(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **3 002 665**

(51) Int. Cl.:

C12Q 1/6881 (2008.01)
C12Q 1/6883 (2008.01)
C12P 19/34 (2006.01)
G01N 33/50 (2006.01)
C12Q 1/68 (2008.01)
C12Q 1/70 (2006.01)
C12Q 1/6869 (2008.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(96) Fecha de presentación y número de la solicitud europea: **24.02.2016 E 19188465 (9)**

(97) Fecha y número de publicación de la concesión europea: **23.10.2024 EP 3591074**

(54) Título: **Métodos para diagnosticar enfermedades infecciosas y para determinar el estado de HLA mediante secuenciación del repertorio inmunitario**

(30) Prioridad:

24.02.2015 US 201562120249 P
05.05.2015 US 201562157249 P
08.09.2015 US 201562215630 P

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
07.03.2025

(73) Titular/es:

ADAPTIVE BIOTECHNOLOGIES CORP. (50.00%)
1551 Eastlake Avenue East, Suite 200
Seattle, Washington 98102, US y
FRED HUTCHINSON CANCER CENTER (50.00%)

(72) Inventor/es:

EMERSON, RYAN O.;
RIEDER, MARK;
ROBINS, HARLAN S.;
DEWITT III, WILLIAM y
CARLSON, CHRISTOPHER

(74) Agente/Representante:

UNGRÍA LÓPEZ, Javier

Observaciones:

Véase nota informativa (Remarks, Remarques o Bemerkungen) en el folleto original publicado por la Oficina Europea de Patentes

Aviso: En el plazo de nueve meses a contar desde la fecha de publicación en el Boletín Europeo de Patentes, de la mención de concesión de la patente europea, cualquier persona podrá oponerse ante la Oficina Europea de Patentes a la patente concedida. La oposición deberá formularse por escrito y estar motivada; sólo se considerará como formulada una vez que se haya realizado el pago de la tasa de oposición (art. 99.1 del Convenio sobre Concesión de Patentes Europeas).

DESCRIPCIÓN

Métodos para diagnosticar enfermedades infecciosas y para determinar el estado de HLA mediante secuenciación del repertorio inmunitario

Antecedentes de la invención

El sistema inmunitario adaptativo celular transmite una amplia protección contra la infección por patógenos a través del desarrollo de un repertorio vasto y muy diverso de genes de receptores de linfocitos T (TCR), que codifican receptores de linfocitos T de superficie celular con especificidad antigénica aleatorizada. La capacidad del sistema inmunitario adaptativo de un sujeto para abordar adecuadamente una infección incipiente se basa en la activación de un receptor de linfocitos T específico de antígeno (TCR) apropiado. La interacción TCR-antígeno está mediada por la presentación de péptidos extraños en la superficie celular por células infectadas con patógenos en el entorno de proteínas del complejo principal de histocompatibilidad (MHC). Específicamente, los linfocitos T $CD8^+$ reconocen el antígeno en el entorno de las proteínas del MHC de clase I. Dado que las proteínas del MHC de clase I están codificadas por los loci A, B y C del antígeno leucocitario humano (HLA), que son muy polimórficos, la especificidad antigénica de un TCR se modula entre individuos por el entorno del HLA.

Cuando se ha encontrado un antígeno, los linfocitos T activados proliferan por expansión clonal y residen en el espacio de linfocitos T de memoria durante muchos años como una población clonal de células (clones) con genes TCR reordenados idénticos por descendencia (Arstila TP, *et al.* A direct estimate of the human alphabeta T cell receptor diversity. Science 286: 958-961, 1999). La protección contra la exposición futura a patógenos causantes de enfermedades se confiere por la capacidad de los linfocitos T activados para formar respuestas de memoria duraderas.

La mayor parte de la diversidad de los TCR reside en la cadena β del heterodímero alfa/beta de los TCR. Cada clon de linfocito T está codificado por un único alelo TCRβ que se ha reordenado aleatoriamente del locus TCRβ de la línea germinal para formar un gen TCRβ maduro. Se genera una inmensa diversidad combinando los segmentos génicos de TCRβ variable (V), de diversidad (D) y de unión (J), que codifican colectivamente la región CDR3, la región primaria del locus TCRβ para determinar la especificidad del antígeno. La deleción y la inserción de nucleótidos independiente del molde durante el reordenamiento en las uniones Vβ-Dβ y Dβ-Jβ se suman a la diversidad potencial de receptores que se pueden codificar (Cabaniols JP, *et al.* Most alpha/beta T cell receptor diversity is due to terminal deoxynucleotidyl transferase. J Exp Med 194: 1385-1390, 2001). Típicamente, en un momento dado, un adulto sano expresa aproximadamente 10 millones de cadenas únicas de TCRβ en sus 10^{12} linfocitos T en circulación (Robins HS, *et al.* (2009) Comprehensive assessment of T-cell receptor beta-chain diversity in alphabeta T cells. Blood 114: 4099-4107). Sin embargo, observar la misma cadena de TCRβ de forma independiente en dos individuos es miles de veces más común de lo que se esperaría si todos los reordenamientos fueran igualmente probables (Robins HS, *et al.* (2010) Overlap and effective size of the human CD8+ T cell receptor repertoire. Science Translational Medicine 2: 47ra64). Se espera que haya muchas secuencias de TCRβ (especialmente aquellas con pocas inserciones o ninguna) que estén presentes en los repertorios sin exposición previa de la mayoría de los individuos y que estas secuencias de TCRβ proliferen de manera confiable tras la exposición a su antígeno asociado (V. Venturi, *et al.*, The molecular basis for public T-cell responses? Nature reviews. Immunology 8, 231-238 (2008); publicado en línea EpubMar (10.1038/nri2260)). Esta sobrerrepresentación de reordenamientos específicos de la secuencia de TCRβ en el repertorio de linfocitos T sin exposición previa forma la base de las respuestas públicas de los linfocitos T.

Las respuestas públicas de los linfocitos T se producen cuando se observa que los linfocitos T que portan receptores de linfocitos T (TCR) idénticos dominan la respuesta al mismo epítopo antigénico en múltiples individuos. Se sabe que muchos antígenos patógenos inducen tal respuesta pública de linfocitos T, en la que un antígeno patógeno es la diana por la misma secuencia del receptor de linfocitos T (y se encuentra que es inmunodominante) en múltiples individuos con isotipos de HLA específicos. H. Li *et al.*, Determinants of public T cell responses. Cell research 22, 33-42 (2012); publicado en línea EpubJan (10.1038/cr.2012.1); H. Li, *et al.*, Recombinatorial biases and convergent recombination determine inter-individual TCRbeta sharing in murine thymocytes. Journal of immunology (Baltimore, Md.: 1950) 189, 2404-2413 (2012); publicado en línea EpubSep 1 (10.4049/jimmunol. 112087). En otras palabras, las respuestas públicas de linfocitos T se observan cuando el espacio de cadenas TCRβ de alta avidez potencial que se podrían unir a un complejo antígeno-MHC particular incluyen una o más cadenas TCRβ que también tienen una alta probabilidad de existir en el repertorio sin exposición previa en un momento dado. Por tanto, las secuencias asociadas con una respuesta pública de linfocitos T solo estarán presentes de forma intermitente en el espacio sin exposición previa de sujetos que no han sido expuestos a un antígeno particular; sin embargo, dichos clones deberían aparecer de forma constante en el repertorio de linfocitos T de sujetos que han estado expuestos al antígeno, habiendo experimentado expansión clonal después de la estimulación del antígeno.

