(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 784 343**

(51) Int. CI.:

C12N 15/10 (2006.01)
C12Q 1/68 (2008.01)
C12Q 1/6809 (2008.01)
C12Q 1/6881 (2008.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(86) Fecha de presentación y número de la solicitud internacional: **29.10.2015 PCT/US2015/058035**

(87) Fecha y número de publicación internacional: **06.05.2016 WO16069886**

(96) Fecha de presentación y número de la solicitud europea: **29.10.2015 E 15854358 (7)**

(97) Fecha y número de publicación de la concesión europea: **25.03.2020 EP 3212790**

(54) Título: **Detección simultánea altamente multiplexada de ácidos nucleicos que codifican heterodímeros de receptores inmunes adaptativos emparejados de muchas muestras**

(30) Prioridad:

29.10.2014 US 201462072162 P

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
24.09.2020

(73) Titular/es:

ADAPTIVE BIOTECHNOLOGIES CORP. (50.0%)
1551 Eastlake Avenue East, Suite 200
Seattle, Washington 98102, US y
FRED HUTCHINSON CANCER RESEARCH CENTER (50.0%)

(72) Inventor/es:

EMERSON, RYAN, O.;
SHERWOOD, ANNA, M. y
ROBINS, HARLAN, S.

(74) Agente/Representante:

UNGRÍA LÓPEZ, Javier

DESCRIPCIÓN

Detección simultánea altamente multiplexada de ácidos nucleicos que codifican heterodímeros de receptores inmunes adaptativos emparejados de muchas muestras

Antecedentes de la invención

Campo técnico

La presente divulgación se refiere en general a métodos y composiciones útiles para la amplificación de ácido nucleico multiplexado y la secuenciación de alto rendimiento de moléculas de ácido nucleico del receptor inmunitario adaptativo y al emparejamiento de moléculas de ácido nucleico que codifican cadenas de polipéptidos afines de heterodímeros del receptor inmunitario adaptativo de un gran número de muestras biológicas.

Descripción de la técnica relacionada

El sistema inmunitario adaptativo protege a los organismos superiores contra infecciones y otros eventos patológicos que pueden atribuirse a sustancias extrañas. Con el uso de receptores inmunes adaptativos, las proteínas de reconocimiento específicas de antígeno son expresadas por las células hematopoyéticas del linaje linfoide y son capaces de distinguir las moléculas propias de las no propias en el hospedador. Estos linfocitos se pueden encontrar en la circulación y los tejidos de un hospedador, y se ha descrito su recirculación entre la sangre y los linfáticos, incluyendo su extravasación a través de las vénulas endoteliales altas de los ganglios linfáticos, así como en sitios de infección, inflamación, lesión tisular y otras lesiones clínicas. Véase, por ejemplo, Stein et al., 2005 Immunol. 116:1-12; DeNucci et al., 2009 Crit. Rev. Immunol. 29:87-109; Marelli-Berg et al., 2010 Immunol. 130:158; Ward et al., 2009 Biochem. J. 418:13; Gonzalez et al., 2011 Ann. Rev. Immunol. 29:215; Kehrl et al., 2009 Curr. Top. Microb. Immunol. 334:107; Steinmetz et al., 2009 Front. Biosci. (Schol. Ed.) 1:13.

Por consiguiente, la naturaleza dinámica del movimiento de los linfocitos en todo el organismo hospedador se refleja en los cambios cualitativos (por ejemplo, la especificidad antigénica del receptor inmunitario adaptativo expresado clonalmente (inmunoglobulina o receptor de linfocitos T), linfocitos T frente a linfocitos B, linfocitos T colaboradores (T_h) frente a linfocitos T reguladores (T_{reg}), linfocitos T efectores frente a linfocitos T de memoria, etc.) y distribución cuantitativa de linfocitos entre tejidos, en función de los cambios en el estado inmunitario del hospedador.

El sistema inmune adaptativo emplea varias estrategias para generar un repertorio de receptores de antígeno de linfocitos T y B con suficiente diversidad para reconocer el universo de posibles patógenos.

Las inmunoglobulinas (Ig) expresadas por los linfocitos B son proteínas que consisten en cuatro cadenas de polipéptidos, dos cadenas pesadas (cadenas H) y dos cadenas ligeras (cadenas L), que forman una estructura H_2L_2. Cada par de cadenas H y L contiene un dominio hipervariable, que consiste en una región V_L y una V_H, y un dominio constante. Las cadenas H de las Ig son de varios tipos: μ, δ, γ, α y ε. La diversidad de las Ig dentro de un individuo está determinada principalmente por el dominio hipervariable. El dominio V de las cadenas H se crea mediante la unión combinatoria de tres tipos de segmentos de genes de la línea germinal, los segmentos V_H, D_H y J_H. La diversidad de la secuencia del dominio hipervariable aumenta aún más mediante la adición y eliminación independiente de nucleótidos en las uniones V_H-D_H, D_H-J_H, y V_H-J_H durante el proceso de reordenamiento génico de la Ig. A este respecto, la inmunocompetencia se refleja en la diversidad de las Ig.

Los receptores de linfocitos T (TCR, por sus siglas en inglés) se expresan mediante linfocitos T αβ o linfocitos T γδ. Los TCR expresados por los linfocitos T αβ son proteínas que consisten en dos cadenas de polipéptidos transmembrana (α y β), expresadas a partir de los genes TCRA y TCRB, respectivamente. Las proteínas TCR similares se expresan en los linfocitos T γδ, de los loci TCRD y TCRG. Cada péptido de TCR contiene regiones determinantes de complementariedad (CDR) variables, así como regiones marco (FR) y una región constante. La diversidad de secuencia de los linfocitos T αβ está determinada en gran medida por la secuencia de aminoácidos de los bucles de la tercera región determinante de complementariedad (CDR3) de los dominios variables de la cadena α y β, cuya diversidad es el resultado de la recombinación entre los segmentos génicos variable (Vβ), de diversidad (Dβ) y de unión (Jβ) en el locus de la cadena β, y entre análogos de los segmentos génicos $V_α$ y $J_α$ en el locus de la cadena α, respectivamente. La recombinación de los segmentos génicos variable, de diversidad y de unión en los loci de cadena α y β del TCR permite codificar un gran número de secuencias de la CDR3 distintas. La diversidad de secuencias de la CDR3 aumenta aún más mediante la adición y eliminación independiente de nucleótidos en las uniones $V_β$-$D_β$, $D_β$-$J_β$, y $V_α$-$J_α$ durante el proceso de reordenamiento génico del TCR. A este respecto, la inmunocompetencia se refleja en la diversidad de los TCR.

El γδ TCR se distingue del αβ TCR en que codifica un receptor que interactúa estrechamente con el sistema inmunitario innato. E TCRγδ se expresa temprano en el desarrollo, tiene distribución anatómica especializada, tiene especificidades únicas de patógenos y moléculas pequeñas, y tiene un amplio espectro de interacciones celulares innatas y adaptativas. Un patrón sesgado de la expresión del segmento V y J del TCRγ se establece temprano en la ontogenia a medida que los subconjuntos restringidos de células TCRγδ pueblan la boca, la piel, el intestino, la

vagina y los pulmones prenatalmente. En consecuencia, el diverso repertorio de TCRγ en tejidos adultos es el resultado de una extensa expansión periférica después de la estimulación por exposición ambiental a patógenos y moléculas tóxicas.

Los reordenamientos de VDJ están mediados por un complejo de enzima recombinasa en el que las proteínas RAG1 y RAG2 juegan un papel clave al reconocer y cortar el ADN en las secuencias de señal de recombinación (RSS), que se encuentran aguas abajo de los segmentos génicos V, a ambos lados de los segmentos génicos D, y aguas arriba de los segmentos génicos J. Las RSS divergentes reducen o incluso evitan completamente los reordenamientos. La secuencia de señal de recombinación (RSS) consta de dos secuencias conservadas (el heptámero, 5'-CACAGTG-3', y el nonámero, 5'-ACAAAAACC-3'), separados por un espaciador de 12 +/- 1 pb ("señal 12") o de 23 +/- 1 pb ("señal 23"). Se han identificado varias posiciones de nucleótidos como importantes para la recombinación, incluido el dinucleótido CA en la posición uno y dos del heptámero, y también se ha demostrado que una C en la posición tres del heptámero es muy preferida, así como un nucleótido A en las posiciones 5, 6, 7 del nonámero. (Ramsden et al. 1994 Nucl. Ac. Res. 22:1785; Akamatsu et al. 1994 J. Immunol. 153:4520; Hesse et al. 1989 Genes Dev. 3:1053). Las mutaciones de otros nucleótidos tienen efectos mínimos o inconsistentes. El espaciador, aunque más variable, también tiene un impacto en la recombinación, y se ha demostrado que las sustituciones de un solo nucleótido tienen un impacto significativo en la eficacia de la recombinación (Fanning et al. 1996 Cell. Immunol. Immunopath. 79:1, Larijani et al. 1999 Nucl. Ac. Res. 27:2304; Nadel et al. 1998 J. Immunol. 161:6068; Nadel et al. 1998 J. Exp. Med. 187:1495). Se han descrito criterios para identificar secuencias RSS de polinucleótidos que tienen eficacias de recombinación significativamente diferentes (Ramsden et al. 1994 Nucl. Ac. Res. 22:1785; Akamatsu et al. 1994 J. Immunol. 153:4520; Hesse et al. 1989 Genes Dev. 3:1053, y Lee et al., 2003 PLoS 1(1):E1).

El proceso de reordenamiento somático generalmente comienza con un reordenamiento de D a J seguido de un reordenamiento de V a D-J en el caso de genes de cadena pesada de Ig (IgH), TCR beta (TCRB) y TCR delta (TCRD) o implica reordenamientos directos de V a J en el caso de los genes Ig kappa (IgK), Ig lambda (IgL), TCR alfa (TCRA) y TCR gamma (TCRG). Las secuencias entre los segmentos de genes reorganizados generalmente se eliminan en forma de un producto de escisión circular, también llamado círculo de escisión de TCR (TREC) o círculo de escisión del receptor de linfocitos B (BREC).

Las muchas combinaciones diferentes de los segmentos génicos V, D y J representan el llamado repertorio combinatorio, que se estima en al menos de 10^6 moléculas; por ejemplo, ~$2x10^6$ para las moléculas de Ig, ~$3x10^6$ para las moléculas de TCRαβ y ~$5x10^3$ para las moléculas de TCRγδ. En los sitios de unión de los segmentos génicos V, D y J, la deleción e inserción aleatoria de nucleótidos ocurre durante el proceso de reordenamiento, que da como resultado regiones de unión altamente diversas, que contribuyen significativamente al repertorio total de moléculas de Ig y TCR, estimado en > 10^{12}.

Los linfocitos B maduros extienden aún más su repertorio de Ig tras el reconocimiento de antígeno en los centros foliculares mediante hipermutación somática, un proceso, que conduce a la maduración por afinidad de las moléculas de Ig. El proceso de hipermutación somática se centra en el exón V-(D-) J de los genes de cadena ligera IgH e Ig y se refiere a mutaciones de un solo nucleótido y, a veces, también a inserciones o deleciones de nucleótidos. Los genes IG mutados somáticamente también se encuentran en tumores malignos maduros de linfocitos B de origen folicular o post-folicular.

Se han empleado varias estrategias diferentes para secuenciar ácidos nucleicos que codifican receptores inmunes adaptativos cuantitativamente a alto rendimiento, y estas estrategias se pueden distinguir, por ejemplo, por la estrategia que se usa para amplificar las regiones que codifican la CDR3, y por la elección de secuenciar el ADN genómico (ADNg) o el ARN mensajero (ARNm). Ciertos métodos convencionales de alto rendimiento secuencian solo una cadena de un receptor inmune adaptativo a la vez, haciendo imposible determinar que el ADN o el ARN que codifica ambas cadenas de un heterodímero TCR o IG se originó a partir de la misma célula linfoide. Para reconstituir los receptores inmunes adaptativos para el análisis funcional, uso terapéutico, o modelado o unión antígeno-receptor, los polipéptidos de cadena emparejados de un heterodímero receptor completo deben identificarse como un par.

Se han descrito varias estrategias para emparejar cadenas de receptores inmunes adaptativos en la técnica. Un enfoque es aislar linfocitos B o T individuales y unir físicamente las cadenas pesadas y ligeras mediante PCR de puente antes de la secuenciación. Véase, por ejemplo, Embleton et al., 1992 Nucleic acids research 20:3831-3837; Meijer et al., 2006 J. Mol. Bio. 358:764-772. Como alternativa, las cadenas pesadas y ligeras pueden tener un código de barras en el nivel unicelular. Véase, e.g. DeKosky et al., 2013 Nat. Biotech. 31:166-169; Dash et al., 2011 J. Clin. Invest. 121:288. (25, 27-31). Aunque los métodos unicelulares han mejorado significativamente, todavía son técnicamente desafiantes y tienen un rendimiento limitado. También requieren células individuales intactas, lo que dificulta evaluar las células inmunitarias infiltrantes en tejidos o tumores sólidos.

Los solicitantes han desarrollado recientemente nuevos métodos para emparejar secuencias de heterodímeros de receptores inmunes adaptativos basados en el uso de combinatoria para detectar y caracterizar incluso secuencias de receptores emparejados muy raros en entornos biológicos complejos (véase, por ejemplo, los documentos

WO2013/188831 y WO2014/145992). Si bien estas estrategias son altamente eficaces, ya que son capaces de generar decenas de miles de secuencias de receptores inmunes emparejados, pueden presentar ciertos desafíos técnicos. Por ejemplo, el experimento mínimo necesario para identificar secuencias de receptores inmunes adaptativos emparejados en una pluralidad de muestras biológicas requiere decenas de millones de lecturas de secuencia por muestra individual analizada. Los métodos descritos anteriormente requieren que cada muestra de interés se analice en un experimento de emparejamiento separado. Dichos experimentos grandes pueden imponer considerables cargas de recursos en situaciones en las que se debe analizar simultáneamente un gran número de muestras biológicas diferentes. Además, en algunos casos, los investigadores solo pueden estar interesados en las secuencias del receptor inmunitario adaptativo emparejado de mayor frecuencia en una pluralidad de muestras, y no necesitan información sobre todas las secuencias emparejadas en la pluralidad de muestras.

Claramente, sigue existiendo la necesidad de composiciones y métodos mejorados para un emparejamiento preciso, aunque eficiente, de alto rendimiento de secuencias de receptores inmunes adaptativos en múltiples muestras biológicas. Las realizaciones actualmente descritas abordan esta necesidad y proporcionan otras ventajas relacionadas.

Sumario de la invención

En el presente documento se desvelan métodos para emparejar simultáneamente ácidos nucleicos reorganizados que codifican secuencias de polipéptidos heterodímeros del receptor inmunitario adaptativo a alto rendimiento de un gran número de muestras de fuentes biológicas que contienen células inmunes adaptativas.

La invención proporciona un método para asignar un par de polipéptidos primero y segundo que forman un receptor de linfocitos T (TCR) o un heterodímero de inmunoglobulina (Ig) a una muestra de fuente única entre una pluralidad de muestras de la fuente, incluyendo (1) para cada una de la pluralidad de muestras de la fuente cada una de las cuales comprende linfocitos T o linfocitos B, determinando las primeras secuencias de ácido nucleico reorganizadas que codifican los primeros polipéptidos de los heterodímeros de TCR o Ig presentes en la muestra de la fuente y asignar las primeras secuencias de ácido nucleico reorganizadas a la muestra de la fuente; (2) agrupar la pluralidad de muestras de la fuente para formar una población combinada de células; (3) determinar a partir de la población combinada de células, una pluralidad de pares afines de secuencias de ácido nucleico reorganizadas primera y segunda que codifican los polipéptidos primero y segundo de los heterodímeros de TCR o Ig; (4) comparar las primeras secuencias de ácido nucleico reorganizadas determinadas en cada una de las muestras de la fuente en (1) con las primeras secuencias de ácido nucleico reorganizadas determinadas a partir de la pluralidad de pares afines de secuencias de ácido nucleico reorganizadas en (3) para asignar cada primera secuencia de ácido nucleico reordenada presente en la población combinada con una sola muestra de la fuente; y (5) para cada primera secuencia de ácido nucleico reordenada asignada a una única muestra de la fuente en la etapa (4), asignando la segunda secuencia de ácido nucleico reordenada afín del par afín identificado en la etapa (3) a la misma muestra única de la fuente.

En determinadas realizaciones de la invención, la primera secuencia de ácido nucleico reordenada es una secuencia de ácido nucleico reordenada de TCRB, una secuencia de ácido nucleico reordenada de TCRA, una secuencia de ácido nucleico reordenada de cadena pesada de inmunoglobulina (IGH), o una secuencia de ácido nucleico reordenada de cadena ligera de inmunoglobulina (IGK de IGL).

En otra realización de la invención, la etapa de determinar una primera secuencia de ácido nucleico reordenada que codifica el primer polipéptido del heterodímero TCR o Ig presente en la muestra fuente incluye las etapas de: para cada muestra de la fuente, amplificar moléculas de ácido nucleico reorganizadas extraídas de la muestra de la fuente en una única reacción en cadena de la polimerasa (PCR) multiplexada usando una pluralidad de cebadores del segmento V y una pluralidad de cebadores del segmento J para producir una pluralidad de amplicones de ácido nucleico reorganizados, y secuenciar dicha pluralidad de amplicones de ácido nucleico reorganizados para determinar las secuencias de las primeras secuencias de ácido nucleico reorganizadas en cada muestra de la fuente. En otra realización más de la invención, la PCR multiplexada simple produce al menos 10^4 amplicones distintos que representan una diversidad de secuencias de la CDR3 del TCR o IG reorganizadas presentes en cada una de las muestras.

