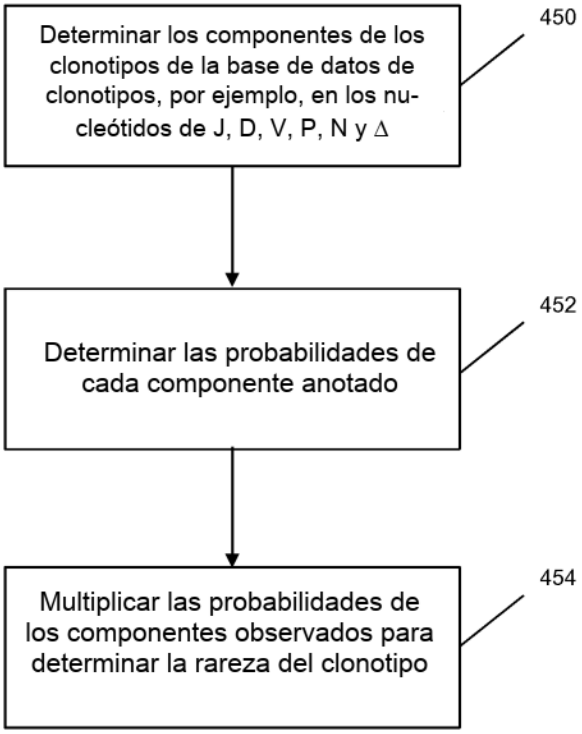

Determinar los componentes de los clonotipos de la base de datos de clonotipos, por ejemplo, en los nucleótidos de J, D, V, P, N y Δ
450
Determinar las probabilidades de cada componente anotado
452
Multiplicar las probabilidades de los componentes observados para determinar la rareza del clonotipo
454

Fig. 4B