El trabajo anterior sobre las respuestas públicas de los linfocitos T ha identificado ejemplos individuales de respuestas públicas de los linfocitos T a las enfermedades (incluido el CMV, EBV, gripe, esclerosis múltiple y otras enfermedades neoplásicas y afecciones autoinmunes), (Venturi V, *et al.* (2008) J Immunol 181: 7853-7862). Estos estudios se vieron limitados por la profundidad de secuenciación y el tamaño de las cohortes de investigación. Típicamente, estas respuestas públicas de linfocitos T se estudiaron en el entorno de antígenos individuales en un único entorno de HLA, normalmente, usando tetrámeros de antígeno-MHC para purificar linfocitos T específicos de antígeno. Sin embargo,

debido a limitaciones técnicas, hasta la fecha, solo se ha identificado un número relativamente pequeño de respuestas públicas de linfocitos T. No es sorprendente que estos resultados se hayan limitado a las respuestas públicas de linfocitos T de alta frecuencia y fácilmente observables que dominan la respuesta inmunitaria a sus antígenos diana. Por lo tanto, estas respuestas públicas representan solo secuencias de TCRβ que son comunes al espacio de linfocitos T sin exposición previa de casi todos los sujetos adultos (es decir, disponibles para respuesta de antígeno), y omiten ambos reordenamientos de linfocitos T raros que están presentes solo en algunos pacientes experimentados con antígenos debido a la rareza entre las secuencias de linfocitos T sin exposición previa y a las respuestas de los linfocitos T que no son inmunodominantes y, por lo tanto, nunca comprenden una fracción suficiente del repertorio total de linfocitos T para ser observados de forma fiable con una profundidad de secuenciación limitada.

Además, la interacción TCR-antígeno está mediada por la presentación de péptidos extraños en la superficie celular por células infectadas con patógenos en el entorno de proteínas del complejo principal de histocompatibilidad (MHC) de clase I. Dado que las proteínas del MHC de clase I están codificadas por los loci A, B y C del antígeno leucocitario humano (HLA), que son muy polimórficos, la especificidad antigénica de un TCR se modula entre individuos por el entorno del HLA.

La unión de los receptores de los linfocitos T a los antígenos está mediada por proteínas del MHC, que presentan el antígeno en la superficie de las células. El MHC están codificado en humanos por los loci HLA, que son muy polimórficos. Este polimorfismo da lugar a respuestas heterogéneas de linfocitos T al mismo antígeno en todos los individuos y a una selección diferencial positiva y negativa de secuencias de receptores de linfocitos T específicos durante el entrenamiento tímico. La determinación de los alelos HLA de un individuo (tipo HLA) tiene varias aplicaciones clínicas. Un ejemplo de una aplicación clínica para la tipificación de HLA es probar la idoneidad de un individuo como donante de trasplante de médula ósea.

Existe la necesidad de métodos mejorados para diagnosticar y/o predecir el estado de una persona para una enfermedad infecciosa, tal como CMV, EBV, HPV, viruela, y otros, con mayor sensibilidad y precisión. Se necesitan métodos de diagnóstico que aprovechen la información sobre el perfil de la secuencia del receptor de linfocitos T de un individuo, incluida la presencia de clones de linfocitos T públicos, y evaluar el estado de enfermedad infecciosa del individuo en función del perfil de la secuencia del receptor de linfocitos T. También existe la necesidad de inferir el tipo de HLA de un individuo basándose en el perfil de secuencia del receptor de linfocitos T del individuo. La presente invención satisface estas necesidades y proporciona ventajas adicionales relacionadas.

Sumario de la Invención

La invención proporciona un método para predecir la presencia o ausencia de una enfermedad infecciosa determinada en un sujeto con estado de infección desconocido, comprendiendo el método:

a) realizar amplificación y secuenciación de alto rendimiento de ADN genómico obtenido de una muestra que comprende linfocitos T obtenida del sujeto para determinar un perfil de TCR que comprende secuencias de aminoácidos de CDR3 únicos;
b) comparar el perfil de TCR con una base de datos de secuencias de receptores de linfocitos T públicos de diagnóstico previamente identificado que de manera estadísticamente significativa están asociadas a la enfermedad infecciosa determinada;
c) generar una primera puntuación para el sujeto, en donde la primera puntuación es la proporción de secuencias de TCR únicas en el perfil del sujeto que coincide con las secuencias de TCR públicas en la base de datos;
d) introducir la primera puntuación de c) en un algoritmo, en donde el algoritmo compara la primera puntuación del sujeto y el estado infeccioso de una pluralidad de sujetos de estado infeccioso desconocido, en donde el algoritmo es un modelo de regresión logística; y
e) determinar una probabilidad estimada de estado infeccioso del sujeto como resultado del algoritmo

La presente divulgación está basada, en parte, en métodos para determinar o predecir la presencia o ausencia de uno o más agentes de enfermedades infecciosas en un sujeto con estado de infección conocido o desconocido mediante técnicas de cuantificación inmunitarias combinadas con modelos matemáticos.

En algunas realizaciones no reivindicadas, el método comprende, además, una etapa inicial de obtención de una muestra que comprenda linfocitos T del sujeto.

En realizaciones particulares, la presente divulgación proporciona adicionalmente un método para predecir la presencia o ausencia de una o más infecciones víricas en un sujeto con estado de infección desconocido. El método comprende determinar un perfil de secuencias de TCR únicos de una muestra obtenida del sujeto (por ejemplo, una muestra que comprende linfocitos T) e ingresar estas secuencias de TCR únicos en uno o más algoritmos. En una realización, los uno o más algoritmos se generan determinando al menos 100.000 secuencias de TCR únicos de cada uno de una pluralidad de sujetos de estado de infección conocido para cada una de las una o más infecciones. En algunas realizaciones no reivindicadas, el método comprende además una etapa inicial de obtener una muestra del sujeto. Se identifican estadísticamente secuencias de TCR únicos que se correlacionan con la presencia o ausencia de cada una de las una o más infecciones, generando así una puntuación que predice la presencia o ausencia de

cada una de las una o más infecciones. Las puntuaciones se introducen luego en un modelo de regresión logística entrenado en cada uno de la pluralidad de sujetos de estado de infección conocido para cada una de las una o más infecciones, cuyo resultado es la predicción de si el sujeto es positivo o negativo para cada una de las una o más infecciones.

En algunas realizaciones, una o más infecciones son por citomegalovirus (CMV), un virus de Epstein-Barr (EBV), un virus del herpes simple (HSV) o un virus de la viruela.