En otra realización más de la invención, la pluralidad de cebadores del segmento V y la pluralidad de cebadores del segmento J consta de 15 a 50 nucleótidos. En otra realización más de la invención, los cebadores del segmento V incluyen una primera secuencia y una segunda secuencia, en donde la primera secuencia es complementaria a una porción de una primera región de un segmento V que codifica TCR o IG, la primera región ubicada inmediatamente en 5' a una segunda región del segmento V codificante donde se producen deleciones sin molde durante la reordenación del gen del TCR o la IG, en donde la segunda región del segmento V codificante es adyacente al extremo 5' a una secuencia señal de recombinación de V (V-RSS) del segmento V codificante, en donde la primera secuencia está ubicada en 3' a la segunda secuencia en el cebador del segmento V, en donde la segunda secuencia comprende una secuencia de cebador universal, y en donde cada uno de los cebadores del segmento J tiene una primera secuencia y una segunda secuencia, en donde la primera secuencia es complementaria a una porción de una primera región de un segmento J que codifica el TCR o la IG, la primera región ubicada inmediatamente en 3' a

una segunda región del segmento J codificante donde se producen deleciones sin molde durante la reordenación del gen del TCR o la IG, en donde la segunda región del segmento J es adyacente y en 3' a una secuencia señal de recombinación J (J-RSS) del segmento J codificante, en donde la primera secuencia está ubicada en 3' con la segunda secuencia en el cebador del segmento J, en donde la segunda secuencia comprende una secuencia de cebador universal.

En otra realización, los métodos incluyen realizar una segunda reacción de amplificación hibridando cebadores de cola a regiones dentro de los amplicones de ácido nucleico reorganizados. En otra realización más, el cebador de colas incluye una secuencia de cebador universal, una secuencia de código de barras única, una secuencia de oligonucleótidos aleatoria y una secuencia adaptadora. En otra realización más, la secuencia de código de barras única se utiliza para identificar una muestra particular de la fuente.

En otro aspecto de la invención, la etapa de determinar a partir de la población combinada de células, una pluralidad de pares afines de secuencias de ácido nucleico reorganizadas primera y segunda que codifican los polipéptidos primero y segundo de los heterodímeros de TCR o de Ig incluye las etapas de: distribuir células de la población combinada de células en una pluralidad de envases, comprendiendo cada envase una subpoblación de células, generar una biblioteca de amplicones para cada una de las pluralidades de envases mediante la realización de una única PCR multiplexada de moléculas de ADNc que se han transcrito inversamente a partir de moléculas de ARNm obtenidas de la subpoblación de células, realizar una secuenciación de alto rendimiento de la biblioteca de amplicones para obtener un conjunto de datos de una pluralidad de secuencias primera y segunda de amplicón del receptor inmunitario adaptativo para cada una de las pluralidades de los envases, determinar un patrón de ocupación del envase para cada secuencia única de amplicón del receptor inmune del primer adaptador mediante la asignación de cada secuencia única de amplicón del receptor inmune del primer adaptador a uno o más envases, y determinar un patrón de ocupación del envase para cada segunda secuencia de amplicón del receptor inmune del adaptador único asignando cada segundo adaptador de la secuencia del amplicón del receptor inmune a uno o más envases, para cada posible emparejamiento de una primera y segunda secuencia única de amplicón del receptor inmunitario adaptativo para formar un supuesto par afín, calcular una probabilidad estadística de observar los patrones de ocupación del envase e identificar una pluralidad de supuestos pares afines basados en la probabilidad estadística.

En otra realización de la invención, la etapa de identificar una pluralidad de supuestos pares afines se basa en que dicha probabilidad estadística tiene una puntuación inferior a un límite de probabilidad predeterminado. En otra realización más, los métodos incluyen para cada supuesto par afín identificado, determinar una estimación de la tasa de falso descubrimiento para un posible emparejamiento falso de la secuencia única de amplicón del receptor inmune del primer adaptador único y la secuencia única de amplicón del receptor inmune del segundo adaptador único; e identificar una pluralidad de pares afines de secuencias únicas de receptores inmunes adaptativos primero y segundo como pares afines verdaderos que codifican los receptores inmunes adaptativos en la muestra basándose en la probabilidad estadística y la estimación de la tasa de falso descubrimiento.

En otra realización, la pluralidad de las primeras secuencias de amplicón del receptor inmunitario adaptativo incluye una secuencia de codificación de región variable (V) única, una secuencia de codificación de región J única o una secuencia de codificación de región J única y una secuencia de codificación de región C única, al menos una secuencia de código de barras, al menos una secuencia de adaptador universal, y una secuencia de marcaje de plataforma de secuenciación, y la pluralidad de segundas secuencias de amplicón de receptor inmunitario adaptativo incluye una secuencia de codificación de región V única, una secuencia de codificación de región J única o una secuencia de codificación de región J única y una secuencia de codificación de región C única, al menos una secuencia de código de barras, al menos una secuencia de adaptador universal y una secuencia de marcaje de plataforma de secuenciación.

En otras realizaciones, la pluralidad de muestras de la fuente puede incluir muestras biológicas de diferentes sujetos humanos. En otras realizaciones más, las muestras biológicas pueden provenir de sangre completa, muestras de tejido sólido, tejidos cancerosos o no cancerosos. En otras realizaciones, la pluralidad de muestras de la fuente puede incluir de aproximadamente 10 a aproximadamente 100 muestras, de aproximadamente 100 a 1000 muestras, o aproximadamente 100 muestras. En otra realización, cada envase tiene un número de células sustancialmente equivalente. En otra realización más, hay al menos 10^4 células en cada envase.

Breve descripción de las diversas vistas de los dibujos

La **Figura 1** representa los resultados de un ejemplo de experimento de emparejamiento de TCRA/TCRB realizado en un conjunto combinado de muestras de tumores humanos.
La **Figura 2** representa un análisis de tasa de falsos descubrimientos realizado en el experimento representado en la Figura 1.
La **Figura 3** representa la tasa de falsos descubrimientos (TFD) predicha frente a la empírica para 18 muestras tumorales multiplexadas. La TFD predicha se suministra por un modelo estadístico, y la TFD empírica se estima por la fracción de pares de muestras cruzadas. La TFD empírica se calculó añadiendo los resultados de tres pares de placas pairSEQ replicadas.

La **Figura 4** representa histogramas de rendimientos de emparejamiento para 18 muestras entre los clones más frecuentes en cada tumor. Los 10 mejores clones (**Figura 4A**) y los 100 mejores clones (**Figura 4B**) se identificaron como las secuencias de repertorio de TCRB más comunes que tenían reordenamientos de VDJ en el marco y se observaron en los datos de ADNc de pairSEQ (expresados). Cada histograma incluye 18 puntos de datos (uno por muestra de tumor).

Descripción detallada de la invención

La presente invención proporciona métodos inesperadamente ventajosos para el emparejamiento simultáneo preciso y eficaz de secuencias de receptores inmunes adaptativos de un gran número de muestras biológicas. En ciertas realizaciones, los métodos de la presente invención pueden aplicarse a múltiples muestras biológicas. En algunas realizaciones, las muestras se obtienen de diferentes fuentes y contienen células de interés (por ejemplo, linfocitos T o linfocitos B).

De acuerdo con una realización de los métodos de la invención, se realiza una PCR multiplexada y un experimento de secuenciación de alto rendimiento en cada una de las muestras de la fuente para determinar las secuencias de ácido nucleico reorganizadas que codifican las regiones determinantes de complementariedad (CDR) de una sola (es decir, una "primera") cadena de polipéptidos de los heterodímeros del receptor inmunitario adaptativo presentes en una muestra. En ciertas realizaciones, esto se denomina "secuenciación de un solo locus". En algunas realizaciones, cada una de las primeras cadenas de polipéptidos de los heterodímeros del receptor inmunitario adaptativo se observa en una y solo una muestra. Cada una de las primeras secuencias de ácido nucleico reorganizadas determinadas se asigna a su respectiva muestra de la fuente.

En otra etapa, se realiza un único ensayo de emparejamiento para determinar pares afines de secuencias primera y segunda de ácido nucleico reorganizadas que codifican las CDR de las cadenas polipeptídicas primera y segunda de los heterodímeros del receptor inmunitario adaptativo. Las submuestras de las diferentes muestras de la fuente se agrupan para proporcionar una población combinada de células. La población combinada de células se distribuye en una pluralidad de pocillos. En cada pocillo, se realiza una PCR multiplexada para amplificar las secuencias de ácido nucleico reorganizadas que codifican las regiones determinantes de complementariedad (CDR) de la primera y segunda cadena de polipéptidos de heterodímeros de receptores inmunes adaptativos. Los amplicones de cada pocillo se secuencian usando métodos de secuenciación de alto rendimiento. Basado en las lecturas de secuencia determinadas de cada uno de los pocillos, se usa un método de emparejamiento para determinar qué secuencias de ácido nucleico reorganizadas codifican la primera y segunda cadenas de polipéptidos de un heterodímero de receptor inmunitario adaptativo.

Los pares de secuencias afines resultantes se asignan a una sola muestra de la fuente de origen, entre la totalidad de las muestras. Esto se logra comparando las primeras secuencias de ácido nucleico reorganizadas determinadas en cada una de las muestras de la fuente no combinadas con las primeras secuencias de ácido nucleico reorganizadas determinadas en la población combinada de células para asignar cada primera secuencia de ácido nucleico reorganizada presente en la población combinada de células a la muestra de la fuente única en la que se expresa de manera única entre todas las muestras de la fuente. En otras palabras, se determina una coincidencia entre una primera secuencia de ácido nucleico reordenada obtenida de una muestra de la fuente y una primera secuencia de ácido nucleico reordenada asignada a un par afín en la población combinada de células. Después, para cada primera secuencia de ácido nucleico reordenada de la muestra combinada que se asigna a una muestra fuente única, su segunda secuencia de ácido nucleico reordenada afín determinada a partir de la muestra combinada también se asigna a la misma muestra de la fuente única.

Por lo tanto, en esta realización, la presente invención proporciona ventajosamente un método en el que se puede usar un único ensayo de emparejamiento para determinar los polipéptidos heterodímeros del receptor inmunitario adaptativo emparejado presentes en un gran número de muestras diferentes. Los métodos de la presente invención proporcionan un aumento significativo en la eficacia sobre los métodos de emparejamiento conocidos en la técnica, en el que la información de emparejamiento se limita a los heterodímeros del receptor inmunitario adaptativo expresados en una sola muestra de fuente biológica. La ausencia de "multiplexación de muestras" en los métodos descritos anteriormente requiere que cada muestra de interés sea analizada en un experimento de emparejamiento separado. Esta deficiencia da como resultado un aumento proporcional en el tiempo y el coste del análisis de la muestra a medida que aumenta el número de muestras a procesar. En gran contraste, los métodos de la presente invención permiten el análisis de emparejamiento simultáneo de un número significativo de muestras únicas, lo que reduce drásticamente la entrada de tiempo y recursos y mejora la eficacia. Estas y otras realizaciones de la presente invención se describen en mayor detalle en el presente documento.

Definiciones

Los términos utilizados en las reivindicaciones y en la memoria descriptiva se definen tal como se establece a continuación, a menos que se especifique lo contrario.

Tal y como se usa en el presente documento, el receptor inmune adaptativo (AIR, por sus siglas en inglés) se refiere

a un receptor de células inmunitarias, por ejemplo, un receptor de linfocitos T (TCR) o un receptor de inmunoglobulina (Ig) hallado en células de mamíferos. En ciertas realizaciones, el receptor inmune adaptativo está codificado por un gen o segmento génico TCRB, TCRG, TCRA, TCRD, IGH, IGK y IGL.

El término "cebador", tal y como se usa en el presente documento, se refiere a una secuencia de oligonucleótidos capaz de actuar como un punto de inicio de la síntesis de ADN en condiciones adecuadas. Un cebador es complementario (o hibrida con) un molde diana (por ejemplo, un molde de ADN, ADNc o ARNm). Dichas condiciones incluyen aquellas en las que se induce la síntesis de un producto de extensión de cebador complementario a una cadena de ácido nucleico en presencia de cuatro nucleósidos trifosfatos diferentes y un agente de extensión (por ejemplo, una ADN polimerasa o transcriptasa inversa) en un tampón apropiado y a una temperatura adecuada.

En algunas realizaciones, tal y como se usa en el presente documento, el término "gen" se refiere al segmento de ADN involucrado en la producción de una cadena polipeptídica, tal como todo o una porción de un polipéptido del TCR o de la Ig (por ejemplo, un polipéptido que contiene CDR3); incluye regiones que preceden y siguen a la región de codificación "líder y de avance", así como secuencias intermedias (intrones) entre segmentos de codificación individuales (exones), elementos reguladores (por ejemplo, promotores, potenciadores, sitios de unión a represores y similares), o secuencias señal de recombinación (RSS), tal como se describe en el presente documento.

Los ácidos nucleicos de las presentes realizaciones, también denominado en el presente documento como polinucleótidos, e incluyendo oligonucleótidos, puede estar en forma de ARN o en forma de ADN, incluyendo ADNc, ADN genómico y ADN sintético. El ADN puede ser bicatenario o monocatenario, y si es monocatenario puede ser la cadena codificante o la cadena no codificante (antisentido). Una secuencia de codificación que codifica un TCR o una Ig o una región del mismo (por ejemplo, una región V, un segmento D, una región J, una región C, etc.) para su uso de acuerdo con las presentes realizaciones puede ser idéntica a la secuencia codificante conocida en la técnica para cualquier región de gen del TCR o de la inmunoglobulina o de dominios de polipéptidos dados (por ejemplo, los dominios de la región V, los dominios CDR3, etc.), o puede ser una secuencia codificante diferente, que como resultado de la redundancia o degeneración del código genético, codifica la misma región o polipéptido del TCR o de la inmunoglobulina.

El término porcentaje de "identidad", en el contexto de dos o más ácidos nucleicos o secuencias polipeptídicas, se refiere a dos o más secuencias o subsecuencias que tienen un porcentaje específico de nucleótidos o de restos de aminoácidos que son iguales, cuando se compara y se alinea para una correspondencia máxima, tal como se mide utilizando uno de los algoritmos de comparación de secuencias que se describen a continuación (por ejemplo, BLASTP y BLASTN u otros algoritmos disponibles para personas expertas) o mediante inspección visual. Dependiendo de la aplicación, el porcentaje de "identidad" puede existir sobre una región de la secuencia que se compara, por ejemplo, sobre un dominio funcional, o, como alternativa, existen en toda la longitud de las dos secuencias a comparar.

Para comparar secuencias, típicamente una secuencia actúa como una secuencia de referencia con la cual se comparan las secuencias de prueba. Cuando se usa un algoritmo de comparación de secuencias, las secuencias de prueba y de referencia se introducen en un ordenador, se designan las coordenadas de subsecuencia, en caso necesario, y se designan los parámetros del programa del algoritmo de secuencias. A continuación, el algoritmo de comparación de secuencias calcula el porcentaje de identidad de secuencia para la(s) secuencia(s) problema con respecto a la secuencia de referencia, basándose en los parámetros designados del programa.

Se puede realizar una alineación óptima de secuencias para la comparación, por ejemplo, por el algoritmo de homología local de Smith y Waterman, Adv. Appl. Math. 2:482 (1981), mediante el algoritmo de alineación de homología de Needleman y Wunsch, J. Mol. Biol. 48:443 (1970), mediante la búsqueda del método de similitud de Pearson y Lipman, Proc. Nat'1. Acad. Sci. EE.UU. 85:2444 (1988), mediante implementaciones computarizadas de estos algoritmos (GAP, BESTFIT, FASTA y TFASTA en el paquete informático de Wisconsin Genetics, Genetics Computer Group, 575 Science Dr., Madison, Wis.), o mediante inspección visual (véase de manera general Ausubel et al., citado a continuación).

Un ejemplo de un algoritmo que es adecuado para determinar el porcentaje de identidad de secuencia y la similitud de secuencia es el algoritmo BLAST, que se describe en Altschul et al., J. Mol. Biol. 215:403-410 (1990). El programa informático para realizar el análisis BLAST está disponible públicamente a través del Centro Nacional de Información Biotecnológica (www.ncbi.nlm.nih.gov/).

La expresión "cantidad suficiente" significa una cantidad suficiente para producir un efecto deseado, por ejemplo, una cantidad suficiente para modular la agregación de proteínas en una célula.

La expresión "cantidad terapéuticamente eficaz" es una cantidad que es eficaz para mejorar un síntoma de una enfermedad. Una cantidad terapéuticamente eficaz puede ser una "cantidad profilácticamente eficaz", ya que la profilaxis puede considerarse terapia.

A menos que se proporcionen definiciones específicas, la terminología utilizada vinculada con, y los procedimientos

de laboratorio y las técnicas de, biología molecular, química analítica, química orgánica sintética, y química médica y farmacéutica descritas en el presente documento son aquellas bien conocidas y comúnmente utilizadas en la materia. Se pueden utilizar técnicas convencionales para tecnología recombinante, biología molecular, microbiología, síntesis química, análisis químico, preparación, formulación y administración de compuestos farmacéuticos, y tratamiento de pacientes.