En realizaciones particulares, la presente divulgación también proporciona un método para predecir la presencia o ausencia de una infección por CMV en un sujeto con estado de infección desconocido. En una realización, el ADN genómico de una muestra que comprende linfocitos T obtenidos de un sujeto se somete a amplificación y secuenciación de alto rendimiento para determinar un perfil de TCR que comprende secuencias de aminoácidos de CDR3 únicos. En algunas realizaciones no reivindicadas, el método comprende además una etapa inicial de obtener una muestra que comprende linfocitos T del sujeto. En otras realizaciones, el perfil de TCR se compara luego con una base de datos de secuencias de receptores de linfocitos T públicos de diagnóstico previamente identificados que se sabe que están asociados estadísticamente de manera significativa con la infección por CMV. En algunas realizaciones, a continuación, se genera una puntuación de carga del CMV para el sujeto determinando la proporción de secuencias de TCR únicos en el perfil del sujeto que coinciden con las secuencias de TCR públicos en la base de datos. En otras realizaciones, la puntuación de carga del CMV calculada se introduce en un modelo de regresión logística que compara la carga del CMV y el estado de infección por CMV de una pluralidad de sujetos con estado de infección por CMV conocido. En una realización, a continuación, se determina una probabilidad estimada del estado de infección por CMV para el sujeto como resultado del modelo de regresión logística.

En algunas realizaciones, se determina que la base de datos de secuencias de linfocitos T públicos está asociada estadísticamente mediante la obtención de secuencias de TCR únicos de un grupo de sujetos con infección por CMV y un grupo de sujetos sin infección por CMV. En una realización, cada secuencia de TCR únicos se somete a una prueba exacta de Fisher de una cola basada en la presencia o ausencia de cada secuencia de TCR en sujetos con estado de infección por CMV conocido. En una realización, la hipótesis nula es que cada secuencia de TCR no es más común en sujetos con una infección por CMV que en sujetos sin una infección por CMV. En una realización, se establece un umbral teórico de p-valor, y la tasa de descubrimientos falsos (TDF) se controla mediante la permutación del estado de infección por CMV en cada sujeto para generar una distribución empírica nula de p-valores. En una realización particular, a continuación, se genera una base de datos que comprende secuencias de receptores de linfocitos T que se comparten de forma estadísticamente significativa en sujetos con infección por CMV.

En algunas realizaciones, el umbral teórico de p-valor es menor o igual a $1{,}0*10^{-4}$.

En algunas realizaciones, las secuencias de TCR únicos identificadas estadísticamente que se correlacionan con la presencia o ausencia de infección por CMV comprenden una o más de las SEQ ID NO: 1 a 142. En algunas realizaciones, las secuencias de TCR únicos identificadas estadísticamente que se correlacionan con la presencia o ausencia de infección por CMV comprenden las SEQ ID NO: 1 a 142.

En algunas realizaciones, la etapa de determinar las al menos 100.000 secuencias de TCRβ únicos de la muestra incluye amplificar los ácidos nucleicos reordenados que codifican la región CDR3 de TCRβ en una reacción de PCR multiplexada con una mezcla de cebadores directos específicos de los segmentos del gen Vβ de TCR y cebadores inversos específicos de los segmentos del gen Jβ. En algunas realizaciones, las lecturas de los ácidos nucleicos amplificados se secuencian y las lecturas de secuencia se procesan para eliminar errores en la secuencia primaria de cada lectura y comprimir los datos. En realizaciones particulares, se aplica un algoritmo del vecino más cercano para colapsar los datos en secuencias únicas fusionando secuencias estrechamente relacionadas para eliminar tanto la PCR como los errores de secuenciación.

El método también incluye determinar una asociación de HLA para cada secuencia de TCR único en el perfil de TCR. En algunas realizaciones, la secuencia de TCR único está asociada con un alelo HLA-A y/o HLA-B.

El método descrito en el presente documento, pero no incluido en las reivindicaciones, comprende etapas para predecir el estado de un alelo del antígeno leucocitario humano (HLA) de un sujeto, que comprende (a) determinar un perfil de receptor inmunitario de secuencias de ADN reordenadas del receptor de linfocitos T (TCR) único para cada uno de una pluralidad de sujetos, teniendo cada sujeto un estado de alelo HLA conocido; (b) categorizar la pluralidad de sujetos basándose en (i) dicho estado del alelo HLA conocido del sujeto y (ii) una presencia o ausencia en el perfil del receptor inmunitario del sujeto de un rasgo que comprende una secuencia de ADN reordenada de TCR único; (c) determinar una puntuación estadística para la asociación entre un conjunto de rasgos y un estado de alelo HLA positivo basado en (b); (d) entrenar un modelo de aprendizaje automático usando dicho conjunto de rasgos para definir un conjunto de clasificadores para cada estado de alelo HLA; (e) introducir una o más secuencias de ADN reordenadas del TCR único de un sujeto con un estado de alelo HLA desconocido en dicho modelo de aprendizaje automático para identificar uno o más rasgos que coincidan con el conjunto de clasificadores; y (f) predecir el estado del alelo HLA de dicho sujeto basándose en dicho uno o más rasgos coincidentes.
El método comprende determinar un perfil de receptor inmunitario determinando el número total de secuencias de

TCR únicos y la frecuencia de cada secuencia de TCR único. El método también comprende determinar una puntuación estadística que comprende determinar un p-valor utilizando una prueba exacta de Fisher de dos colas.

En algunos casos del método anterior, el método incluye la determinación de un p-valor de corte para identificar un conjunto de rasgos que están asociados significativamente con un estado de alelo HLA. En otro caso, el método incluye determinar una tasa de descubrimientos falsos (TDF) de la asociación de un rasgo con un estado de alelo HLA. En otros casos, el método incluye determinar una serie de asociaciones de falsos positivos entre dicho rasgo y dicho estado del alelo HLA.

En algunos casos del método anterior, el método incluye entrenar un modelo de aprendizaje automático entrenando un modelo de regresión logística usando dicho conjunto de características identificadas y dichos estados de alelos HLA conocidos de cada sujeto. En un caso, el método incluye entrenar un modelo de aprendizaje automático que comprende realizar un método de validación cruzada de dejar uno fuera. En otro caso, el método comprende realizar dicho método de validación cruzada de dejar uno fuera durante múltiples rondas. En otro caso, la predicción tiene al menos un 80 % de precisión o al menos un 90 % de precisión.

En determinadas realizaciones, la secuencia de ADN reordenada de TCR es una secuencia de ADN reordenada TCRA, TCRB, TCRG o TCRD. En el presente documento se describe, pero no se incluye en las reivindicaciones, que el alelo HLA es un alelo HLA-A2 o un alelo HLA-24.

Breve descripción de los dibujos

Se obtendrá una mejor comprensión de las características novedosas de la invención y las ventajas de la presente invención haciendo referencia a la siguiente descripción y los dibujos adjuntos de los cuales:

La Figura (FIG.) 1 representa once eventos distintos de recombinación de CDR3β VDJ, sin dejar de producir la misma secuencia de nucleótidos y, en última instancia, la misma secuencia de aminoácidos de CDR3β (SEQ ID NO: 149). Las secuencias de nucleótidos en el lado izquierdo de la FIG. 1 son, en orden de arriba hacia abajo, SEQ ID NO: 143, SEQ ID NO: 144, SEQ ID NO: 144, SEQ ID NO: 144, SEQ ID NO: 144, SEQ ID NO: 144, SEQ ID NO: 145, SEQ ID NO: 146, SEQ ID NO: 146, SEQ ID NO: 147, SEQ ID NO:148.

Las FIG. 2A, 2B y 2C proporcionan una descripción general del método para predecir el estado de la enfermedad de un individuo, de acuerdo con una realización de la invención. La FIG. 2A muestra un conjunto de datos de muestras de sangre periférica de 640 sujetos sanos (287 CMV- y 353 CMV+), que fueron analizados mediante perfiles de receptores inmunitarios de TCR de alto rendimiento. En la FIG. 2B, se identificaron secuencias de TCRβ únicos que estaban presentes en significativamente más sujetos CMV+ que sujetos CMV-, controlando la tasa de determinaciones falsas (TDF) por permutación del estado del CMV. La presencia de estas secuencias de TCRβ asociadas al CMV se utilizó para construir un modelo de clasificación. El panel superior de la FIG. 2B representa CASSLIGVSSYNEQFF (SEQ ID NO: 12). En la FIG. 2C, el modelo de clasificación se probó mediante una validación cruzada exhaustiva de dejar uno fuera, en la que una muestra se mantuvo fuera de los cálculos y el proceso se repitió desde el principio. El modelo de clasificación resultante se utilizó para predecir el estado del CMV del sujeto excluido.

La FIG. 3A muestra una representación gráfica de un conjunto de datos de sujetos, en donde 640 sujetos fueron fenotipados específicamente para el tipo de HLA y el estado del CMB para ser elegibles como donantes de HCT, y se separaron entre 287 donantes CMV+ y 353 donantes CMV-. Hubo aproximadamente el mismo número de muestras seropositivas y seronegativas para la investigación de las respuestas de los linfocitos T públicos.

La FIG. 3B muestra las características demográficas de los sujetos incluidos en este estudio, clasificados por estado del CMV.

La FIG. 4 muestra una prueba de asociación al CMV de los TCR, en donde se seleccionó e identificó la secuencia de aminoácidos del TCR CASSLIGVSSYNEQFF (SEQ ID NO: 12) en 27 de los sujetos CMV+, y solo 2 de los sujetos CMV-, con un p-valor de $2{,}8E^{-08}$.

La FIG. 5 representa secuencias de aminoácidos de los TCR ilustrativos identificadas en sujetos CMV+ con un umbral de p-valor elegido en 10^{-4}. En orden de arriba hacia abajo, éstas son las SEQ ID NO: 1-5. Se puede ver que a un p-valor más bajo, se pueden identificar un mayor número de secuencias de TCRβ asociadas al CMB.

La FIG. 6 representa la carga del CMV, que se representa como la proporción de TCR únicos en cada sujeto que están significativamente asociados al CMV frente a aquellos TCR que no están significativamente asociados al CMV.

La FIG. 7 describe el método de validación cruzada, que deja fuera una de las 640 muestras iniciales, luego, la base de datos de secuencias de TCRβ específicas del CMV y las estadísticas asociadas se vuelven a entrenar de una manera no supervisada para eliminar el sesgo, y se clasifica el estado serológico del CMV de la muestra excluida. La FIG. 7 representa CASSLIGVSSYNEQFF (SEQ ID NO: 12).

La FIG. 8 describe los resultados del método de validación cruzada que se muestra para todos los sujetos y sujetos con validación cruzada. Los datos se presentan como el área bajo la característica operativa del receptor (ABCOR) en el eje y frente al p-valor en el eje x. Los resultados del método de validación cruzada se muestran como un gráfico de la tasa de falsos positivos frente a la tasa de verdaderos positivos. La FIG. 8 (gráfico superior) muestra datos para el rendimiento de clasificación de todos y los conjuntos de datos de validación cruzada (VC) para cada umbral de p-valor, medido como el área bajo la curva COR (ABCOR). El número sobre cada conjunto de puntos de datos se corresponde con el número de TCRβ asociados al CMV identificados en ese umbral de p-valor, y el rectángulo indica el conjunto de datos seleccionado para el análisis posterior (p-valor = 10^{-4}). La FIG. 8 (gráfico inferior) también muestra

una tasa de descubrimientos falsos (TDF) estimada para cada umbral de p-valor utilizado en la identificación de secuencias de TCRβ asociadas significativamente al CMV, utilizando permutaciones del estado del CMV. El mejor rendimiento se observa con un p-valor de 10^{-4}, que se corresponde con una TDF estimada de ~ 20 %, dando como resultado la identificación de un conjunto de 142 secuencias de TCR-β que se asociaron significativamente con el estado positivo del CMV (enumeradas en la Tabla 1).

La FIG. 9 muestra las curvas COR para los conjuntos de datos totales y de validación cruzada.

Las FIG. 10A y 10B muestran restricción de HLA de secuencias de TCRβ asociados al CMV. La FIG. 10A muestra la distribución de los alelos HLA-A en esta cohorte. La FIG. 10B muestra la distribución de los alelos HLA-B en esta cohorte. Las FIG. 10A y 10B representan las siguientes secuencias: CASSLAPGATNEKLFF (SEQ ID NO:8), CASSLIGVSSYNEQFF (SEQ ID NO:12), CASSPSRNTEAFF (SEQ ID NO:73), CASSLQAGANEQFF (SEQ ID NO: 119) y CASASANYGYTF (SEQ ID NO:118).

Las FIG. 11A y 11B muestran la incidencia de secuencias de los TCRβ reactivos al CMV previamente documentados en esta cohorte. La FIG. 11A muestra la incidencia de cada secuencia de los TCRβ en la cohorte de 640 sujetos representada gráficamente a lo largo del eje horizontal mediante la disminución de la incidencia total, con la incidencia en sujetos CMV+ por encima de la horizontal y la incidencia en sujetos CMV- por debajo de la horizontal. La FIG. 11B muestra un histograma de incidencia de estas secuencias de los TCRβ en la cohorte de 640 sujetos representados gráficamente para cada grupo de secuencias.

La FIG. 12 muestra la concordancia de las secuencias de los TCRB en la cohorte en comparación con las de la bibliografía. La FIG. 12 representa las siguientes secuencias: CASSLAPGATNEKLFF (SEQ ID NO:8), CASSLIGVSSYNEQFF (SEQ ID NO:12), CASSPSRNTEAFF (SEQ ID NO:73), CASSLQAGANEQFF (SEQ ID NO:119) y CASASANYGYTF (SEQ ID NO:118).

La FIG. 13 representa una descripción general de un método descrito en el presente documento pero que no se incluye en las reivindicaciones. La FIG. 13 representa CSARDRGIGNTIYF (SEQ ID NO: 152).

La FIG. 14 representa la selección de rasgos utilizando una prueba exacta de Fisher de dos colas para determinar la significación estadística de la asociación entre un rasgo (secuencia de TCRB único) y el estado del alelo HLA (HLA-A2+ o HLA-A2-) descrito en el presente documento, pero no incluido en las reivindicaciones. Se muestra un listado ilustrativo de características (secuencias de TCRB únicos) y el número de sujetos que tienen una característica particular y si el sujeto es positivo o negativo para el alelo HLA-A2. La FIG. 14 representa, en orden de arriba hacia abajo, las SEQ ID NO:150-161.

La FIG. 15 representa las etapas para la selección de rasgos, descritas en el presente documento pero que no se incluyen en las reivindicaciones. Se selecciona un p-valor como un punto de corte para identificar un conjunto de "rasgos de los TCR" del listado completo de las posibles secuencias de los TCR. La definición de un umbral de p-valor y la permutación del estado del alelo entre individuos proporciona una estimación de la tasa de descubrimientos falsos. Esto se realiza para cada alelo HLA, dando como resultado un conjunto de secuencias de TCRβ asociadas a los alelos para cada alelo HLA. Se utilizó un p-valor de corte de $p \leq 10^{-4}$ y una TDF de 0,1 para identificar 288 secuencias de TCRB que están asociadas positivamente con HLA-A2. Para cada una de las secuencias de TCRB asociadas a los alelos, la frecuencia de la secuencia también se determina en cada sujeto. La FIG. 15 representa, en orden de izquierda a derecha, las SEQ ID NO:162 y 163.