A no ser que el contexto requiera lo contrario, a lo largo de la presente memoria descriptiva y de las reivindicaciones, la palabra "comprende" y sus variaciones, tales como, "comprende" y "que comprende" deben interpretarse de forma abierta, sentido inclusivo, esto es, como "incluyendo, pero sin limitación". Por "que consiste en" se entiende que incluye, y normalmente se limita a, lo que sigue en la frase "que consiste en". Por "que consiste esencialmente en" se entiende que incluye cualquiera elementos enumerados después de la frase, y se limita a otros elementos que no interfieren con o contribuyen a la actividad o acción especificada en la divulgación de los elementos enumerados. Por lo tanto, la frase "que consiste esencialmente en" indica que los elementos enumerados son necesarios u obligatorios, pero que no se requieren otros elementos y pueden o no pueden estar presentes dependiendo de si afectan o no a la actividad o acción de los elementos enumerados.

Cabe señalar que, tal y como se usa en la memoria descriptiva y en las reivindicaciones adjuntas, las formas en singular "un", "una" y "el" o "la" incluyen referentes plurales salvo que el contexto indique claramente lo contrario.

Tal y como se usa en el presente documento, en realizaciones particulares, los términos "sobre" o "aproximadamente" cuando preceden a un valor numérico indican el valor más o menos un intervalo del 5 %, 6 %, 7 %, 8 % o 9 %, o mayor, *etc.* En otras realizaciones, los términos "sobre" o "aproximadamente" cuando preceden a un valor numérico indican el valor más o menos un intervalo del 10 %, 11 %, 12 %, 13 % o 14 %, o mayor, *etc.* En otras realizaciones más, los términos "sobre" o "aproximadamente" cuando preceden a un valor numérico indican el valor más o menos un intervalo del 15 %, 16 %, 17 %, 18 %, 19 % o 20 %, o mayor, *etc.*

La referencia a lo largo de la presente memoria descriptiva a "una realización" o "una realización" o "un aspecto" significa que un aspecto, estructura o característica particular, descrita en relación con la realización se incluye en al menos una realización de la presente invención. Por lo tanto, las apariciones de las frases "en una realización" o "en una realización" en varios lugares a lo largo de la presente memoria descriptiva no se refieren necesariamente a la misma realización. Además, los aspectos, estructuras o características particulares se pueden combinar de cualquier manera adecuada en una o más realizaciones.

Métodos de la invención

1. Muestras y células

En ciertas realizaciones, los métodos de la presente invención se dirigen a los métodos para asignar un par de polipéptidos primero y segundo que forman un receptor de linfocitos T (TCR) o inmunoglobulina (Ig) a una única muestra de la fuente entre una pluralidad de muestras de la fuente. En algunas realizaciones, la pluralidad de muestras de la fuente posee cada una, una composición genética distinta, es decir, cada una de las muestras de la fuente proviene de un sujeto diferente que posee un repertorio único de TCR o de Ig. En ciertas realizaciones, los diferentes sujetos pueden ser mamíferos humanos o no humanos. En una realización preferida, los sujetos son sujetos humanos cuyos TCR o Ig son de interés.

Tal como se describe en el presente documento, "una pluralidad" de muestras de la fuente puede comprender de aproximadamente decenas a cientos de muestras de la fuente. Un experto en la materia reconocerá que el número exacto de muestras de la fuente dependerá de la aplicación particular de los métodos reivindicados y, por lo tanto, se pretende que sea variable. En ciertas realizaciones, la pluralidad de muestras de la fuente puede comprender de aproximadamente 10 a aproximadamente 100 muestras. En otras realizaciones, la pluralidad de muestras de la fuente puede comprender de aproximadamente 100 a aproximadamente 1000 muestras. En otras realizaciones más, la pluralidad de muestras de la fuente puede comprender aproximadamente 100 muestras.

Cualquier tejido periférico puede ser una fuente para tomar muestras de la presencia de linfocitos B o T y, por lo tanto, se contempla su uso en los métodos descritos en el presente documento. Las muestras biológicas pueden incluir, pero no se limitan a una muestra de tejido tumoral sólido, una muestra de biopsia, piel, tejidos epiteliales, colon, bazo, una secreción mucosa, mucosa oral, mucosa intestinal, mucosa vaginal o una secreción vaginal, tejido del cuello uterino, ganglios, saliva, líquido cefalorraquídeo (LCR), médula ósea, sangre de cordón, suero, fluido seroso, plasma, linfa, orina, fluido de ascitis, fluido pleural, fluido pericardial, fluido peritoneal, fluido abdominal, medio de cultivo, medio de cultivo acondicionado o fluido de lavado. La sangre periférica puede ser un tejido preferido ya que se accede fácilmente. Se pueden obtener muestras de sangre periférica por flebotomía de los sujetos. Las células mononucleares de sangre periférica (PBMC) se aíslan mediante técnicas conocidas por los expertos en la materia, por ejemplo, por separación de gradiente de densidad Ficoll-Hypaque®. En ciertas realizaciones, se utilizan PBMC completas para el análisis.

En otras realizaciones, la muestra de la fuente puede ser de un tejido canceroso, tal como una muestra de tumor

sólido de una biopsia de tumor de piel u órgano. El tumor puede ser de sarcomas, carcinomas o linfomas. Los ejemplos incluyen cáncer de ovario, cáncer de mama, cáncer de próstata, cáncer de pulmón, cáncer de hígado, cáncer de páncreas y melanoma, y similares.

Otros ejemplos de muestras fuente incluyen orina, líquido amniótico que rodea a un feto, humor acuoso, bilis, sangre y plasma sanguíneo, cerumen (cera de oído), líquido de Cowper o líquido preeyaculatorio, quilo, quimo, eyaculado femenino, líquido intersticial, linfa, menstruo, leche materna, moco (incluyendo moco y flema), fluido pleural, pus, saliva, sebo (grasa de la piel), semen, suero, sudor, lágrimas, lubricación vaginal, vómito, agua, heces, fluidos corporales internos, incluyendo el líquido cefalorraquídeo que rodea el cerebro y la médula espinal, líquido sinovial que rodea las articulaciones óseas, el líquido intracelular es el líquido dentro de las células, y el humor vítreo los fluidos en el globo ocular, o líquido cefalorraquídeo (LCR).

La muestra de la fuente se puede obtener por un proveedor de atención médica, por ejemplo, un médico, asistente médico, enfermero, veterinario, dermatólogo, reumatólogo, dentista, paramédico, cirujano o técnico de investigación. Se puede obtener más de una muestra de un sujeto.

La muestra de la fuente puede ser una biopsia. La biopsia puede ser, de, por ejemplo, piel, ovario, mama, 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 de la fuente incluye 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 o linfocitos Th efectores), linfocitos T citotóxicos (LTC), linfocitos T de memoria y linfocitos T reguladores. La muestra puede incluir una sola célula en algunas aplicaciones (por ejemplo, una prueba de calibración para definir los linfocitos T relevantes) o, en general, al menos 1.000, al menos 10.000, al menos 100.000, al menos 250.000, al menos 500.000, al menos 750.000, o al menos 1.000.000 de linfocitos T.

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). La muestra puede incluir una sola célula en algunas aplicaciones (por ejemplo, una prueba de calibración para definir los linfocitos B relevantes) o, en general, al menos 1.000, al menos 10.000, al menos 100.000, al menos 250.000, al menos 500.000, al menos 750.000, o al menos 1.000.000 de linfocitos B.

En ciertas realizaciones relacionadas, se pueden preparar preparaciones que comprenden predominantemente linfocitos (por ejemplo, linfocitos T y B) o que comprenden predominantemente linfocitos T o predominantemente linfocitos B. En otras realizaciones relacionadas, se pueden aislar subpoblaciones específicas de linfocitos T o B antes del análisis utilizando los métodos descritos en el presente documento. En la técnica se conocen diversos métodos y kits disponibles comercialmente para aislar diferentes subpoblaciones de linfocitos T y B e incluyen, aunque sin limitación, la selección de subconjuntos por separación inmunomágnetica con perlas o la clasificación inmunocitométrica de flujo de células usando anticuerpos específicos para uno o más de una variedad de marcadores de superficie de linfocitos T y B conocidos. Los marcadores ilustrativos incluyen, aunque sin limitación, uno o una combinación de CD2, CD3, CD4, CD8, CD14, CD19, CD20, CD25, CD28, CD45RO, CD45RA, CD54, CD62, CD62L, Cdw137 (41BB), CD154, GITR, FoxP3, CD54 y CD28. Por ejemplo, y como sabe la persona experta, los marcadores de superficie celular, tales como CD2, CD3, CD4, CD8, CD14, CD19, CD20, CD45RA y CD45RO pueden usarse para determinar el linaje T, B, y de monocitos y las subpoblaciones en citometría de flujo. De manera similar, dispersión de luz directa, los marcadores de dispersión lateral y/o de superficie celular como CD25, CD62L, CD54, CD137 y CD154 pueden usarse para determinar el estado de activación y las propiedades funcionales de las células.

Las combinaciones ilustrativas útiles en algunos de los métodos descritos en el presente documento pueden incluir $CD8^+CD45RO^+$ (linfocitos T citotóxicos de memoria), $CD4^+CD45RO^+$ (T colaboradores de memoria), $CD8^+CD45RO^-$ ($CD8^+CD62L^+CD45RA^+$ (linfocitos T citotóxicos sin exposición previa); $CD4^+CD25^+CD62L^{hi}GITR^+FoxP3^+$ (linfocitos T reguladores). Los anticuerpos ilustrativos para su uso en separaciones inmunomagnéticas de células o en la clasificación inmunocitométrica de flujo de células incluyen anticuerpos antihumanos marcados con fluorescencia, por ejemplo, CD4 FITC (clon M-T466, Miltenyi Biotec), CD8 PE (clon RPA-T8, BD Biosciences), CD45RO ECD (clon UCHL-1, Beckman Coulter), y CPA CD45RO (clon UCHL-1, BD Biosciences). La tinción de PBMC totales se puede hacer con la combinación apropiada de anticuerpos, seguida del lavado de las células antes del análisis. Los subconjuntos de linfocitos pueden aislarse mediante clasificación celular activada por fluorescencia (FACS), por ejemplo, mediante un sistema de clasificación de células BD FACSAria ™ (BD Biosciences) y mediante el análisis de resultados con el programa informático FlowJo ™ (Treestar Inc.), y también mediante métodos conceptualmente similares que implican anticuerpos específicos inmovilizados en superficies o perlas.

Tal como se describe en el presente documento, las muestras de la fuente pueden agruparse para formar una población combinada de células. En ciertas realizaciones, el volumen completo de cada muestra se agrupa para formar la población combinada, mientras que en otras realizaciones, se agrupan porciones de cada muestra para formar la población combinada. Un experto en la materia reconocerá que el volumen de cada muestra agrupada dependerá del número de células de cada muestra deseada para formar la población combinada de células. El número de células deseado de cada muestra y el número de células en la población combinada dependerá del diseño del experimento particular y se pretende que sea un número flexible. En ciertas realizaciones, el número de células presentes en la muestra combinada es de aproximadamente 10^5 a unos 10^7. En otra realización, el número de células en la muestra combinada es de aproximadamente 10^6.

La muestra de la fuente puede incluir ácido nucleico, por ejemplo, ADN (por ejemplo, ADN genómico o ADN mitocondrial) o ARN (por ejemplo, ARN mensajero o microARN). El ácido nucleico puede ser ADN o ARN sin células. En los métodos de la invención que se proporciona, la cantidad de ARN o ADN de un sujeto que puede analizarse incluye, por ejemplo, tan bajo como una 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 algunas realizaciones, el ADN genómico total se puede extraer de las células mediante métodos conocidos por los expertos en la materia. Los ejemplos incluyen el uso del mini kit de ADN sanguíneo QIAamp® (QIAGEN®). La masa aproximada de un genoma haploide único es de 3 pg. Preferentemente, se utilizan al menos 100.000 a 200.000 células para el análisis de la diversidad, es decir, aproximadamente de 0,6 a 1,2 µg de ADN de linfocitos T diploides. Con el uso de PBMC como fuente, el número de linfocitos T puede estimarse en aproximadamente el 30 % del total de células.

En algunas realizaciones, el ARN se puede extraer de las células en una muestra, tal como una muestra de sangre, linfa, tejido u otra muestra de un sujeto que se sabe que contiene células linfoides, utilizando métodos convencionales o kits disponibles comercialmente conocidos en la técnica. En otras realizaciones, el ADNc se puede transcribir a partir del ARNm obtenido de las células y luego usarse como moldes en una PCR multiplexada.

Como alternativa, el ácido nucleico total puede aislarse de las células, incluyendo tanto ADN genómico como ARNm. Si la diversidad se va a medir a partir de ARNm en el extracto de ácido nucleico, el ARNm se puede convertir en ADNc antes de la medición. Esto se puede hacer fácilmente por los métodos de un experto en la materia, por ejemplo, utilizando transcriptasa inversa de acuerdo con procedimientos conocidos.

2. Caracterización de alto rendimiento de ácidos nucleicos reorganizados que codifican cadenas de polipéptidos individuales de heterodímeros de receptores inmunes adaptativos (secuenciación de locus único)

En ciertas realizaciones, los métodos de la presente invención incluyen la etapa de determinar las primeras secuencias de ácido nucleico reorganizadas que codifican los primeros polipéptidos de los heterodímeros del TCR o de la Ig presentes en cada una de una pluralidad de muestras de la fuente. En Robins et al., se describen métodos para la detección cuantitativa de secuencias de sustancialmente todos los posibles reordenamientos de genes del TCR o de la Ig presentes en una muestra que contiene ADN de células linfoides., 2009 Blood 114, 4099; Robins et al., 2010 Sci. Translat. Med. 2:47ra64; Robins et al., 2011 J. Immunol. Meth. doi:10.1016/j.jim.2011.09. 001; Sherwood et al. 2011 Sci. Translat. Med.3:90ra61; US Pub. N.º 2012/0058902, US Pub. N.º 2010/0330571, WO/2010/151416, W0/2011/106738, W02012/027503. La presente invención no pretende limitarse a ningún método y contempla que muchos métodos conocidos en la técnica pueden ser adecuados para poner en práctica la invención reivindicada. En realizaciones preferidas, los métodos para determinar la diversidad del repertorio de TCR y/o Ig son aquellos descritos en solicitudes pendientes de los solicitantes, US2010/0330571, presentada el 4 de junio de 2010, US2012/0058902, presentada el 24 de agosto de 2011 y W02014/055561, presentada el 1 de octubre de 2013. A modo de ilustración pero no de limitación, una realización ejemplar de los métodos de la invención se resume en el presente documento a continuación.

Los métodos de esta realización de la invención incluyen 1) la construcción sofisticada de cebadores y métodos para la amplificación de la reacción en cadena de polimerasa (PCR) multiplexada, controlada e imparcial de todas las regiones CDR3 posibles que podrían estar presentes en el ADN genómico (o ADNc) derivado de un locus de receptor inmunitario dado (Ig o TCR) dentro de cada linfocito en una muestra de sangre, de médula ósea o de tejido, 2) la secuenciación de firma paralela masiva de alto rendimiento de los productos amplificados, y 3) el análisis computacional refinado y formidable de la salida de datos de secuencia sin procesar para eliminar el "ruido", extraer la señal, solucionar problemas de artefactos tecnológicos y validar el control del proceso desde la recepción de muestras hasta la entrega de secuencias.

Los métodos actuales implican un único método de PCR multiplexada que utiliza un conjunto de cebadores directos que hibridan específicamente con segmentos V y un conjunto de cebadores inversos que hibridan específicamente con los segmentos J de un locus de TCR o de IG, en donde una única reacción de PCR multiplexada usando los

cebadores permite la amplificación de todas las combinaciones posibles de VJ (y VDJ) dentro de una población dada de linfocitos T o B.

Los ejemplos de cebadores de segmento V y segmento J se describen en los documentos US2012/0058902, US2010/0330571, WO/2010/151416, WO/2011/106738, WO2012/027503.

Se puede usar un único sistema de PCR multiplexada para amplificar los loci de receptores de células inmunes adaptativas reorganizados a partir de ADN genómico, preferentemente de una región CDR3. En ciertas realizaciones, la región CDR3 se amplifica a partir de una región CDR3 de TCRα, TCRβ, TCRγ o TCRδ o similar de un locus de IgH o IgL (lambda o kappa). Se proporcionan composiciones que comprenden una pluralidad de cebadores del segmento V y del segmento J que son capaces de promover la amplificación en una reacción en cadena de polimerasa (PCR) multiplexada de sustancialmente todas las regiones que codifican el receptor inmunitario adaptativo CDR3 reorganizado productivamente en la muestra para una clase dada de tales receptores (por ejemplo, TCRγ, TCRβ, IgH, etc.) para producir una multiplicidad de moléculas de ADN reorganizadas amplificadas a partir de una población de linfocitos T (para TCR) o linfocitos B (para Ig) en la muestra. En ciertas realizaciones, los cebadores están diseñados para que cada molécula de ADN reordenada amplificada tenga menos de 600 nucleótidos de longitud, excluyendo así los productos de amplificación de loci de receptores inmunes adaptativos no reorganizados.