La FIG. 16 representa un proceso de aprendizaje automático para ajustar un modelo de regresión logística, de acuerdo con una realización de la invención.

La FIG. 17 representa una validación cruzada exhaustiva de dejar uno fuera, de acuerdo con una realización de la invención.

La FIG. 18 muestra los resultados del experimento de validación cruzada e ilustra la precisión del método.

Descripción detallada de la invención

La invención proporciona métodos para predecir y diagnosticar enfermedades infecciosas en un sujeto que son específicos de la enfermedad (específicos para cada estado de enfermedad en particular), tienen un soporte universal (no requieren procesos/reactivos diferentes para cada estado de enfermedad) y son multiplexados (son capaces de analizar múltiples estados de enfermedad simultáneamente).

En una realización, el método incluye etapas para la inmunosecuenciación de alto rendimiento de genes TCR reorganizados en sujetos sanos con estado conocido del CMV. Usando los resultados de la inmunosecuenciación, el método incluye buscar secuencias de TCRβ que están presentes en múltiples sujetos e identificar un conjunto de secuencias de TCRβ que están significativamente asociadas con el estado positivo del CMV. El método también incluye calcular un p-valor para la asociación de cada secuencia de TCRβ con el estado del CMV usando una prueba exacta de Fisher, controlar la tasa de descubrimientos falsos (TDF) mediante la permutación del estado del CMV e identificar un listado de secuencias de TCRβ asociadas al CMV (para una determinada TDF y p-valor). Se calcula una puntuación del CMV para cada sujeto como la proporción de todas las secuencias de TCRβ de ese sujeto que están representadas en el catálogo de secuencias de TCRβ asociadas al CMV. La puntuación del CMV se utiliza para distinguir entre sujetos CMV+ y CMV-.

Los agentes infecciosos incluyen patógenos, virus, bacterias, parásitos y/o microorganismos. En algunas realizaciones, los virus incluyen, pero sin limitación, miembros de la familia del virus del herpes (tales como los virus del herpes simple 1 y 2, virus de la varicela-zóster, EBV (virus de Epstein-Barr), citomegalovirus humano (CMV), virus del herpes humano 6, virus del herpes humano 7 y virus del herpes asociado al sarcoma de Kaposi), virus de la

hepatitis B (VHB), virus de la hepatitis C (VHC), virus de inmunodeficiencia humana (VIH) I y II, virus de la gripe A, virus de la gripe B, virus respiratorios sinciciales (VRS) A y B, y metaneumovirus humano (MPV). Otros ejemplos incluyen virus linfocitotrófico de linfocitos T humanos, virus del papiloma humano, ortomixovirus, paramixovirus, adenovirus, coronavirus, rabdovirus, virus de la polio, togavirus, buniavirus, arenavirus, virus de la rubéola, reovirus, norovirus, metaneumovirus humano (MPV), virus del Nilo Occidental, virus de la fiebre amarilla, virus de la rabia, rinovirus, virus de la fiebre del Valle del Rift, virus de Marburgo, virus de las paperas, virus del sarampión, virus del papiloma humano (VPH), virus del ébola, virus de la fiebre por garrapatas de Colorado (CTFV) y/o rinovirus.

Otros organismos infecciosos incluyen Escherichia coli, Salmonella, Shigella, Campylobacter, Klebsiella, Pseudomonas, Listeria monocytogenes, Mycobacterium tuberculosis, Mycobacterium avium-intracellulare, Yersinia, Francisella, Pasteurella, Brucella, Clostridia, Bordetella pertussis, Bacteroides, Staphylococcus aureus, Streptococcus pneumonia, estreptococo hemolítico del grupo B, Corynebacteria, Legionella, Micoplasma, Ureaplasma, Chlamydia, Clostridium difficile, Gardnerella, Trichomonas vaginalis, Neisseria gonorrhea, Neisseria meningitides, Hemophilus influenza, Enterococcus faecalis, Proteus vulgaris, Proteus mirabilis, Helicobacter pylori, Treponema palladium, Borrelia burgdorferi, Borrelia recurrentis, patógenos rickettsiales, Nocardia, Acitnomycetes y/o Acinetobacter.

En otras realizaciones más, los agentes infecciosos fúngicos incluyen, pero sin limitación, Cryptococcus neoformans, Blastomyces dermatitidis, Histoplasma capsulatum, Coccidioides immitis, Paracoccicioides brasiliensis, Candida albicans, Aspergillus fumigautus, Phycomycetes (Rhizopus), Sporothrix schenckii, Chromomycosis y/o Maduromycosis.

En más realizaciones, los agentes parasitarios incluyen, pero sin limitación, Plasmodium falciparum, Plasmodium malaria, Plasmodium vivax, Plasmodium ovale, Onchoverva volvulus, Leishmania, Trypanosoma spp., Schistosoma spp., Entamoeba histolytica, Cryptosporidum, Giardia spp., Trichimonas spp., Balatidium coli, Wuchereria bancrofti, Toxoplasma spp., Enterobius vermicularis, Ascaris lumbricoides, Trichuris trichiura, Dracunculus medinesis, trematódos, Diphyllobothrium latum, Taenia spp., Pneumocystis carinii y/o Necator americanis.

El complejo principal de histocompatibilidad (MHC) es un conjunto de moléculas de superficie celular codificadas por una gran familia de genes que controla una parte importante del sistema inmunitario en todos los vertebrados. La función principal de los complejos principales de histocompatibilidad es unirse a fragmentos de péptidos que provienen de patógenos y mostrarlos en la superficie celular para que los reconozcan los linfocitos T apropiados.

El complejo principal de histocompatibilidad (MHC) contiene dos tipos de genes polimórficos del MHC, los genes de clase I y de clase II, que codifican dos grupos de proteínas estructuralmente distintas pero homólogas, y otros genes no polimórficos cuyos productos están implicados en la presentación de antígenos.

El MHC humano se llama antígeno leucocitario humano (HLA). Las proteínas HLA están codificadas por genes del MHC. Los antígenos HLA de clase I incluyen HLA-A, HLA-B y HLA-C. Los antígenos HLA de clase II incluyen HLA-DR, HLA-DQ, HLA-DP, HLA-DM, HLA-DOA, HLA-DOB. Los HLA correspondientes al MHC de clase I (A, B y C) presentan péptidos del interior de la célula. Los HLA correspondientes al MHC de clase II (DP, DM, DOA, DOB, DQ y DR) presentan antígenos del exterior de la célula a los linfocitos T. Las moléculas de MHC median la unión de un receptor de linfocitos T dado a un antígeno dado, por tanto, el polimorfismo del MHC entre individuos modula la respuesta de los linfocitos T a un antígeno determinado. Las aplicaciones clínicas de la tipificación de HLA incluyen ensayos de vacunas, asociaciones de enfermedades, reacciones adversas a fármacos, transfusión de plaquetas y trasplante de órganos y células madre.