En algunas realizaciones, el método utiliza dos grupos de cebadores para proporcionar una reacción de PCR altamente multiplexada de un solo tubo. Un conjunto de cebadores "directos" puede incluir una pluralidad de cebadores oligonucleotídicos del segmento V usados como cebadores "directos" y una pluralidad de cebadores oligonucleotídicos del segmento J usados como cebadores "inversos". En otras realizaciones, los cebadores del segmento J se pueden usar como cebadores "directos" y el cebador del segmento V se pueden usar cebadores "inversos". En algunas realizaciones, se puede usar un cebador oligonucleotídico que es específico para (por ejemplo, que tiene una secuencia de nucleótidos complementaria a una región de secuencia única de) cada segmento codificante de la región V ("segmento V") en el respectivo locus del gen de TCR o de Ig. En otras realizaciones, los cebadores dirigidos a una región altamente conservada se usan para amplificar simultáneamente múltiples segmentos V o múltiples segmentos J, reduciendo así el número de cebadores requeridos en la PCR multiplexada. En ciertas realizaciones, los cebadores del segmento J se combinan con una secuencia conservada en el segmento de unión ("J").

Cada cebador puede diseñarse de manera que se obtenga un segmento de ADN amplificado respectivo que incluya una porción de secuencia de longitud suficiente para identificar cada segmento J sin ambigüedad en función de las diferencias de secuencia entre segmentos de genes que codifican la región J conocidos en la base de datos del genoma humano, y también para incluir una porción de secuencia con la que puede hibridar un cebador específico del segmento J para la resecuenciación. Este diseño de cebadores específicos del segmento V y J permite la observación directa de una gran fracción de los reordenamientos somáticos presentes en el repertorio de genes del receptor inmunitario adaptativo dentro de un individuo. Esta característica a su vez permite una comparación rápida de los repertorios de TCR y/o de Ig en individuos antes del trasplante y después del trasplante, por ejemplo.

En una realización, la presente divulgación proporciona una pluralidad de cebadores del segmento V y una pluralidad de cebadores del segmento J, en donde la pluralidad de cebadores del segmento V y la pluralidad de cebadores del segmento J amplifican todas o sustancialmente todas las combinaciones de los segmentos V y J de un locus de receptor inmunitario reordenado. En algunas realizaciones, el método proporciona amplificación de sustancialmente todas las secuencias del receptor inmunitario adaptativo (AIR) reorganizadas en una célula linfoide y es capaz de cuantificar la diversidad del repertorio de TCR o IG de al menos 10^6, 10^5, 10^4 o 10^3 secuencias AIR reordenadas únicas en una muestra. "Sustancialmente todas las combinaciones" puede referirse al menos al 90 %, 91 %, 92 %, 93 %, 94 %, 95 %, 96 %, 97 %, 98 %, 99 % o más de todas las combinaciones de los segmentos V y J de un locus del receptor inmunitario reordenado. En ciertas realizaciones, la pluralidad de cebadores del segmento V y la pluralidad de cebadores del segmento J amplifican todas las combinaciones de los segmentos V y J de un locus de receptor inmunitario adaptativo reordenado.

En general, un sistema de PCR multiplexada puede usar 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24 o 25, y en ciertas realizaciones, al menos 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38 o 39, y en otras realizaciones 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 65, 70, 75, 80, 85 o más cebadores directos, en donde cada cebador directo hibrida específicamente o es complementario con una secuencia correspondiente a uno o más segmentos de la región V. El sistema de PCR multiplexada también utiliza al menos 2, 3, 4, 5, 6 o 7, y en ciertas realizaciones, 8, 9, 10, 11, 12 o 13 cebadores inversos, o 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24 o 25 o más cebadores, en donde cada cebador inverso hibrida específicamente con o es complementario con una secuencia correspondiente a uno o más segmentos de la región J. En algunas realizaciones, cada cebador J inverso es específico de un segmento génico J diferente. En otras realizaciones, no hay un cebador J común que se una a todos los segmentos génicos J.

Se pueden usar varias combinaciones de cebadores del segmento V y J para amplificar la diversidad completa de secuencias TCR e IG en un repertorio. Para obtener detalles sobre el sistema de PCR multiplexada, incluyendo

secuencias de oligonucleótidos cebadores para amplificar sustancialmente todas las secuencias de TCR e IG, véase, por ejemplo, Robins et al., 2009 Blood 114, 4099; Robins et al., 2010 Sci. Translat. Med. 2:47ra64; Robins et al., 2011 J. Immunol. Meth. doi:10.1016/j.jim.2011.09. 001; Sherwood et al. 2011 Sci. Translat. Med. 3:90ra61; US2012/0058902, US2010/033571, WO/2010/151416, WO/2011/106738, WO2012/027503.

Los oligonucleótidos o polinucleótidos que son capaces de hibridar o de hibridar específicamente con una secuencia de ácido nucleico diana mediante la complementariedad de las bases de nucleótidos pueden hacerlo en condiciones de rigurosidad moderada a alta. Para fines de ilustración, las condiciones de rigurosidad moderada a alta adecuadas para la amplificación por PCR específica de una secuencia de ácido nucleico diana estarían entre 25 y 80 ciclos de PCR, con cada ciclo que consiste en una etapa de desnaturalización (por ejemplo, aproximadamente 10-30 segundos (s) a más de aproximadamente 95 °C), una etapa de hibridación (por ejemplo, de aproximadamente 10-30 s a aproximadamente 60-68 °C), y una etapa de extensión (por ejemplo, de aproximadamente 10-60 s a aproximadamente 60-72 °C), opcionalmente de acuerdo con ciertas realizaciones con las etapas de hibridación y extensión combinadas para proporcionar una PCR de dos etapas. Como reconocería un experto en la materia, se pueden agregar o cambiar otros reactivos de PCR en la reacción de PCR para aumentar la especificidad de la hibridación y la amplificación del cebador, tal como alterar la concentración de magnesio, opcionalmente añadiendo DMSO, y/o el uso de cebadores bloqueados, nucleótidos modificados, ácidos nucleicos peptídicos, y similares.

En ciertas realizaciones, se pueden usar las técnicas de hibridación de ácido nucleico para evaluar la especificidad de hibridación de los cebadores descritos en el presente documento. Las técnicas de hibridación son bien conocidas en la técnica de la biología molecular. Para fines de ilustración, las condiciones moderadamente rigurosas adecuadas para probar la hibridación de un polinucleótido como se proporciona en el presente documento con otros polinucleótidos incluyen prelavado en una solución de 5 X SSC, SDS al 0,5 %, EDTA 1,0 mM (pH 8,0); hibridación a 50 °C-60 °C, SSC a 5 X, durante toda la noche; seguido de lavar dos veces a 65 °C durante 20 minutos con cada una de SSC a 2X, 0,5X y 0,2X que contiene SDS al 0,1 %. Un experto en la materia comprenderá que la rigurosidad de la hibridación puede manipularse fácilmente, tal como alterando el contenido de sal de la solución de hibridación y/o la temperatura a la que se realiza la hibridación. Por ejemplo, en otra realización, las condiciones de hibridación altamente rigurosas adecuadas incluyen las descritas anteriormente, con la excepción de que la temperatura de hibridación aumenta, por ejemplo, a 60 °C - 65 °C o 65 °C - 70 °C.

En ciertas realizaciones, los cebadores están diseñados para no cruzar un límite intrón/exón. Los cebadores directos en ciertas realizaciones hibridan con los segmentos V en una región de conservación de secuencia relativamente fuerte entre segmentos V para maximizar la conservación de la secuencia entre estos cebadores. Por consiguiente, esto minimiza el potencial de propiedades de hibridación diferencial de cada cebador, y de modo que la región amplificada entre los cebadores V y J contiene suficiente información de secuencia V de TCR o Ig para identificar el segmento específico V del gen utilizado. En una realización, los cebadores del segmento J hibridan con un elemento conservado del segmento J y tienen una resistencia de hibridación similar. En una realización particular, los cebadores del segmento J hibridan con el mismo motivo de región marco conservado. En ciertas realizaciones, los cebadores del segmento J tienen un intervalo de temperatura de fusión dentro de 10 °C, 7,5 °C, 5 °C, o 2,5 °C o menos.

Los oligonucleótidos (por ejemplo, cebadores) se pueden preparar por cualquier método adecuado, incluyendo la síntesis química directa por un método como el método del fosfotriéster de Narang et al., 1979, Meth. Enzymol. 68:90-99; el método del fosfodiéster de Brown et al., 1979, Meth. Enzymol. 68:109-151; el método de dietilfosforamidita de Beaucage et al., 1981, Tetrahedron Lett. 22:1859-1862; y el método de soporte sólido de la patente de EE.UU. N ° 4.458.066. Se proporciona una revisión de los métodos de síntesis de conjugados de oligonucleótidos y nucleótidos modificados en Goodchild, 1990, Bioconjugate Chemistry 1(3): 165-187.

Un cebador es preferentemente un oligonucleótido monocatenario. La longitud apropiada de un cebador depende del uso previsto del cebador, pero generalmente varía de 6 a 50 nucleótidos, 15-50 nucleótidos, o en ciertas realizaciones, de 15-35 nucleótidos. Las moléculas de cebador cortas generalmente requieren temperaturas más frías para formar complejos híbridos suficientemente estables con el molde. Un cebador no necesita reflejar la secuencia exacta del ácido nucleico del molde, pero debe ser lo suficientemente complementario para hibridar con el molde. El diseño de cebadores adecuados para la amplificación de una secuencia diana dada es bien conocido en la técnica y se describe en la bibliografía citada en el presente documento.

Tal como se describe en el presente documento, los cebadores pueden incorporar características adicionales que permiten la detección o inmovilización del cebador, pero que no alteren la propiedad básica del cebador, la de actuar como punto de inicio de la síntesis de ADN. Por ejemplo, los cebadores pueden contener una secuencia adicional de ácido nucleico en el extremo 5', que no hibrida con el ácido nucleico diana, pero que facilita la clonación, detección o secuenciación del producto amplificado. La región del cebador que es suficientemente complementaria al molde para hibridar se denomina en el presente documento la región de hibridación.

Tal y como se usa en el presente documento, un cebador es "específico" para una secuencia diana si, cuando se usa en una reacción de amplificación en condiciones suficientemente rigurosas, el cebador hibrida principalmente con el ácido nucleico diana. Normalmente, un cebador es específico para una secuencia diana si la estabilidad del

dúplex cebador-diana es mayor que la estabilidad de un dúplex formado entre el cebador y cualquier otra secuencia encontrada en la muestra. Un experto en la materia reconocerá que varios factores, tales como las condiciones de sal, así como la composición base del cebador y la ubicación de los desacoplamientos, afectará a la especificidad del cebador, y esa confirmación experimental de rutina de la especificidad del cebador será necesaria en muchos casos. Se pueden elegir las condiciones de hibridación bajo las cuales el cebador puede formar dúplex estables solo con una secuencia diana. Por lo tanto, el uso de cebadores específicos de diana en condiciones de amplificación adecuadamente rigurosas permite la amplificación selectiva de aquellas secuencias diana que contienen los sitios de unión del cebador a la diana. En otros términos, los cebadores de la invención son cada uno complementario a una secuencia diana y pueden incluir 1,2 o más desacoplamientos sin reducir la complementariedad o la hibridación del cebador con la secuencia diana.

En realizaciones particulares, los cebadores para su uso en los métodos descritos en el presente documento comprenden o consisten en un ácido nucleico de al menos aproximadamente 15 nucleótidos de longitud que tiene la misma secuencia que, o es sustancialmente complementario a, una secuencia contigua de ácido nucleico del segmento V o J diana. Los cebadores más largos, por ejemplo, los de unos 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 45 o 50 nucleótidos de longitud que tienen la misma secuencia o secuencia complementaria a, una secuencia contigua del segmento V o J diana, también será de uso en ciertas realizaciones. Se pueden contemplar varios desacoplamientos (1, 2, 3 o más) con la secuencia diana en los cebadores, mientras se preserva la complementariedad con el segmento V o J diana. Todas las longitudes intermedias de los cebadores mencionados anteriormente se contemplan para su uso en el presente documento. Como reconocería un experto en la materia, los cebadores pueden tener una secuencia adicional agregada (por ejemplo, nucleótidos que no pueden ser iguales o complementarios al segmento V o J diana), tales como sitios de reconocimiento de enzimas de restricción, secuencias adaptadoras para secuenciación, secuencias de códigos de barras y similares (véase, por ejemplo, secuencias de cebadores proporcionadas en el presente documento y en el listado de secuencias). Por lo tanto, la longitud de los cebadores puede ser más larga, tal como de 55, 56, 57, 58, 59, 60, 65, 70, 75 u 80 nucleótidos de longitud o más, dependiendo del uso específico o necesidad.

Por ejemplo, en una realización, los cebadores directo e inverso se modifican en el extremo 5' con la secuencia del cebador directo universal compatible con una secuencia de ácido nucleico de secuenciación de ADN. Dichas secuencias de cebadores universales se pueden adaptar a las utilizadas en el sistema de secuenciación de lectura de extremo único Illumina GAII. Se proporcionan ejemplos de secuencias de cebadores universales y oligonucleótidos de secuenciación en los documentos US2012/0058902, US2010/0330571, WO2012/027503. En algunas realizaciones, los cebadores directo e inverso se modifican en el extremo 5' con una secuencia adaptadora que no es complementaria con el segmento V, el segmento J o el segmento C (secuencia diana) y se puede usar como una región complementaria a un segundo conjunto de cebadores o un oligonucleótido de secuenciación.

Como reconocería un experto en la materia, en ciertas realizaciones, se pueden hacer otras modificaciones a los cebadores, tales como la adición de sitios de enzimas de restricción, marcadores fluorescentes y similares, dependiendo de la aplicación específica.

También se contemplan variantes de cebador oligonucleotídico del segmento V o del segmento J del receptor inmune adaptativo que pueden compartir un alto grado de identidad de secuencia con los cebadores oligonucleotídicos. Por lo tanto, en éstas y en realizaciones relacionadas, las variantes del cebador oligonucleotídico del segmento V o del segmento J del receptor inmunitario adaptativo pueden tener una identidad sustancial frente a las secuencias cebadoras del oligonucleótido del segmento V o del segmento J del receptor inmunitario adaptativo descritas en el presente documento. Por ejemplo, tales variantes de cebador oligonucleotídico pueden comprender al menos un 70 % de identidad de secuencia, preferentemente al menos un 75 %, 80 %, 85 %, 90 %, 91 %, 92 %, 93 %, 94 %, 95 %, 96 %, 97 %, 98 % o 99 % o más de identidad de secuencia en comparación con una secuencia de polinucleótidos de referencia, como las secuencias de cebadores de oligonucleótidos descritas en el presente documento, utilizando los métodos descritos en el presente documento (por ejemplo, Análisis BLAST utilizando parámetros estándar). Un experto en esta técnica reconocerá que estos valores pueden ajustarse adecuadamente para determinar la capacidad correspondiente de una variante de cebador oligonucleotídico para hibridar con un polinucleótido codificador del segmento del receptor inmune adaptativo teniendo en cuenta la degeneración de codones, el posicionamiento del marco de lectura y similares. Normalmente, las variantes de cebador oligonucleotídico contendrán una o más sustituciones, adiciones, deleciones y/o inserciones, preferentemente de tal manera que la capacidad de hibridación la variante de oligonucleótido no disminuya sustancialmente en relación con la de una secuencia de cebador oligonucleotídico del segmento V o del segmento J del receptor inmune adaptativo que se expone específicamente en el presente documento. Como también se señaló en otra parte del presente documento, en realizaciones preferidas, los cebadores oligonucleotídicos del segmento V y del segmento J adaptativos inmunes están diseñados para ser capaces de amplificar una secuencia del TCR o de la IGH reorganizada que incluye la región codificante para CDR3.

De acuerdo con determinadas realizaciones, los cebadores para su uso en los métodos de PCR multiplexada de la presente divulgación pueden bloquearse funcionalmente para evitar el cebado no específico de secuencias no de linfocitos T o B. Por ejemplo, los cebadores se pueden bloquear con modificaciones químicas tal como se describe en la Publicación de los Estados Unidos N.º. 2010/0167353.

En algunas realizaciones, se usan los cebadores del segmento V y J para producir una pluralidad de amplicones a partir de la reacción de PCR multiplexada. En ciertas realizaciones, los cebadores de segmento J y los cebadores del segmento V pueden producir al menos 10^4, 10^5, 10^6 o más amplicones que representan la diversidad de moléculas de CDR3 reorganizadas en TCR o IG en la muestra. En algunas realizaciones, los amplicones varían en tamaño desde 10, 20, 30, 40, 50, 75, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, 1100, 1200, 1300, 1400, 1500 a 1600 nucleótidos en longitud. En realizaciones preferidas, los amplicones tienen un tamaño entre 50-600 nucleótidos de longitud.

Según la teoría no limitativa, estas realizaciones explotan la comprensión actual en la técnica de que una vez una célula inmune adaptativa (por ejemplo, un linfocito T o B) ha reorganizado sus genes codificantes del receptor inmune adaptativo (por ejemplo, TCR o Ig), sus células de progenie poseen el mismo reordenamiento del gen codificante del receptor inmune adaptativo, dando lugar a una población clonal que puede identificarse de forma única por la presencia en ella de segmentos génicos V y J reorganizados (por ejemplo, que codifican CDR3) que pueden amplificarse mediante una combinación por pares específica de cebadores oligonucleotídicos específicos de V y J como se desvela en el presente documento.

Control de sesgo de amplificación

Los ensayos de PCR multiplexada pueden dar lugar a un sesgo en el número total de amplicones producidos a partir de una muestra, dado que ciertos conjuntos de cebadores son más eficaces en la amplificación que otros. Para superar el problema de dicha utilización sesgada de subpoblaciones de cebadores de amplificación, pueden usarse métodos que proporcionan composiciones de moldes sintéticos para estandarizar las eficacias de amplificación de los miembros de un conjunto de cebadores oligonucleotídicos, donde el conjunto de cebadores es capaz de amplificar ADN reordenado que codifica una pluralidad de receptores inmunes adaptativos (TCR o Ig) en una muestra biológica que comprende ADN de células linfoides.