Existen numerosos alelos en cada locus del gen HLA (A1, A2, A3, etc.). Cada persona hereda un conjunto completo de alelos de HLA (haplotipo) de cada padre, y esta combinación de proteínas codificadas constituye el tipo HLA de una persona (por ejemplo, diferentes antígenos de A23, A31, B7, B44, C7, C8, DR4, DR7, DQ2, DQ7, DP2, DP3). Hay más de 50.000 tipos diferentes de HLA conocidos.

Tal 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 determinadas realizaciones, el receptor inmune adaptativo está codificado por un gen o segmento génico de TCRB, TCRG, TCRA, TCRD, IGH, IGK e IGL.

El término "cebador", tal 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. 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.

Tal 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 cola", así como secuencias intermedias (intrones) entre segmentos de codificación individuales (exones), y también puede

incluir elementos reguladores (por ejemplo, promotores, potenciadores, sitios de unión del represor y similares), y también puede incluir 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 inmunoglobulina o una región de la misma (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, dominios de la región V, 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.

A menos que se proporcionen definiciones específicas, la nomenclatura utilizada en relación con los procedimientos y técnicas de laboratorio de, biología molecular, química analítica, química orgánica sintética y química medicinal y farmacéutica descritas en el presente documento son las bien conocidas y comúnmente utilizadas en la técnica. Se pueden usar técnicas estándar para tecnología recombinante, biología molecular, microbiológica, síntesis química, análisis químicos, preparación farmacéutica, formulación y administración y tratamiento de pacientes.

A menos que el contexto requiera lo contrario, a lo largo de la presente especificación y reivindicaciones, la palabra "comprender" y variaciones de la misma, tales como, "comprende" y "que comprende" se deben interpretar en un sentido abierto, inclusivo, es decir, como "que incluye, pero no se limita a". Por "que consiste en" se entiende que incluye, y normalmente se limita a, lo que sigue en la expresión "que consiste en". Por "que consiste esencialmente en" se entiende que incluye cualquiera elementos enumerados después de la expresión, 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 tanto, la expresión "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.

En esta memoria descriptiva y en las reivindicaciones adjuntas, las formas en singular "un", "una", "el" y "la" incluyen referencias en plural, a menos que el contenido indique claramente otra cosa.

La referencia a lo largo de la presente memoria descriptiva a "una sola realización" o "una realización" o "un aspecto" significa que un rasgo, estructura o característica particular descrita en relación con la realización se incluye en al menos una realización de la presente invención. Por tanto, las apariciones de las expresiones "en una sola realización" o "en una realización" en diversos 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

Células

Se puede obtener una muestra que contiene moléculas de ácido nucleico linfoide (ADN genómico, ADNc o, como alternativa, ARN mensajero) de un sujeto. El sujeto es un sujeto mamífero, tal como un ser humano.

Los linfocitos (linfocitos B y/o linfocitos T) se pueden obtener a partir de una muestra biológica, tal como a partir de una variedad de muestras de tejidos y fluidos biológicos. Estos incluyen pero no se limitan a la médula ósea, timo, glándulas linfáticas, ganglios linfáticos, tejidos periféricos y sangre o muestras de tejido sólido. Se puede tomar una muestra de cualquier tejido periférico para detectar la presencia de linfocitos B y T y, por lo tanto, se contempla su uso en los métodos descritos en el presente documento. 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 determinadas realizaciones, se utilizan PBMC completas para el análisis.

Extracción de ácido nucleico

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 sangre de ADN QIAamp® (QIAGEN®) o un kit de extracción de sangre Qiagen DNeasy (Qiagen, Gaithersburg, MD, EE. UU.). 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 se puede estimar en aproximadamente el 30 % del total de células. Como alternativa, el ácido nucleico total se puede aislar de las células, incluyendo tanto ADN genómico como ARNm. En otras realizaciones, el ADNc se transcribe a partir del ARNm y luego se usa como molde para la amplificación.

PCR cuantitativa multiplexada

La PCR cuantitativa multiplexada se realizó tal como se describe en el presente documento y 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; los documentos US 2012/0058902, US 2010/0330571, WO/2010/151416, WO/2011/106738 (PCT/US2011/026373), US 2015/0299785, WO2012/027503 (PCT/US2011/049012), US 2013/0288237, US 9.181.590, US 9.181.591 y US 2013/0253842. Los métodos actuales implican un 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 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 de segmento J se describen en los documentos US2012/0058902, US 2010/033057, WO/2010/151416, WO/2011/106738 (PCT/US2011/026373), US 2015/0299785, WO2012/027503 (PCT/US2011/049012), US 2013/0288237, US 9.181.590, US 9.181.591, US 2013/0253842, WO 2013/188831 (PCT/US2013/045994).

Se puede utilizar un sistema de PCR multiplexada para amplificar los loci reordenados de receptores de células inmunitarias adaptativas. En determinadas realizaciones, la región CDR3 se amplifica a partir de una región CDR3 de TCRA, TCRB, TCRG o TCRD o de manera similar a partir de un locus IgH o IgL (lambda o kappa). Se utiliza una pluralidad de cebadores del segmento V y del segmento J para amplificar sustancialmente todas (por ejemplo, más del 90 %, del 91 %, del 92 %, del 93 %, del 94 %, del 95 %, del 96 %, del 97 %, del 98 % o del 99 %) las regiones que codifican la CDR3 del receptor inmune adaptativo reordenadas para producir una multiplicidad de moléculas de ADN reordenadas amplificadas. En determinadas 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 reordenados.

En algunas realizaciones, Se utilizan dos grupos de cebadores en una sola reacción de PCR altamente multiplexada. El grupo de cebadores "directos" puede incluir una pluralidad de cebadores de oligonucleótidos del segmento V y el grupo inverso puede incluir una pluralidad de cebadores de oligonucleótidos del segmento J. En algunas realizaciones, hay un cebador que es específico (por ejemplo, que tiene una secuencia de nucleótidos complementaria a una región de secuencia única de) cada segmento de la región V y para cada segmento de la región J en el respectivo locus del gen de TCR o Ig. En otras realizaciones, un cebador puede hibridar con uno o más segmentos V o segmentos J, reduciendo así el número de cebadores requeridos en la PCR multiplexada. En determinadas 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.

En una realización, la presente divulgación proporciona una pluralidad de cebadores del segmento V y una pluralidad de cebadores del segmento J. 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 AIR reordenadas 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 80 %, el 85 %, el 90 %, el 91 %, el 92 %, el 93 %, el 94 %, el 95 %, el 96 %, el 97 %, el 98 %, el 99 % o más de todas las combinaciones de los segmentos V y J de un locus del receptor inmunitario reordenado. En determinadas 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 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. 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 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; US 2012/0058902, US 2010/033057, WO/2010/151416, WO/2011/106738 (PCT/US2011/026373), US 2015/0299785, WO2012/027503 (PCT/US2011/049012), US 2013/0288237, US 9.181.590, US 9.181.591, US 2013/0253842, WO 2013/188831 (PCT/US2013/045994).

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. En una realización, las condiciones de rigurosidad moderada a alta adecuadas para la amplificación por PCR específica de una secuencia de ácido nucleico diana puede estar 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. Tal como reconocería el 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.

Un cebador es preferentemente un ADN monocatenario. La longitud apropiada de un cebador depende del uso previsto del cebador, pero típicamente varía de 6 a 50 nucleótidos, o en ciertas realizaciones, de 15-35 nucleótidos de longitud. 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.

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 determinadas realizaciones, los amplicones varían en tamaño de 10, 20, 30, 40, 50, 75, 100, 200, 300, 400, 500, 600, 700, 800 o más nucleótidos de longitud. En realizaciones preferidas, los amplicones tienen un tamaño entre 50-600 nucleótidos de longitud.

Según la teoría no limitativa, estas realizaciones aprovechan la comprensión actual en la técnica (también descrita anteriormente) de que una vez que una célula inmunitaria adaptativa (por ejemplo, un linfocito T o B) ha reordenado sus genes codificantes del receptor inmunitario adaptativo (por ejemplo, genes de TCR o Ig), sus células de progenie poseen el mismo reordenamiento del gen codificante del receptor inmune adaptativo, dando así lugar a una población clonal (clones) que puede ser identificada de forma única por la presencia en ella de segmentos génicos V y J reordenados (por ejemplo, codificantes de CDR3) que se pueden amplificar mediante una combinación específica por pares de cebadores de oligonucleótidos específicos de V y J tal como se describe en el presente documento.

En algunas realizaciones, los cebadores del segmento V y los cebadores del segmento J incluyen, cada uno, una segunda secuencia en el extremo 5' del cebador que no es complementaria al segmento V o J diana. La segunda secuencia puede comprender un oligonucleótido que tiene una secuencia que se selecciona de (i) una secuencia oligonucleotídica del adaptador universal, y (ii) una secuencia oligonucleotídica específica de la plataforma de secuenciación que está unida a y posicionada en 5' a una primera secuencia oligonucleotídica del adaptador universal. Ejemplos de secuencias de oligonucleótidos del adaptador universal pueden ser las secuencias del adaptador pGEX directa y pGEX inversa.

En algunas realizaciones, los amplicones resultantes que usan los cebadores del segmento V y del segmento J descritos anteriormente incluyen segmentos V y J amplificados y las secuencias de oligonucleótidos del adaptador universal. La secuencia del adaptador universal puede ser complementaria a una secuencia de oligonucleótidos que se encuentra en un cebador de cola. Los cebadores de colas se pueden usar en una segunda reacción de PCR para generar un segundo conjunto de amplicones. En algunas realizaciones, los cebadores de cola pueden tener la fórmula general:

5'-P-S-B-U-3' (III),

en donde P comprende un oligonucleótido específico de la plataforma de secuenciación,
en donde S comprende una secuencia de oligonucleótidos que contiene una etiqueta de plataforma de secuenciación;
en donde B comprende una secuencia de código de barras de oligonucleótidos y en donde dicha secuencia de código de barras de oligonucleótidos se puede usar para identificar una fuente de muestra, y
en donde U comprende una secuencia que es complementaria a la secuencia de oligonucleótidos del adaptador universal o es la misma que la secuencia de oligonucleótidos del adaptador universal.

La descripción adicional sobre secuencias de oligonucleótidos adaptadores universales, los códigos de barras y los

cebadores de colas se encuentran en el documento WO 2013/188831 (PCT/US13/45994).

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, se pueden usar métodos que proporcionan una composición de molde 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 molde constituyentes 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 determinadas 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, independientemente, 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.

Se proporciona una descripción adicional sobre los métodos de control de sesgo en los documentos US 2013/0253842, US 9.150.905, US 2015/0203897 y WO 2013/169957 (PCT/US2013/040221).

Secuenciación

La secuenciación se puede realizar usando cualquiera de una variedad de máquinas y sistemas de secuenciación de una sola molécula de alto rendimiento disponibles. Los sistemas de secuencia ilustrativos incluyen sistemas secuencia por síntesis, tales como el analizador del genoma de Illumina y los equipos asociados (Illumina HiSeq) (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 específicos de la plataforma de secuenciación que hibridan con una región definida dentro de las moléculas de ADN amplificadas. Los oligonucleótidos específicos de la plataforma de secuenciación están diseñados para secuenciar amplicones, de manera que los segmentos génicos que codifican V y J se pueden identificar de forma única por las secuencias que se generan. Véanse, por ejemplo, los documentos US 2012/0058902; US 2010/033057; WO 2011/106738 (PCT/US2011/026373); US 2015/0299785; o WO 2012/027503 (PCT/US2011/049012).

En algunas realizaciones, los datos de secuencia sin procesar se preprocesan para eliminar errores en la secuencia primaria de cada lectura y comprimir los datos. Se puede utilizar un algoritmo del vecino más cercano para contraer los datos en secuencias únicas fusionando secuencias estrechamente relacionadas, para eliminar errores de secuenciación y de PCR. Véanse, por ejemplo, US 2012/0058902; US 2010/033057; WO 2011/106738 (PCT/US2011/026373); US 2015/0299785; o WO 2012/027503 (PCT/US2011/049012).

Estimación de la abundancia del molde de la PCR

Para estimar la cobertura de lectura promedio por molde de entrada en el enfoque de secuenciación y PCR multiplexada, se puede utilizar un conjunto de moldes sintéticos de TCR (o BCR) (tal como se describe anteriormente), que comprende cada combinación de segmentos génicos Vβ y Jβ. Estas moléculas sintéticas pueden ser las descritas en la fórmula general (I) anterior y en el documento US 2013/0253842, US 9.150.905, US 2015/0203897 y WO 2013/169957 (PCT/US2013/040221).

Estas moléculas sintéticas se pueden incluir en cada reacción de PCR a una concentración muy baja de modo que solo se observen algunos de los moldes sintéticos. Utilizando la concentración conocida del conjunto de moldes sintéticos, se puede simular la relación entre el número de moléculas sintéticas únicas observadas y el número total

de moléculas sintéticas añadidas a la reacción (esto es casi uno a uno a las bajas concentraciones que se utilizaron). Las moléculas sintéticas permiten calcular para cada reacción de PCR el número medio de lecturas de secuenciación obtenidas por molécula de molde de PCR y una estimación del número de linfocitos T en el material de entrada que lleva cada reordenamiento de TCR único.

Descubrimiento de respuestas de linfocitos T públicos de diagnóstico

Dada una gran población de sujetos con y sin una determinada enfermedad infecciosa, las respuestas de los linfocitos T públicos de diagnóstico para el estado patológico se pueden determinar aplicando el siguiente método estadístico. Se determina un repertorio de receptores inmunes de secuencias únicas de receptores de linfocitos T, usando los métodos descritos anteriormente, para un grupo de sujetos que han sido identificados como que tienen o que no tienen una enfermedad infecciosa, tal como CMV o viruela. Es posible, a continuación, determinar cuáles de esas secuencias de receptores de linfocitos T son significativamente más comunes (es decir, están presentes en más individuos) entre sujetos con el estado de enfermedad que en sujetos sin el estado de enfermedad. Cuando un TCR común se une a un antígeno, se llama respuesta de linfocitos T públicos. Las respuestas de las linfocitos T públicos son específicas para una enfermedad o antígeno en particular, están presentes en muchos individuos y están codificadas en un formato común (secuencias de receptores reorganizadas específicas) independientemente de la enfermedad.