En algunas realizaciones, se utiliza una composición de molde para estandarizar las diversas eficacias de amplificación de los conjuntos de cebadores. La composición de molde puede comprender una pluralidad de oligonucleótidos de molde diversos de fórmula general (I):

5'-U1-B1-V-B2-R-J-B3-U2-3' (I)

Los oligonucleótidos de molde constituyentes, de los cuales se compone la composición del molde, son diversos con respecto a las secuencias de nucleótidos de los oligonucleótidos molde individuales. Los oligonucleótidos molde individuales pueden variar considerablemente en la secuencia de nucleótidos entre sí en función de una variabilidad de secuencia significativa entre el gran número de posibles polinucleótidos de la región variable (V) y la región de unión (J) de TCR o BCR. Las secuencias de especies de oligonucleótidos de molde individuales también pueden variar entre sí en función de las diferencias de secuencia en U1, U2, B (B1, B2 y B3) y R oligonucleótidos que se incluyen en un molde particular dentro de la pluralidad diversa de moldes.

En ciertas realizaciones, V es un polinucleótido que comprende al menos 20, 30, 60, 90, 120, 150, 180 o 210, y no más de 1000, 900, 800, 700, 600 o 500 nucleótidos contiguos de la secuencia de un gen codificante de una región variable (V) del receptor inmunitario adaptativo, o el complemento de la misma, y en cada una de la pluralidad de secuencias de oligonucleótidos del molde de V comprende una secuencia de oligonucleótidos única.

En algunas realizaciones, J es un polinucleótido que comprende al menos 15-30, 31-60, 61-90, 91-120 o 120-150, y no más de 600, 500, 400, 300 o 200 nucleótidos contiguos de la secuencia de un gen codificante de una región de unión (J) o el complemento de la misma, y en cada una de la pluralidad de secuencias de oligonucleótidos del molde J comprende una secuencia de oligonucleótidos única.

U1 y U2 pueden ser cada uno nada o cada uno comprende un oligonucleótido que tiene, de manera independiente, una secuencia que se selecciona de (i) una secuencia de oligonucleótidos con adaptador universal, y (ii) una secuencia de oligonucleótidos específica de plataforma de secuenciación que está unida y posicionada en 5' a la secuencia de oligonucleótidos con adaptador universal.

B1, B2 y B3 pueden ser nada o cada uno comprende un oligonucleótido B que comprende una primera y una segunda secuencia de código de barras de oligonucleótidos de 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 25, 30, 35, 40, 45, 50, 60, 70, 80, 90, 100, 200, 300, 400, 500, 600, 700, 800, 900 o 1000 nucleótidos contiguos (incluidos todos los valores enteros entre ellos), en donde en cada una de la pluralidad de secuencias de oligonucleótidos de molde B comprende una secuencia de oligonucleótidos única en la que (i) la primera secuencia de código de barras identifica de manera única la secuencia de oligonucleótidos V única del oligonucleótido de molde y (ii) la segunda secuencia de código de barras identifica de manera única el oligonucleótido J único secuencia del oligonucleótido de molde.

R puede ser nada o comprende un sitio de reconocimiento de enzimas de restricción que comprende una secuencia

de oligonucleótidos que está ausente de V, J, U1, U2, B1, B2 y B3.

Los métodos se utilizan con la composición de molde para determinar el potencial de amplificación de ácido nucleico no uniforme entre los miembros de un conjunto de cebadores de amplificación de oligonucleótidos que son capaces de amplificar ADN reorganizado productivamente que codifica uno o una pluralidad de receptores inmunes adaptativos en una muestra biológica que comprende ADN de células linfoides de un sujeto. El método puede incluir las etapas de: (a) amplificación de ADN de una composición de molde para estandarizar la eficiencia de amplificación de un conjunto de cebador oligonucleotídico en una reacción en cadena de polimerasa (PCR) multiplexada que comprende: (i) la composición del molde (I) descrita anteriormente, en donde cada oligonucleótido de molde en la pluralidad de oligonucleótidos de molde está presente en una cantidad sustancialmente equimolar; (ii) un conjunto de cebadores de amplificación de oligonucleótidos que es capaz de amplificar de manera productiva ADN reordenado que codifica uno o una pluralidad de receptores inmunes adaptativos en una muestra biológica que comprende ADN de células linfoides de un sujeto.

El conjunto de cebador puede incluir: (1) en cantidades sustancialmente equimolares, una pluralidad de cebadores oligonucleotídicos del segmento V que son independientemente capaces de hibridar específicamente con al menos un polinucleótido que codifica un polipéptido de la región V del receptor inmunitario adaptativo o con el complemento del mismo, en donde cada cebador del segmento V comprende una secuencia de nucleótidos de al menos 15 nucleótidos contiguos que es complementaria a al menos un segmento génico funcional que codifica la región V del receptor inmunitario adaptativo funcional y en el que la pluralidad de cebadores del segmento V hibrida específicamente con prácticamente todos los segmentos génicos que codifican la región V del receptor inmunitario adaptativo que están presentes en la composición del molde, y (2) en cantidades sustancialmente equimolares, una pluralidad de cebadores oligonucleotídicos del segmento J que son independientemente capaces de hibridar específicamente con al menos un polinucleótido que codifica un polipéptido de la región J del receptor inmune adaptativo o con el complemento del mismo, en donde cada cebador del segmento J comprende una secuencia de nucleótidos de al menos 15 nucleótidos contiguos que es complementaria a al menos un segmento génico funcional que codifica la región J del receptor inmune adaptativo funcional y en donde la pluralidad de cebadores del segmento J hibrida específicamente con prácticamente todos los segmentos génicos que codifican la región J del receptor inmunitario adaptativo que están presentes en la composición del molde.

Los cebadores oligonucleotídicos del segmento V y del segmento J son capaces de promover la amplificación en dicha reacción en cadena de polimerasa (PCR) multiplexada de sustancialmente todos los oligonucleótidos del molde en la composición del molde para producir una multiplicidad de moléculas de ADN del molde amplificadas, siendo dicha multiplicidad de moléculas de ADN del molde amplificadas suficiente para cuantificar la diversidad de los oligonucleótidos del molde en la composición del molde, y en donde cada molécula de ADN del molde amplificada en la multiplicidad de moléculas de ADN del molde amplificada es menor de 1000, 900, 800, 700, 600, 500, 400, 300, 200, 100, 90, 80 o 70 nucleótidos de longitud.

El método también incluye las etapas de: (b) secuenciar todas o una porción suficiente de cada una de dicha multiplicidad de moléculas de ADN del molde amplificadas para determinar, para cada molécula de ADN del molde única en dicha multiplicidad de moléculas de ADN del molde amplificada, (i) una secuencia de ADN de oligonucleótido específica del molde y (ii) una frecuencia relativa de aparición del oligonucleótido del molde; y (c) comparar la frecuencia relativa de ocurrencia para cada secuencia de ADN del molde única de dicha composición del molde, en donde una frecuencia de aparición no uniforme para una o más secuencias de ADN del molde indica potencial de amplificación de ácido nucleico no uniforme entre los miembros del conjunto de cebadores de amplificación de oligonucleótidos. Las cantidades para cada conjunto de cebadores del segmento V y del segmento J utilizados en ensayos de amplificación posteriores se pueden ajustar para reducir el sesgo de amplificación en los conjuntos de cebadores en función de la frecuencia relativa de aparición de cada secuencia de ADN del molde única en la composición del molde.

Se proporciona una descripción adicional sobre los métodos de control de sesgo en el documento WO2013/169957, presentado el 8 de mayo de 2013.

Secuenciación

La secuenciación puede realizarse usando cualquiera de una variedad de máquinas y sistemas de secuenciación de moléculas individuales de alto rendimiento disponibles. Los sistemas de secuenciación ilustrativos incluyen sistemas de secuencia por síntesis, como el analizador de genoma Illumina y los instrumentos asociados (Illumina, Inc., San Diego, CA), el sistema de análisis genético Helicos (Helicos BioSciences Corp., Cambridge, MA), Pacific Biosciences PacBio *RS* (Pacific Biosciences, Menlo Park, CA) u otros sistemas con capacidades similares. La secuenciación se logra utilizando un conjunto de oligonucleótidos de secuenciación que hibridan con una región definida dentro de las moléculas de ADN amplificadas. Los oligonucleótidos de secuenciación están diseñados de tal manera que los segmentos génicos codificantes V y J pueden identificarse de manera única por las secuencias que se generan, basándose en la presente divulgación y en vista de las secuencias génicas del receptor inmunitario adaptativo conocidas que aparecen en bases de datos disponibles públicamente. Se describen ejemplos de oligonucleótidos de secuenciación en Robins et al., 2009 Blood 114, 4099; Robins et al., 2010 Sci. Translat. Med. 2:47ra64; Robins et

al., 2011 J. Immunol. Meth. doi:10.1016/j.jim.2011.09. 001; Sherwood et al. 2011 Sci. Translat. Med. 3:90ra61; US2012/0058902, US2010/0330571, WO/2010/151416, WO/2011/106738, WO2012/027503.

Cualquier técnica para secuenciar ácido nucleico conocida por los expertos en la técnica puede usarse en los métodos de la invención proporcionada. Las técnicas de secuenciación de ADN incluyen reacciones de secuenciación didesoxi clásica (método Sanger) usando terminadores o cebadores marcados y separación por gel en plancha o capilares, secuenciación mediante síntesis usando nucleótidos marcados inversamente terminados, pirosecuenciación, secuenciación 454, hibridación específica de alelos con respecto a una biblioteca de sondas de oligonucleótidos marcadas, secuenciación mediante síntesis usando hibridación específica de alelos con respecto a una biblioteca de clones marcados que está seguida de unión, control en tiempo real de la incorporación de nucleótidos marcados durante una etapa de polimerización, secuenciación de colonias y secuenciación SOLiD. La secuenciación de las moléculas separadas ha demostrado más recientemente mediante reacciones de extensión secuencial o única usando polimerasas o ligasas así como mediante hibridaciones diferenciales únicas o secuenciales con bibliotecas de sondas. Estas reacciones se han llevado a cabo sobre muchas secuencias clonales en paralelo incluyendo demostraciones en aplicaciones comerciales actuales de sobre 100 millones de secuencias en paralelo. Estos enfoques de secuenciación pueden, de este modo, usarse para estudiar el repertorio de receptor de linfocitos T (RLT) y/o receptor de linfocitos B (RLB).

La técnica de secuenciación utilizada en los métodos de la invención puede generar al menos 1000 lecturas por ejecución, al menos 10.000 lecturas por ejecución, al menos 100.000 lecturas por ejecución, al menos 500.000 lecturas por ejecución, o al menos 1.000.000 de lecturas por ejecución. La técnica de secuenciación usada en los métodos 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, 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. La técnica de secuenciación utilizada en los métodos de la invención puede generar al menos 30, 40, 50, 60, 70, 80, 90, 100, 110, 120, 150, 200, 250, 300, 350, 400, 450, 500, 550 o 600 pb por lectura.

Los ejemplos de métodos de secuenciación incluyen, aunque sin limitación, secuenciación de molécula única verdadera (tSMS), secuenciación 454 (Roche), secuenciación SOLiD (Applied Biosystems), secuenciación SOLEXA (Illumina), secuenciación SMRT (Pacific Biosciences), secuenciación de nanoporos, secuenciación de matriz de transitorios de efecto de campo quimiosensible, o secuenciación por microscopio electrónico u otros métodos de secuenciación de alto rendimiento conocidos por los expertos en la materia.

Procesamiento de datos de secuenciación

Como se desvela actualmente, también se proporcionan métodos para analizar las secuencias del conjunto diverso de regiones codificantes de la CDR3 reorganizadas de forma única que se generan usando las composiciones y métodos que se describen en el presente documento. Como se ha descrito anteriormente, el sesgo de amplificación se puede corregir utilizando moldes sintéticos de control de sesgo.

También se pueden hacer correcciones para errores de PCR y para estimar la distribución verdadera de clonotipos específicos (por ejemplo, un TCR o IG que tiene una secuencia CDR3 reorganizada de forma única) en sangre o en una muestra derivada de otro tejido periférico o fluido corporal.

En algunas realizaciones, las lecturas secuenciadas se filtran para aquellas que incluyen secuencias CDR3. El procesamiento de datos del secuenciador implica una serie de etapas para eliminar errores en la secuencia primaria de cada lectura y para comprimir los datos. Un filtro de complejidad elimina aproximadamente el 20 % de las secuencias que son lecturas erróneas del secuenciador. En algunas realizaciones, se requiere que las secuencias tengan un mínimo de una coincidencia de seis bases con una de las regiones J de TCR o IG y una de las regiones V de TCR o IG. Aplicando el filtro al carril de control que contiene la secuencia de fagos, en promedio, solo una secuencia en 7-8 millones pasó estas etapas. Por último, se utiliza un algoritmo del vecino más cercano para contraer los datos en secuencias únicas fusionando secuencias estrechamente relacionadas, para eliminar tanto el error de PCR como el error de secuenciación.

Analizando los datos, la relación de secuencias en el producto de PCR se obtiene trabajando hacia atrás desde los datos de secuencia antes de estimar la distribución verdadera de los clonotipos (por ejemplo, secuencias clonales únicas) en la sangre. Para cada secuencia observada un número dado de veces en los datos del presente documento, se estima la probabilidad de que se muestree esa secuencia de un grupo de PCR de tamaño particular. Debido a que las regiones CDR3 secuenciadas se muestrean aleatoriamente a partir de un grupo masivo de productos de PCR, el número de observaciones para cada secuencia se extrae de las distribuciones de Poisson. Los parámetros de Poisson se cuantifican de acuerdo con el número de genomas de linfocitos T que proporcionaron el molde para la PCR. Un modelo simple de mezcla de Poisson estima estos parámetros y coloca una probabilidad por pares para cada secuencia que se extrae de cada distribución. Este es un método de maximización de expectativas,

que reconstruye la abundancia de cada secuencia que se extrajo de la sangre.

En algunas realizaciones, para estimar el número total de secuencias únicas de CDR3 del receptor inmunitario adaptativo que están presentes en una muestra, se puede emplear un enfoque computacional que emplee la fórmula de "especies invisibles" (Efron y Thisted, 1976 Biometrika 63, 435-447). Este enfoque estima el número de especies únicas (por ejemplo, secuencias únicas de receptores inmunes adaptativos) en una gran población de complejos (por ejemplo, una población de células inmunitarias adaptativas como los linfocitos T o los linfocitos B), basado en el número de especies únicas observadas en una muestra finita y aleatoria de una población (Fisher et al., 1943 J. Anim. Ecol. 12:42-58; Ionita-Laza et al., 2009 Proc. Nat. Acad. Sci. EE.UU. 106:5008). El método emplea una expresión que predice el número de especies "nuevas" que se observarían si se analizase una segunda muestra finita y aleatoria del mismo tamaño y de la misma población. Las especies "invisibles" se refieren al número de nuevas secuencias de receptores inmunes adaptativos que se detectarían si las etapas de amplificar las secuencias codificantes de receptores inmunes adaptativos en una muestra y determinar la frecuencia de aparición de cada secuencia única en la muestra se repitieran un número infinito de veces. A modo de teoría no limitante, se supone operativamente a los fines de estas estimaciones que las células inmunes adaptativas (por ejemplo, linfocitos T, linfocitos B) circulan libremente en el compartimento anatómico del sujeto que es la fuente de la muestra a partir de la cual se estima la diversidad (por ejemplo, sangre, linfa, etc.).

Para aplicar esta fórmula, los clonotipos de receptores inmunes adaptativos únicos (por ejemplo, TCRβ, TCRα, TCRγ, TCRδ, IgH) reemplazan a las especies. La solución matemática proporciona que para S, el número total de receptores inmunes adaptativos que tienen secuencias únicas (por ejemplo, "especies" o clonotipos de TCRβ, TCRγ, IgH, que en ciertas realizaciones pueden ser secuencias CDR3 únicas), un experimento de secuenciación observa x_s copias de la secuencia s. Para todos los clonotipos no observados, x_s es igual *0,* y cada clonotipo de TCR o Ig se "captura" en el trascurso de la obtención de una muestra aleatoria (por ejemplo, una extracción de sangre) según un proceso de Poisson con el parámetro λ_s. El número de genomas de linfocitos T o B secuenciados en la primera medición se define como *1,* y el número de genomas de linfocitos T o B secuenciados en la segunda medición se define como *t.*

Debido a que hay una gran cantidad de secuencias únicas, se usa una integral en lugar de una suma. Si *G (λ)* es la función de distribución empírica de los parámetros $\lambda_1,\ldots, \lambda_S$, y n_x es la cantidad de clonotipos (por ejemplo, secuencias únicas de TCR o Ig, o secuencias únicas de CDR3) observadas exactamente x veces, entonces el número total de clonotipos, es decir, la medición de la diversidad E, viene dada por la siguiente fórmula (I):

$$E(n_x) = S\int_0^\infty \left(\frac{e^{-\lambda}\lambda^x}{x!} \right) dG(\lambda) \quad . \qquad \text{(I)}$$

Por consiguiente, la fórmula (I) se puede utilizar para estimar la diversidad total de especies en toda la fuente de la que se toman las muestras de tamaño idéntico. Sin desear quedar ligados a teoría alguna, el principio es que el número de clonotipos muestreados en una muestra de cualquier tamaño dado contiene información suficiente para estimar la distribución subyacente de clonotipos en toda la fuente. El valor para Δ(t), el número de *nuevos* clonotipos observados en una segunda medición, se puede determinar, preferiblemente usando la siguiente ecuación (II):

$$\Delta(t) = \sum_x E(n_x)_{msmt1+msmt2} - \sum_x E(n_x)_{msmt1} = S\int_0^\infty e^{-\lambda}\left(1-e^{-\lambda t}\right) dG(\lambda)$$

(II)

en la cual *msmt1* y *msmt2* son el número de clonotipos de las medidas 1 y 2, respectivamente. Expansión de Taylor de $1\text{-}e^{-\lambda t}$ y la sustitución en la expresión de Δ(t) produce:

$$\Delta(t) = E(x_1)t - E(x_2)t^2 + E(x_3)t^3 - \ldots, \qquad \text{(III)}$$

que se puede aproximar reemplazando las expectativas $(E(n_X))$ con las secuencias de números reales observadas exactamente x veces en la primera medición de la muestra. La expresión para *Δ(t)* oscila ampliamente, ya que *t* tiende al infinito, entonces *Δ(t)* se regulariza para producir un límite inferior para Δ(∞), por ejemplo, utilizando la transformación de Euler (Efron et al., 1976 Biometrika 63:435).

En un ejemplo, utilizando los números observados en una primera medición de la diversidad de secuencia de TCRβ en una muestra de sangre, esta fórmula (II) predijo que se deben observar 1,6 * 10^5 nuevas secuencias únicas en una segunda medición. El valor real de la segunda medición fue de 1,8 * 10^5 nuevas secuencias de TCRβ, que sugería según la teoría no limitativa que la predicción proporcionaba un límite inferior válido en la diversidad de secuencia total de TCRβ en el sujeto del que se extrajo la muestra.

Una descripción adicional sobre el modelo de especies no vistas y los datos de la secuencia de procesamiento se describen en Robins et al., 2009 Blood 114, 4099; Robins et al., 2010 Sci. Translat. Med. 2:47ra64; Robins et al., 2011 J. Immunol. Meth. doi:10.1016/j.jim.2011.09. 001; Sherwood et al. 2011 Sci. Translat. Med. 3:90ra61; US2012/0058902, US2010/0330571, WO/2010/151416, WO/2011/106738, WO2012/027503.

3. Emparejamiento de alto rendimiento de secuencias de ácido nucleico reorganizadas que codifican polipéptidos de heterodímero de receptor inmunitario adaptativo

En ciertas realizaciones, los métodos de la presente invención incluyen la etapa de determinar a partir de la población combinada de células, una pluralidad de pares afines de secuencias primera y segunda de ácido nucleico reorganizadas que codifican los polipéptidos primero y segundo de los heterodímeros del receptor inmunitario adaptativo. La presente invención no pretende limitarse a ningún método de emparejamiento y contempla que muchos métodos conocidos en la técnica, incluidos los desvelados en el presente documento, pueden ser adecuados para practicar la invención reivindicada.

En una realización preferida, los métodos para determinar pares de TCR y/o heterodímeros de Ig son los descritos en el documento WO2014/145992, presentado el 17 de marzo de 2014. Otros métodos para emparejar cadenas de polipéptidos de TCR y/o heterodímeros de Ig se describen en el documento WO2012/188831, presentado el 14 de junio de 2013. A modo de ilustración, pero sin limitación, una realización ejemplar de los métodos de la invención se resume en el presente documento a continuación.

El método de la invención se basa en la observación de que las secuencias de nucleótidos primera y segunda reorganizadas son casi únicas para cada población clonal de células inmunes adaptativas. Las secuencias primera y segunda distintivas surgen a través de la recombinación de segmentos génicos y la deleción o inserción de nucleótidos independientes del molde en las uniones V-J, VD y D-J en células somáticas durante el desarrollo de los linfocitos. Esta diversidad extraordinaria significa que los ARNm que codifican las cadenas de polipéptidos heterodiméricos de un clon de células inmunes adaptativas específicas generalmente estarán presentes solo en conjuntos de células que incluyen ese clon. Esta diversidad extrema se puede aprovechar dividiendo una muestra de células inmunes adaptativas en múltiples subconjuntos y luego secuenciando las moléculas de ARNm primera y segunda para determinar la presencia o ausencia de cada cadena de polipéptidos en cada subconjunto. La primera y la segunda secuencia de un clon deben verse en los mismos subconjuntos de células inmunes adaptativas, y solo en esos subconjuntos.

En algunas realizaciones, el método puede implicar extraer ADN genómico, en lugar de ARNm de las células en una muestra, para amplificar las cadenas de polipéptidos de un heterodímero de receptor inmunitario adaptativo específico.

El emparejamiento de las cadenas de polipéptidos heterodiméricos se convierte en un problema estadístico: para declarar un emparejamiento único, uno debe demostrar que es altamente improbable que un clon determinado ocupe la misma colección de subconjuntos de células inmunes adaptativas que otro clon. La probabilidad de que un clon determinado ocupe la misma colección de subconjuntos de células inmunes adaptativas que otro clon es cercana a cero para miles de clones en un experimento que utiliza los métodos de la invención.

En otras realizaciones, el método de la invención puede ajustarse para emparejar cadenas de receptores inmunes adaptativos afines en cualquier intervalo de frecuencia deseado simplemente cambiando el número de células inmunes adaptativas de entrada por pocillo. Otras realizaciones también pueden analizar pares afines de múltiples bandas de frecuencia en un solo experimento estratificando el número de células inmunes adaptativas de entrada en subconjuntos.

Como se ha descrito anteriormente, el método se puede utilizar para emparejar con precisión secuencias de TCR o IG a alto rendimiento. Por ejemplo, los métodos de la invención pueden usarse para emparejar una primera cadena de polipéptidos de un heterodímero de receptor inmunitario adaptativo que comprende una cadena alfa de TCR (TCRA) y un segundo polipéptido del heterodímero de receptor inmunitario adaptativo que comprende una cadena beta de TCR (TCRB). Además, los métodos de la invención pueden usarse para emparejar un primer polipéptido del heterodímero del receptor inmune adaptativo que comprende una cadena gamma TCR (TCRG) y un segundo polipéptido del heterodímero del receptor inmune adaptativo que comprende una cadena delta TCR (TCRD). En otro ejemplo, los métodos de la invención pueden usarse para emparejar un primer polipéptido de un heterodímero de receptor inmunitario adaptativo que comprende una cadena pesada de inmunoglobulina (IGH) y un segundo polipéptido del heterodímero de receptor inmunitario adaptativo que se selecciona de una cadena ligera de inmunoglobulina IGL o IGK.

El método proporciona etapas para identificar una pluralidad de pares afines que comprenden un primer polipéptido y un segundo polipéptido que forman un heterodímero de receptor inmunitario adaptativo, dicho heterodímero de receptor inmunitario adaptativo comprende un receptor de linfocitos T (TCR) o inmunoglobulina (IG) de un solo clon en una muestra, comprendiendo la muestra una pluralidad de células linfoides de un sujeto mamífero. Como se ha descrito anteriormente, el método incluye etapas para distribuir una pluralidad de células linfoides entre una pluralidad de envases, comprendiendo cada envase una pluralidad de células linfoides; generar una biblioteca de amplicones en la pluralidad de envases mediante la realización de la PCR multiplexada de moléculas de ADNc que se han transcrito inversamente a partir de moléculas de ARNm obtenidas de la pluralidad de células linfoides. La biblioteca de amplicones incluye: i) una pluralidad de primeros amplicones de receptores inmunes adaptativos que codifican el primer polipéptido, comprendiendo, cada uno, una secuencia codificante de región variable (V) única, una secuencia de codificación de región J única o una secuencia de codificación de región J única y una secuencia de codificación de región C única, al menos una secuencia de código de barras, al menos una secuencia de adaptador universal y una secuencia de marcaje de plataforma de secuenciación, y ii) una pluralidad de segundos amplicones de receptores inmunes adaptativos que codifican el segundo polipéptido, comprendiendo, cada uno, una secuencia de codificación de región V única, una secuencia de codificación de región J única o una secuencia de codificación de región J única y una secuencia de codificación de región C única, al menos una secuencia de código de barras, al menos una secuencia de adaptador universal y una secuencia de marcaje de plataforma de secuenciación. El método también incluye etapas para realizar la secuenciación de alto rendimiento de la biblioteca de amplicones para obtener un conjunto de datos de una pluralidad de secuencias primera y segunda de amplicón de receptor inmunitario adaptativo.

Además, el método incluye determinar un patrón de ocupación del envase para cada secuencia única de amplicón del receptor inmune del primer adaptador mediante la asignación de cada secuencia única de amplicón del receptor inmune del primer adaptador a uno o más envases, y un patrón de ocupación del envase para cada segunda secuencia de amplicón del receptor inmune del adaptador único asignando cada segundo adaptador de la secuencia del amplicón del receptor inmune a uno o más envases, en donde cada secuencia de código de barras en las secuencias primera o segunda únicas de amplicón del receptor inmune del adaptador está asociada con un envase particular.

Para cada posible emparejamiento de una primera y segunda secuencia única de amplicón del receptor inmunitario adaptativo para formar un supuesto par afín, el método implica calcular una probabilidad estadística de observar los patrones de ocupación del envase, u observar cualquier proporción mayor de envases compartidos de lo esperado por casualidad, dado que las secuencias primera y segunda de amplicón del receptor inmunitario adaptativo no se originan a partir de la misma población clonal de células linfoides, e identificar una pluralidad de supuestos pares afines basados en la probabilidad estadística que tiene una puntuación inferior a un límite de probabilidad predeterminado.

Después, para cada supuesto par afín identificado, se puede determinar una estimación de la tasa de falso descubrimiento para un posible emparejamiento falso de la primera secuencia única de amplicón del receptor inmune adaptativo y la segunda secuencia única de amplicón del receptor inmune adaptativo. El método incluye etapas para identificar una pluralidad de pares afines de secuencias primera y segunda únicas de receptores inmunes adaptativos como pares afines verdaderos que codifican dichos receptores inmunes adaptativos en dicha muestra basándose en dicha probabilidad estadística y dicha estimación de tasa de falso descubrimiento.

En algunas realizaciones, la puntuación estadística puede ser un p-valor calculado para emparejar cada supuesto par afín de secuencias primera y segunda únicas de amplicón de receptor inmunitario adaptativo. En una realización, calcular la puntuación estadística comprende calcular la probabilidad de que la primera y segunda secuencias únicas de amplicón del receptor inmunitario adaptativo ocupen conjuntamente tantos o más envases de los que se observa que ocupan conjuntamente, suponiendo que no existe un emparejamiento afín verdadero y dado el número de envases ocupados por dicha primera secuencia única de amplicón del receptor inmune adaptativo y el número de envases ocupados por la segunda secuencia única de amplicón del receptor inmune adaptativo.

Esencialmente, dado cualquiera de las dos secuencias de receptores inmunes adaptativos, el método analiza si las dos secuencias se producen conjuntamente en más envases de los que cabría esperar por casualidad. Dado un total de N envases, una primera secuencia de receptor inmunitario adaptativo (A) observada en un total de X envases, una segunda secuencia de receptor inmunitario adaptativo (B) observada en un total de Y envases, y Z envases en los que se observan ambas secuencias de receptor inmunitario adaptativo (A) y (B), el método establece que la secuencia dada (A) se encuentra en X de N envases (X I N) y la secuencia (B) se encuentra en Y de N (Y I N) envases, un cálculo de la probabilidad de que ambas secuencias se encuentren en Z o más envases.

En algunas realizaciones, cuanto menor sea la probabilidad de que el número observado de envases superpuestos entre las secuencias A y B pueda ocurrir por casualidad, más probable es que su coincidencia no sea casual, pero en cambio se debe a un verdadero emparejamiento afín.

A continuación, la identificación de una pluralidad de supuestos pares afines que tienen una alta probabilidad de

emparejamiento basado en la probabilidad estadística puede comprender para cada secuencia primera única de amplicón del receptor inmunitario adaptativo que identifica la secuencia segunda única de amplicón del receptor inmunitario adaptativo que tiene la puntuación de p-valor más baja de coincidencia, o para cada secuencia segunda única de amplicón del receptor inmunitario adaptativo que encuentra la secuencia primera única de amplicón del receptor inmunitario adaptativo que tiene la puntuación de p-valor más baja de coincidencia.

En otras realizaciones, determinar una estimación de tasa de falsos descubrimientos comprende: calcular los p-valores para cada una de la pluralidad de supuestos pares afines identificados en la muestra; comparar los p-valores para toda la pluralidad de supuestos pares afines con una distribución esperada del p-valor, dicha distribución esperada del p-valor calculada para representar un experimento en el que no hay verdaderos pares afines presentes; y determinar para cada supuesto par afín, una proporción esperada de resultados falsos positivos, de modo que todos los p-valores en o por debajo del p-valor del supuesto par afín se determinan para representar un emparejamiento verdadero afín.

En ciertas realizaciones, el cálculo de la distribución esperada del p-valor comprende: permutar los envases en los que se ha observado cada primera y segunda secuencia del receptor inmunitario adaptativo en un experimento idéntico sin pares verdaderos afines, y calcular la distribución de los p-valores asociados con cada supuesto par afín.

El método incluye identificar una pluralidad de pares afines de secuencias primera y segunda únicas de receptores inmunes adaptativos como pares afines verdaderos seleccionando una pluralidad de supuestos pares afines que tienen p-valores por debajo de un umbral calculado basándose en la estimación de la tasa de falso descubrimiento.

En una realización, el par identificado de secuencias primera y segunda de amplicón de receptor inmunitario adaptativo tiene una estimación de tasa de falsos descubrimientos de menos del 1 %. En otras realizaciones, el par identificado de secuencias primera y segunda de amplicón del receptor inmunitario adaptativo tiene una estimación de tasa de falsos descubrimientos de menos del 2 %, 3 %, 4 %, 5 %, 6 %, 7 %, 8 %, 9 % o 10 %.

El método también puede incluir poner en contacto cada una de dicha pluralidad de envases, bajo condiciones y durante un tiempo suficiente para promover la transcripción inversa de moléculas de ARNm obtenidas a partir de dicha pluralidad de células linfoides, con un primer conjunto de cebadores de transcripción inversa. En ciertas realizaciones, el (A) primer conjunto de cebadores de transcripción inversa de oligonucleótidos comprende cebadores capaces de transcribir inversamente una pluralidad de secuencias de ARNm que codifican la pluralidad de primer y segundo polipéptidos del receptor inmunitario adaptativo para generar una pluralidad de primer y segundo amplicones de ADNc del receptor inmunitario adaptativo transcrito de forma inversa, en donde la pluralidad de los primeros amplicones de ADNc del receptor inmunitario adaptativo transcrito de forma inversa que codifican el primer polipéptido del receptor inmunitario adaptativo comprenden 1) una secuencia génica codificante de la región V única, y 2) una secuencia génica codificante de la región J única o tanto una secuencia génica codificante de la región J única y una secuencia génica codificante de la región C, y en donde la pluralidad de segundos amplicones de ADNc de receptor inmunitario adaptativo transcrito de forma inversa que codifican el segundo polipéptido del receptor inmunitario adaptativo comprende 1) una secuencia génica codificante de la región V única, y 2) una secuencia génica codificante de la región J única o tanto una secuencia génica codificante de una región J única como una secuencia génica codificante de una región C única.

El primer y segundo amplicón de ADNc del receptor inmunitario adaptativo transcrito de forma inversa se amplifican en una segunda reacción. La reacción comienza poniendo en contacto cada uno de dicha pluralidad de envases, bajo condiciones y durante un tiempo suficiente para promover una amplificación por PCR multiplexada del primer y segundo amplicones de ADNc de receptor inmunitario adaptativo transcrito de forma inversa con un segundo (B) y un tercer (C) conjuntos de cebadores de oligonucleótidos. En algunos aspectos, el (B) segundo conjunto de cebadores oligonucleotídicos comprende cebadores directos e inversos capaces de amplificar la pluralidad de primeros amplicones de ADNc del receptor inmunitario adaptativo transcrito de forma inversa, en donde dichos cebadores directo e inverso son capaces de hibridar con los primeros amplicones de ADNc de receptor inmunitario adaptativo transcrito de forma inversa.

Cada par de cebadores directos e inversos en el segundo conjunto de cebadores de oligonucleótidos es capaz de amplificar los primeros amplicones de ADNc del receptor inmunitario adaptativo transcrito de forma inversa. Los cebadores directos en el segundo conjunto de cebadores oligonucleotídicos comprenden una primera secuencia adaptadora universal y una región complementaria a la secuencia génica que codifica la región V. Los cebadores inversos en el segundo conjunto de cebadores oligonucleotídicos comprenden una segunda secuencia adaptadora universal y una región complementaria a la secuencia génica que codifica la región J o la secuencia del gen que codifica la región C.

El tercer conjunto de cebadores de oligonucleótidos (C) comprende cebadores directos e inversos capaces de amplificar la pluralidad de los segundos amplicones de ADNc del receptor inmunitario adaptativo transcrito de forma inversa. Cada par de cebadores directo e inverso en el tercer conjunto de cebadores de oligonucleótidos es capaz de amplificar los segundos amplicones de ADNc del receptor inmunitario adaptativo transcrito de forma inversa. En un aspecto, los cebadores directos en el tercer conjunto de cebadores oligonucleotídicos comprenden una primera

secuencia adaptadora universal y una región complementaria a la secuencia génica que codifica la región V. Los cebadores inversos en el tercer conjunto de cebadores oligonucleotídicos comprenden una segunda secuencia adaptadora universal y una región complementaria a la secuencia del gen que codifica la región J o complementaria a la secuencia del gen que codifica la región C.

El método también incluye generar i) una pluralidad de terceros amplicones de receptores inmunes adaptativos, cada uno de los cuales comprende una secuencia génica que codifica una región V única, o complemento de la misma, una secuencia génica única que codifica la región J o una secuencia génica única que codifica la región J y una secuencia génica única que codifica la región C, o complemento de la misma, y la primera y segunda secuencia adaptadora universal, y ii) una pluralidad de cuartos amplicones de receptores inmunes adaptativos que comprende, cada uno, una secuencia génica que codifica una región V única, o un complemento de la misma, una secuencia génica única que codifica la región J o una secuencia génica única que codifica la región J y una secuencia génica única que codifica la región C, o complemento de la misma, y la primera y segunda secuencias del adaptador universal.

La pluralidad de terceros amplicones de receptores inmunes adaptativos y la pluralidad de cuartos amplicones de receptores inmunes adaptativos se amplifican con cebadores adicionales. El método incluye poner en contacto cada uno de la pluralidad de envases, bajo condiciones y durante un tiempo suficiente para promover una segunda amplificación por PCR multiplexada de la pluralidad del tercer y cuarto amplicones de los receptores inmunitarios adaptativos con un cuarto conjunto de cebador oligonucleotídico (D) y un quinto conjunto de cebador oligonucleotídico (E).

En una realización, el (D) cuarto conjunto de cebadores oligonucleotídicos comprende cebadores directos e inversos capaces de amplificar la pluralidad de amplicones del tercer receptor inmunitario adaptativo, en donde los cebadores directo e inverso son capaces de hibridar con los terceros amplicones del receptor inmunitario adaptativo. Cada par de cebadores directos e inversos en el cuarto conjunto de cebadores de oligonucleótidos es capaz de amplificar dichos terceros amplicones del receptor inmunitario adaptativo.

El cebador directo en el cuarto conjunto de cebadores oligonucleotídicos comprende una secuencia de marcaje de plataforma de secuenciación y una región complementaria a la primera secuencia del adaptador universal en la pluralidad de un tercer amplicón del receptor inmunitario adaptativo y el cebador inverso comprende una secuencia de marcaje de plataforma de secuenciación y una región complementaria a la segunda secuencia de adaptador universal en la pluralidad de terceros amplicones de receptores inmunes adaptativos. En otra realización, uno o ambos cebadores directo e inverso en el cuarto conjunto de cebadores de oligonucleótidos comprende una secuencia de código de barras única asociada con el envase en el que se introduce el cuarto conjunto de cebadores de oligonucleótidos.

El (E) quinto conjunto de cebadores de oligonucleótidos comprende cebadores directos e inversos capaces de amplificar la pluralidad de cuartos amplicones de receptores inmunes adaptativos, en donde los cebadores directo e inverso son capaces de hibridar con los cuartos amplicones del receptor inmunitario adaptativo. Cada par de cebadores directo e inverso en dicho cuarto conjunto de cebadores oligonucleotídicos es capaz de amplificar dicha pluralidad del cuarto amplicón del receptor inmunitario adaptativo. El cebador directo en el quinto conjunto de cebadores de oligonucleótidos comprende una secuencia de marcaje de plataforma de secuenciación y una región complementaria a la primera secuencia del adaptador universal en la pluralidad de cuartos amplicones de receptores inmunes adaptativos, y el cebador inverso en el quinto conjunto de cebadores de oligonucleótidos comprende una secuencia de marcaje de plataforma de secuenciación y una región complementaria a la segunda secuencia del adaptador universal en la pluralidad de cuartos amplicones de receptores inmunitarios adaptativos.

Uno o ambos cebadores directo e inverso del cuarto conjunto de cebadores de oligonucleótidos comprende una secuencia de código de barras única asociada con el envase en el que se introduce el cuarto conjunto de cebadores de oligonucleótidos, generando así la biblioteca de amplicones que comprende la pluralidad de primeros amplicones de receptores inmunitarios adaptativos y la pluralidad de segundos amplicones de receptores inmunitarios adaptativos.

A continuación, el método incluye combinar la biblioteca de amplicones de la pluralidad de envases en una mezcla para la secuenciación. Los métodos para la secuenciación de alto rendimiento se describen en detalle anteriormente y en los documentos US2012/0058902, US2010/0330571, WO2011/106738 o WO2012/027503.

En un aspecto, la pluralidad de los primeros amplicones del receptor inmunitario adaptativo comprende una secuencia codificante de la región C. En algunos aspectos, la pluralidad de los segundos amplicones del receptor inmunitario adaptativo comprenden una secuencia codificante de la región C.

En algunos casos, la muestra comprende una muestra de sangre. En otra realización, la muestra comprende una muestra de tejido. En ciertas realizaciones, la muestra comprende una muestra de células linfoides humanas purificadas o cultivadas. En otras realizaciones, el envase comprende al menos 104 células linfoides. En otra realización, la muestra comprende al menos 104 células.

El método es aplicable a varios loci de receptores inmunitarios adaptativos, como se ha descrito anteriormente, tal como el emparejamiento de una cadena TCR alfa (TCRA) y una cadena TCR beta (TCRB), una cadena TCR gamma (TCRG) y una cadena TCR delta (TCRD), o una cadena pesada de inmunoglobulina (IGH) y una cadena ligera de inmunoglobulina IGL o una cadena IGK.

Cuando el primer polipéptido del heterodímero del receptor inmunitario adaptativo es una cadena IGH y el segundo polipéptido del heterodímero del receptor inmunitario adaptativo es tanto IGL como IGK, entonces se usan tres conjuntos de cebadores de amplificación diferentes que comprenden: un primer cebador de amplificación de oligonucleótidos establecido para IGH, un segundo conjunto de cebadores de amplificación de oligonucleótidos establecido para IGK, y un tercer conjunto de cebadores de amplificación de oligonucleótidos establecido para IGL.

Por lo tanto, los métodos y las composiciones de la invención pueden resultar útiles en muchas aplicaciones en inmunología, medicina y desarrollo terapéutico. Los métodos de la invención ofrecen oportunidades para investigar conexiones entre las secuencias primarias de una colección de receptores inmunes seleccionados y la(s) diana(s) (y epítopos) que causaron su selección. Con atención al diseño experimental y al control de variables (por ejemplo, tipo HLA), los métodos de la invención pueden ser un enfoque útil para identificar TCR críticos de linfocitos infiltrantes de tumores, para establecer nuevos criterios de respuesta a la vacunación de rutina o experimental, y para el análisis epidemiológico de exposiciones públicas y respuestas compartidas. Los métodos de la invención también proporcionan información sobre la contribución relativa de cada cadena independiente a una respuesta dada. Además, el enfoque de los presentes inventores proporciona datos sobre si puede haber atributos físicos de la cadena TCR que dirijan una respuesta inmune particular. Por ejemplo, restricciones en la longitud o parámetros biofísicos de una o ambas cadenas para un tipo dado de respuesta a un tipo dado de exposición antigénica. Los métodos de la invención se pueden ejecutar con suministros y equipos de laboratorio convencionales, sin la necesidad de experiencia especializada, y el tipo de muestra inicial tiene un amplio rango potencial (muestras tumorales, células clasificadas, células en suspensión, etc.). Esta tecnología está diseñada para hacerse a escala y que sea accesible a una variedad de laboratorios.

Es importante reconocer que los métodos de la invención pueden aplicarse y funcionarán igualmente bien para TCRγ/δ, y para unir las cadenas pesada y ligera de inmunoglobulina (IGH con IGK o IGL). Dado el interés práctico en el desarrollo de anticuerpos monoclonales, así como la importancia general de la respuesta inmunitaria humoral, los métodos de la invención tienen el potencial de convertirse en una tecnología importante para el descubrimiento biomédico.

4. Asignación de pares afines de primera y segunda moléculas de ácido nucleico reorganizadas que codifican heterodímeros del receptor inmunitario adaptativo a una muestra de fuente única

Usando los métodos descritos anteriormente, las primeras secuencias de ácido nucleico reorganizadas se determinan en cada una de las muestras de la fuente durante la secuenciación de un solo locus como se describe en la Sección 2. Además, las primeras secuencias de ácido nucleico reorganizadas se determinan en la muestra combinada durante un experimento de emparejamiento como se describe anteriormente en la Sección 3. En ciertas realizaciones, los métodos de la invención incluyen comparar las primeras secuencias de ácido nucleico reorganizadas determinadas en cada una de las muestras de la fuente durante la secuenciación de un solo locus con las primeras secuencias de ácido nucleico reorganizadas determinadas en la muestra combinada durante un experimento de emparejamiento para asignar cada primera secuencia de ácido nucleico reorganizada en la población combinada a una sola muestra de la fuente. Posteriormente, cada segunda secuencia de ácido nucleico reordenada afín identificada como emparejada con cada primera secuencia de ácido nucleico reordenada afín a una muestra de fuente única puede asignarse a la misma muestra de fuente única.

Sin desear quedar ligados a teoría alguna, dado que las secuencias de la primera cadena (por ejemplo, TCRB o IgH) son muy únicas, se predice que la mayoría se detectará en una y solo una muestra de fuente biológica. Por lo tanto, cada par afín de secuencias identificadas en la muestra combinada puede asignarse de nuevo a una muestra de fuente única. Al mezclar muchas muestras de la fuente antes de hacer un experimento de emparejamiento, tal como se describe en el presente documento, las secuencias de ácido nucleico que codifican los heterodímeros del receptor inmunitario adaptativo de alta frecuencia pueden detectarse a partir de un solo experimento de emparejamiento. En ciertas realizaciones, los receptores de alta frecuencia están presentes a una frecuencia de aproximadamente 1:100 a aproximadamente 1:1000 en una muestra de fuente única. Al realizar la secuenciación de un solo locus en cada muestra de la fuente en paralelo al experimento de emparejamiento único, se pueden asignar secuencias emparejadas de alta frecuencia a una muestra de fuente única entre una pluralidad de muestras no relacionadas.

Ejemplos

Ejemplo 1: Asignación de secuencias de TRCA y TCRB emparejadas a muestras de origen tumoral

Este ejemplo demuestra la prueba de concepto de una realización de los métodos de la presente invención. A modo

de resumen no limitante, el método de asignar simultáneamente ácidos nucleicos emparejados que codifican heterodímeros del receptor inmunitario incluye las siguientes etapas:

(a) Comenzar con una pluralidad de muestras biológicas, incluyendo cada muestra los linfocitos con TCR emparejado o secuencias de receptor de Ig de interés;
(b) Mezclar volúmenes iguales de solución acuosa incluyendo células de cada muestra biológica para generar una muestra mixta de linfocitos;
(c) Realizar un descubrimiento de alto rendimiento de pares de cadenas de TCR del receptor de antígeno afín o pares de cadenas pesadas/ligeras de Ig para la muestra mixta ("emparejamiento" por ejemplo, tal como se describe en la Sección 3 anterior);
(d) para cada muestra biológica individual, realizar una secuenciación de alto rendimiento de una cadena del par (por ejemplo, cadena pesada sola), sin información de emparejamiento ("secuenciación de un solo locus" por ejemplo, tal como se describe en la Sección 2 anterior), asignando así una secuencia del par (por ejemplo, la secuencia de la cadena pesada) a una muestra biológica individual;
(e) para cada par identificado en (c), examinar la secuencia de la cadena asignada (por ejemplo, la secuencia de la cadena pesada). Cuando las secuencias de cadena asignadas se observan en una y solo una muestra en (d), asignar pares afines identificados en (c) a la muestra biológica original en la que la cadena pesada sola se expresa de manera única.

En el siguiente experimento, se asignaron secuencias de TCRA y TCRB emparejadas a muestras de origen mediante la realización de un solo experimento de emparejamiento en muestras de tumores agrupadas, tal como se resume anteriormente. Brevemente, se obtuvieron 10 muestras de tumor (cinco muestras de tumor de riñón y cinco muestras de tumor de ovario) de diferentes individuos. Los tumores se procesaron de acuerdo con los procedimientos estándar para proporcionar suspensiones celulares disociadas que contienen linfocitos. Se mezclaron volúmenes iguales de las suspensiones celulares acuosas para proporcionar una población combinada de células. El descubrimiento de alto rendimiento de los pares de cadenas pesadas/ligeras del receptor de antígeno afín para la muestra mixta se realizó como se describió anteriormente para identificar secuencias TCRA/TCRB emparejadas que codifican heterodímeros de CDR TCRα:β. En esta parte del experimento, una placa de 96 pozos se dividió en 3 subconjuntos de 32 pozos. A cada subconjunto de pocillos se le asignaron 100, 50 y 10 linfocitos T por tumor, sin embargo, no se asignó el mismo número de linfocitos T a cada pocillo en una placa. Al hacerlo, el objetivo era capturar linfocitos infiltrantes de tumores de diferentes frecuencias.

En paralelo, la secuenciación de un solo locus de ADN genómico se realizó por separado en 25-50 ng de ADN de cada tumor, en cinco réplicas cada una, por separado para TCRA y TCRB, tal como se describe adicionalmente en el presente documento. Las secuencias de TCRA/TCRB emparejadas se mapearon para tumores de origen mediante el uso de coincidencias de cadenas exactas entre las secuencias emparejadas y las secuencias de un solo locus con el requisito de que ambas secuencias en un par se mapeen exclusivamente para una muestra de origen.

La **Figura 1** muestra el número de pares TCRA/B que podrían asignarse nuevamente a la muestra de origen para cada tumor. Notablemente, se leyeron más de 1.500 pares de secuencias TCRA/B una tasa de falso descubrimiento (TFD) del 1 % y cada una de ellas se asignó a una sola muestra tumoral de origen.

Para confirmar experimentalmente la precisión del emparejamiento de secuencias y la predicción de la TFD, las secuencias de TCRA de cada muestra fuente se determinaron mediante secuenciación de un solo locus. Las secuencias de TCRA se usaron para determinar el porcentaje de secuencias TCRA/B afines identificadas en el experimento de emparejamiento que en realidad se asignan a la misma muestra fuente ("pares verdaderos"). De los> 1.500 pares leídos, hubo 731 pares cuyas secuencias de TCRA y TCRB podrían asignarse sin ambigüedad, y se descubrió que la mayoría de estos se mapean en una sola muestra como pares verdaderos. Se determinó que el número restante de pares falsos era seis, casi coincide con la predicción de siete pares falsos proporcionada por el modelo de TFD. Los resultados de este análisis se representan gráficamente en la **Figura 2**. De manera importante, estos resultados demuestran que más de 1.500 pares de secuencias de TCR pueden recuperarse de una mezcla compleja de muestras tumorales y asignarse a una sola fuente con alta confianza mediante la práctica de la estrategia experimental única descrita en el presente documento.

Ejemplo 2: Asignación de secuencias de TRCA y TCRB emparejadas a muestras de origen tumoral

Este ejemplo proporciona validación para un enfoque para multiplexar muestras en una placa de múltiples pocillos, en donde se realizó un experimento de emparejamiento para emparejar los TCR de 18 muestras tumorales en paralelo.

El repertorio de TCRA y TCRB de cada muestra de tumor se secuenciaron primero, y se estimó la frecuencia de cada clon y el número de linfocitos T en cada muestra del repertorio. Esta información se usó después para determinar el número óptimo de células de entrada para obtener de cada tumor para capturar los clones más comunes en un experimento de emparejamiento. Se mezclaron las cantidades indicadas de material de entrada de los 18 tumores, las células se distribuyeron en tres placas de 96 pocillos (réplicas técnicas) y se realizó un ensayo de emparejamiento en cada placa.

Con experimentos multiplexados que incluyen muestras con repertorios de TCRA y TCRB conocidos, es posible medir directamente la tasa de falsos descubrimientos (TFD) contando "pares de muestras cruzadas" - es decir, los pares de TCR se emparejan con una secuencia de TCRA de una muestra y una secuencia de TCRB de otra. La **Figura 3** compara la TFD empírica de pares de muestras cruzadas (eje y) con la TFD del modelo estadístico de los presentes inventores (eje x) para los valores predichos de TFD. Los valores predichos son precisos en umbrales de TFD del 2 % o más bajos y algo conservadores (más bajos que la TFD empírica) en umbrales de TFD del 5 % o más. Estos resultados confirman que el método puede controlar la TFD en una placa multiplexada de múltiples pocillos en un rango de rigurosidades deseadas.

Se pueden lograr altos rendimientos de emparejamiento para todas las muestras de entrada en una reacción multiplexada. La **Figura 4** ilustra los rendimientos de emparejamiento para los 18 tumores multiplexados. Para los 10 más frecuentes, **Figura 4A**, o los 100 más frecuentes, **Figura 4B**, las secuencias productivas de TCRB en el repertorio de cada tumor, los histogramas muestran la cantidad de clones que se emparejaron con alta confianza, TFD ≤ 1 %. Para 12 de los 18 tumores, los diez clones más frecuentes se emparejaron con éxito, y al menos siete de los diez clones principales se emparejaron en 17/18 muestras (**Figura 4A**). De manera similar, al menos 70 de los 100 clones más frecuentes se combinaron en 14/18 muestras (**Figura 4B**), mientras que las muestras restantes tuvieron rendimientos más bajos debido al material de entrada limitado. El clon más frecuente se emparejó con éxito en cada tumor. El número total de pares por muestra osciló entre 32 y 1.204, con una mediana de 404.

REIVINDICACIONES

1. Un método para asignar un par de polipéptidos primero y segundo que forman un receptor de linfocitos T (TCR) o un heterodímero de inmunoglobulina (Ig) a una muestra de fuente única entre una pluralidad de muestras de la fuente, que comprende:

(1) para cada una de una pluralidad de muestras de la fuente cada una de las cuales comprende linfocitos T o linfocitos B, determinando las primeras secuencias de ácido nucleico reorganizadas que codifican los primeros polipéptidos de los heterodímeros de TCR o Ig presentes en la muestra de la fuente y asignar las primeras secuencias de ácido nucleico reorganizadas a la muestra de la fuente;
(2) agrupar la pluralidad de muestras de la fuente para formar una población combinada de células;
(3) determinar a partir de la población combinada de células, una pluralidad de pares afines de secuencias de ácido nucleico reorganizadas primera y segunda que codifican los polipéptidos primero y segundo de los heterodímeros de TCR o Ig;
(4) comparar las primeras secuencias de ácido nucleico reorganizadas determinadas en cada una de las muestras de la fuente en (1) con las primeras secuencias de ácido nucleico reorganizadas determinadas a partir de la pluralidad de pares afines de secuencias de ácido nucleico reorganizadas en (3) para asignar cada primera secuencia de ácido nucleico reordenada presente en la población combinada con una sola muestra de la fuente; y
(5) para cada primera secuencia de ácido nucleico reordenada asignada a una única muestra de la fuente en la etapa (4), asignando la segunda secuencia de ácido nucleico reordenada afín del par afín identificado en la etapa (3) a la misma muestra única de la fuente.

2. El método de la reivindicación 1, en donde la primera secuencia de ácido nucleico reordenada comprende una secuencia de ácido nucleico reordenada de TCRB.

3. El método de la reivindicación 1, en donde la primera secuencia de ácido nucleico reordenada comprende una secuencia de ácido nucleico reordenada de TCRA.

4. El método de la reivindicación 1, en donde la primera secuencia de ácido nucleico reordenada comprende una secuencia de ácido nucleico reordenada de cadena pesada de inmunoglobulina (IGH).

5. El método de la reivindicación 1, en donde la primera secuencia de ácido nucleico reordenada comprende una secuencia de ácido nucleico reordenada de cadena ligera de inmunoglobulina kappa o inmunoglobulina lambda.

6. El método de una cualquiera de las reivindicaciones 1 a 5, en donde la etapa de determinar una primera secuencia de ácido nucleico reordenada que codifica el primer polipéptido del heterodímero de TCR o Ig presente en la muestra fuente incluye las etapas de:

(a) para cada muestra de la fuente, amplificar moléculas de ácido nucleico reorganizadas extraídas de la muestra de la fuente en una única reacción en cadena de polimerasa (PCR) multiplexada usando una pluralidad de cebadores del segmento V y una pluralidad de cebadores del segmento J para producir una pluralidad de amplicones de ácido nucleico reorganizados, y
(b) secuenciar dicha pluralidad de amplicones de ácido nucleico reorganizados para determinar las secuencias de las primeras secuencias de ácido nucleico reorganizadas en cada muestra de la fuente.

7. El método de la reivindicación 6, en donde dicha PCR multiplexada simple produce al menos 10^6 amplicones distintos que representan una diversidad de secuencias del TCR reorganizadas o de CDR3 de la IG presentes en cada una de dichas muestras.

8. El método de la reivindicación 6, en donde cada uno de dichos cebadores del segmento V comprende una primera secuencia y una segunda secuencia, en donde dicha primera secuencia es complementaria con una porción de una primera región de un segmento codificante V del TCR o de la IG, dicha primera región se encuentra inmediatamente en 5' para una segunda región de dicho segmento codificante V donde se producen deleciones sin molde durante la reordenación del gen del TCR o de la IG, en donde dicha segunda región de dicho segmento codificante V es adyacente a y 5' a una secuencia señal de recombinación de V (V-RSS) de dicho segmento codificante V, en donde dicha primera secuencia está ubicada en 3' para dicha segunda secuencia en dicho cebador del segmento V, en donde la segunda secuencia comprende una secuencia de cebador universal, y
en donde cada uno de dichos cebadores del segmento J tiene una primera secuencia y una segunda secuencia, en donde dicha primera secuencia es complementaria con una porción de una primera región de un segmento codificante J del TCR o IG, dicha primera región se localiza inmediatamente en 3' para una segunda región de dicho segmento codificante J donde se producen deleciones sin molde durante la reordenación génica del TCR o de la IG, en donde dicha segunda región de dicho segmento J es adyacente a y 3' para una secuencia señal de recombinación de J (J-RSS) de dicho segmento codificante J, en donde dicha primera secuencia está ubicada en 3' para dicha segunda secuencia en dicho cebador del segmento J, en donde la segunda secuencia comprende una secuencia de cebador universal.

9. El método de una cualquiera de las reivindicaciones 6 a 8, que comprende además realizar una segunda reacción de amplificación hibridando cebadores de colas con regiones dentro de los amplicones de ácido nucleico reorganizados.

10. El método de la reivindicación 9, en donde el cebador de colas comprende una secuencia de cebador universal, una secuencia de código de barras única, una secuencia de oligonucleótidos aleatoria y una secuencia adaptadora.

11. El método de una cualquiera de las reivindicaciones 1 a 10, en donde la etapa de determinar a partir de la población combinada de células una pluralidad de pares afines de secuencias primer y segunda de ácido nucleico reorganizadas que codifican los polipéptidos primero y segundo de los heterodímeros del TCR o de la Ig comprende las etapas de:

(a) distribuir células de la población combinada de células en una pluralidad de envases, comprendiendo cada envase una subpoblación de células;
(b) generar una biblioteca de amplicones para cada una de dichas pluralidades de envases mediante la realización de una única PCR multiplexada de moléculas de ADNc que se han transcrito inversamente a partir de moléculas de ARNm obtenidas de dicha subpoblación de células;
(c) realizar una secuenciación de alto rendimiento de dicha biblioteca de amplicones para obtener un conjunto de datos de una pluralidad de secuencias primera y segunda de amplicón del receptor inmunitario adaptativo para cada una de dichas pluralidades de los envases;
(d) determinar un patrón de ocupación del envase para cada secuencia única de amplicón del receptor inmune del primer adaptador mediante la asignación de cada secuencia única de amplicón del receptor inmune del primer adaptador a uno o más envases, y determinar un patrón de ocupación del envase para cada segunda secuencia de amplicón del receptor inmune del adaptador único asignando cada segundo adaptador de la secuencia del amplicón del receptor inmune a uno o más envases;
(e) para cada posible emparejamiento de una primera y segunda secuencia única de amplicón del receptor inmunitario adaptativo para formar un supuesto par afín, calcular una probabilidad estadística de observar dichos patrones de ocupación de envases; y
(f) identificar una pluralidad de pares de supuestos pares afines basada en dicha probabilidad estadística.

12. El método de la reivindicación 11, en donde la identificación de una pluralidad de supuestos pares afines se basa en que dicha probabilidad estadística tiene una puntuación inferior a un límite de probabilidad predeterminado.

13. El método de la reivindicación 11 o la reivindicación 12, que comprende además:

para cada supuesto par afín identificado, determinar una estimación de la tasa de falso descubrimiento para un posible emparejamiento falso de dicha secuencia única de amplicón del receptor inmune del primer adaptador único y dicha secuencia única de amplicón del receptor inmune del segundo adaptador único; e identificar una pluralidad de pares afines de secuencias únicas de receptores inmunes adaptativos primero y segundo como pares afines verdaderos que codifican dichos receptores inmunes adaptativos en dicha muestra basándose en dicha probabilidad estadística y dicha estimación de la tasa de falso descubrimiento.

14. El método de una cualquiera de las reivindicaciones 11 a 13, en donde:

(i) la pluralidad de las primeras secuencias de amplicón del receptor inmunitario adaptativo comprende una secuencia de codificación de región variable (V) única, una secuencia de codificación de región J única o una secuencia de codificación de región J única y una secuencia de codificación de región C única, al menos una secuencia de código de barras, al menos una secuencia de adaptador universal y una secuencia de marcaje de plataforma de secuenciación, y
(ii) la pluralidad de las segundas secuencias de amplicón del receptor inmunitario adaptativo comprende, cada una, una única secuencia codificante de región V, una secuencia de codificación de región J única o una secuencia de codificación de región J única y una secuencia de codificación de región C única, al menos una secuencia de código de barras, al menos una secuencia de adaptador universal y una secuencia de marcaje de plataforma de secuenciación.

15. El método de una cualquiera de las reivindicaciones 1 a 14, en donde la pluralidad de muestras de la fuente comprende muestras biológicas de diferentes sujetos humanos.

Figura 1

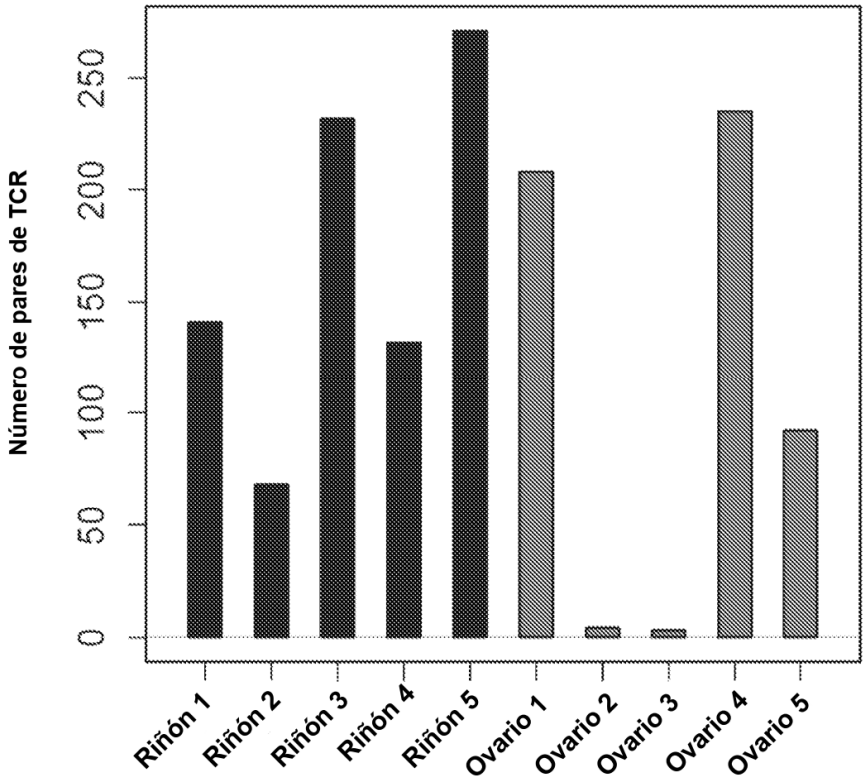
Número de pares de TCR
0
50
100
150
200
250
Riñón 1
Riñón 2
Riñón 3
Riñón 4
Riñón 5
Ovario 1
Ovario 2
Ovario 3
Ovario 4
Ovario 5

Figura 2

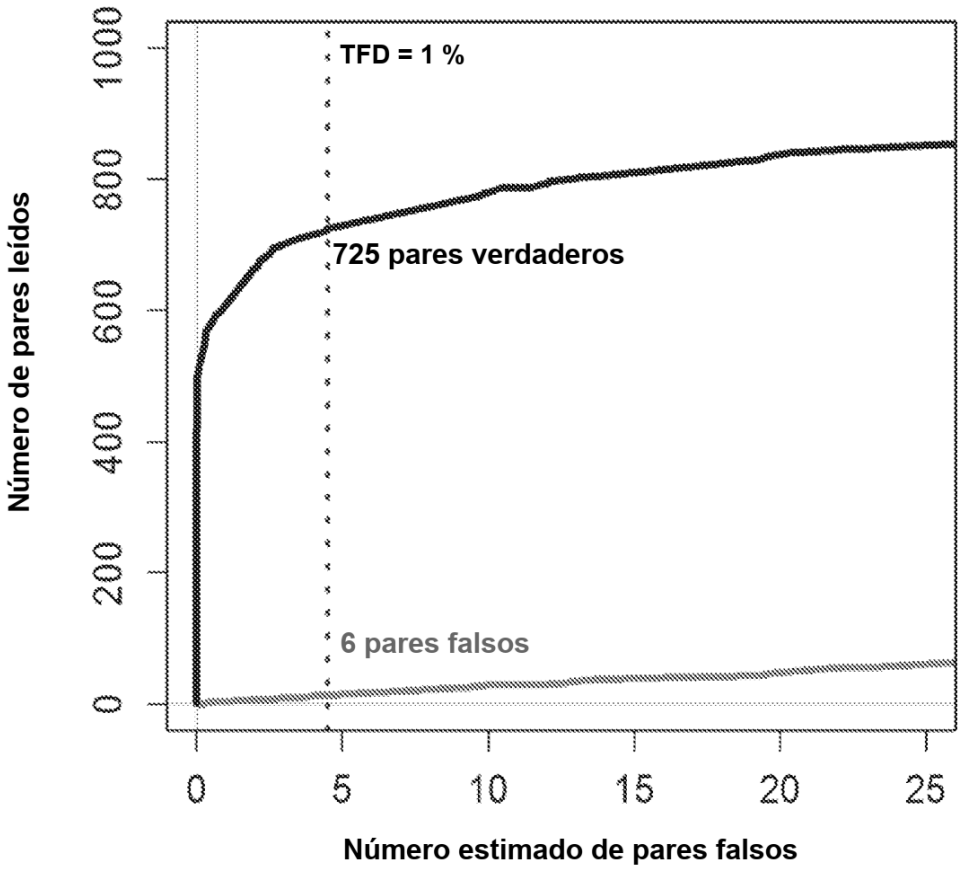
TFD = 1 %
725 pares verdaderos
6 pares falsos
Número de pares leídos
0
200
400
600
800
1000
0
5
10
15
20
25
Número estimado de pares falsos

Figura 3

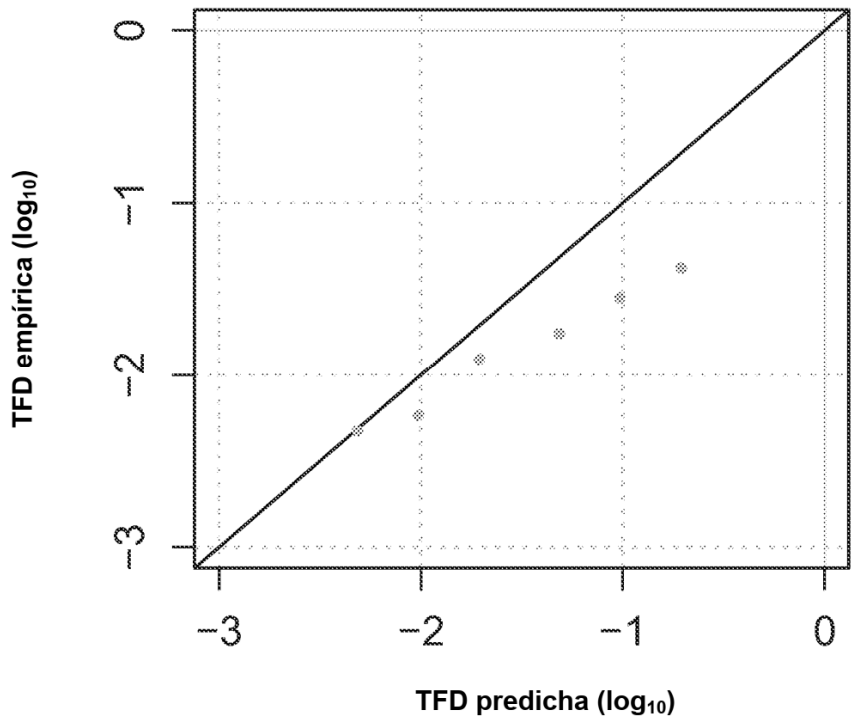

Figura 4

A

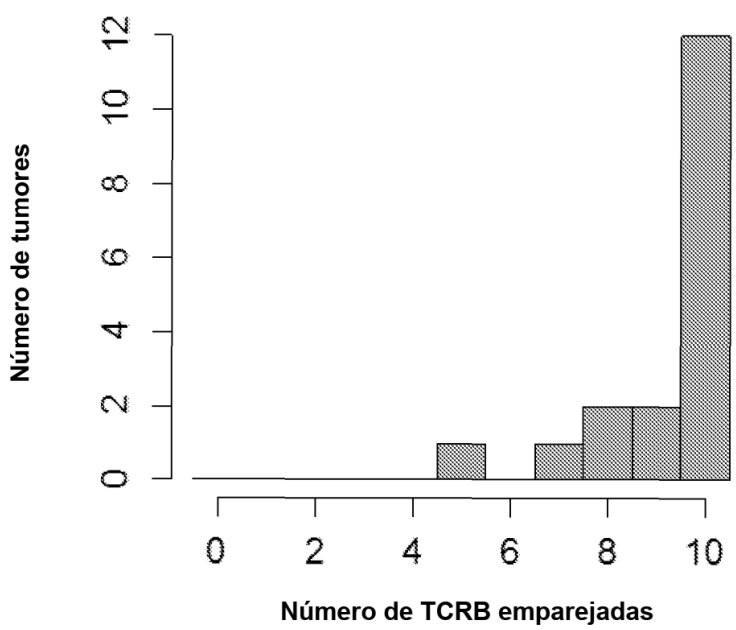
Los 10 mejores clones por tumor
Número de tumores
0
2
4
6
8
10
12
0
2
4
6
8
10
Número de TCRB emparejadas

Figura 4 continuación