Se utiliza una prueba exacta de Fisher de una cola, usando la presencia o ausencia de la secuencia del receptor de linfocitos T en cuestión frente a con o sin estado de enfermedad para construir una tabla de contingencia de 2x2 (que se muestra a continuación) y con la TFD (tasa de descubrimientos falsos) controlada usando una distribución empírica de p-valores nulos determinados por permutaciones del estado de enfermedad. Dado que muchos clones son exclusivos de un solo sujeto (y, en consecuencia, exclusivos de las clases de estado de enfermedad positivo o negativo), es vital controlar la tasa de descubrimientos falsos en la selección de rasgos para evitar un ajuste excesivo a las muchas asociaciones falsas de TCRβ únicos con estado positivo de enfermedad infecciosa. Este proceso genera un listado de secuencias de receptores de linfocitos T significativamente más comunes en sujetos con el estado de enfermedad de interés.

	Enfermedad+	**Enfermedad-**
clon *i* presente	n_{i+}	n_{i-}
clon *i* no presente	$N_{+} - n_{i+}$	$N_{-} - n_{i-}$

Uso de respuestas de diagnóstico de linfocitos T públicos para inferir el estado de la enfermedad

Dada una gran población de sujetos con y sin una enfermedad infecciosa determinada, y dado un listado de respuestas de linfocitos T públicos de diagnóstico generado tal como se describe anteriormente, es posible inferir el estado de la enfermedad en un sujeto cuyo estado de enfermedad se desconoce.

Primero, se calcula una "carga de enfermedad" o medida cuantitativa de la presencia y/o abundancia de dichas respuestas de linfocitos T públicos de diagnóstico para cada sujeto (aquellos con estado conocido así como el sujeto de estado desconocido). La carga de enfermedad o medida cuantitativa es la proporción de secuencias de receptores de linfocitos T únicos en cada sujeto que se encuentran entre el listado de respuestas de linfocitos T públicos de diagnóstico.

Una vez calculada esta medida, luego se determina si el sujeto de estado desconocido tiene una carga de enfermedad consistente con sujetos de los que se sabe que tienen el estado de enfermedad de interés o con sujetos de los que se sabe que no tienen el estado de enfermedad de interés. En el presente documento, el método de comparación es entrenar un modelo de regresión logística de (carga de enfermedad frente a presencia de estado de enfermedad) en todos los sujetos con estado de enfermedad conocido, y usar ese modelo para asignar al sujeto de estado desconocido una probabilidad de tener el estado de enfermedad de interés.

En algunas realizaciones, el método comprende un modelo de clasificación que utiliza una validación cruzada exhaustiva de dejar uno fuera, en la que el conjunto de datos del perfil inmunitario de un individuo se excluye de los cálculos y el proceso se repite desde el principio. En un aspecto, la validación cruzada se utiliza para evaluar la precisión de un modelo de clasificación. El modelo de clasificación resultante se utiliza para predecir el estado de enfermedad infecciosa del sujeto que se ha quedado fuera.

Estos métodos se pueden realizar para determinar el estado de la enfermedad para varios tipos de enfermedad, incluyendo, pero sin limitación, CMV, EBV, HPV, VIH, viruela, otras enfermedades infecciosas, o incluso enfermedades no infecciosas, como enfermedades autoinmunes y neoplasias.

Tipificación de HLA

La tipificación de HLA de sujetos donantes de médula ósea se realizó de acuerdo con el protocolo estándar del Centro

de Investigación del Cáncer Fred Hutchinson, y se puede realizar mediante métodos conocidos por los expertos en la técnica.

Método para determinar el estado de HLA de sujetos desconocidos

La FIG. 13 muestra una descripción general del método para inferir el estado de HLA. Este método se puede aplicar a cualquier tipo de alelo HLA y a cualquier perfil de secuencias de los TCR, que incluyen las secuencias génicas de TCRA, TCRB, TCRG o TCRD.

Primero, el tipo de alelo HLA se determina para cada sujeto de un grupo, de acuerdo con métodos estándar en la técnica para la tipificación de HLA. A continuación, se secuencia un locus del repertorio inmunitario del sujeto. Por ejemplo, el locus TCRB se amplifica y secuencia, usando los métodos descritos anteriormente.

Para cada sujeto, se determina el número total de secuencias de TCRB único y la frecuencia de cada secuencia de TCRB único. Para un alelo HLA dado, se determina qué secuencias de TCRB están asociadas significativamente con el alelo de HLA. En el presente documento se describe, pero no se incluye en las reivindicaciones, que para cada secuencia de TCRB único, se determina cuántos sujetos que son positivos para un alelo HLA tienen la secuencia de TCRB y cuántos sujetos que son negativos para el alelo HLA tienen la secuencia de TCRB. Además, se puede determinar el número de sujetos que son positivos para un alelo HLA y negativos para la secuencia de TCRB y el número de sujetos que son negativos para el alelo HLA y positivos para la secuencia de TCRB. La siguiente tabla muestra la categorización de los sujetos según la presencia o ausencia de una secuencia de TCRB y la presencia o ausencia de un alelo de HLA, HLA-A2.

	HLA-A2+	**HLA-A2-**
secuencia *i* de TCRβ presente	n_{i+}	n_{i-}
secuencia *i* de TCRβ no presente	$N_+ - n_{i+}$	$N_- - n_{i-}$

La FIG. 14 muestra un listado de secuencias de TCRB únicos ilustrativos y el número de sujetos que son positivos o negativos para un alelo HLA-A2 que tienen una secuencia de TCRB particular.

Se determina un p-valor para la asociación de cada secuencia de TCRB con un estado de HLA utilizando una prueba exacta de Fisher (de dos colas). El p-valor para la asociación de cada TCR con el estado del alelo usando una prueba exacta de Fisher (dos colas) se calcula de la siguiente manera:

$$\Pr(\text{tabla}) = \frac{\binom{n_+ + n_-}{n_+}\binom{N_+ + N_- - n_+ - n_-}{n_+ + n_-}}{\binom{N_+ + N_-}{N_+}}$$

La FIG. 14 también muestra p-valores ilustrativos que se calculan para la asociación de una secuencia de TCRB particular con un tipo de HLA (HLA-A2).

Tal como se muestra en la FIG. 15, se selecciona un p-valor como un punto de corte para identificar un conjunto de "rasgos de los TCR" del listado completo de las posibles secuencias de los TCR. La definición de un umbral de p-valor y la permutación del estado del alelo entre individuos proporciona una estimación de la tasa de descubrimientos falsos. Esto se realiza para cada alelo HLA, dando como resultado un conjunto de secuencias de TCRβ asociadas a alelos para cada alelo HLA presente en los datos de entrenamiento. En la FIG. 15, se usa un valor de corte del p-valor de $p \leq 10^{-4}$ y una TDF de 0,1 para identificar 288 secuencias de TCRB que están asociadas positivamente con HLA-A2. Para cada una de las secuencias de TCRB asociadas a los alelos, la frecuencia de la secuencia también se determina en cada sujeto.

La etapa de selección de rasgos va seguida de un proceso de aprendizaje automático. Tal como se muestra en la FIG. 16, para cada alelo HLA, se entrena un modelo de regresión logística utilizando el conjunto de vectores de características sobre todos los sujetos, junto con el estado conocido de la presencia de ese alelo.

Tal como se muestra en la FIG. 17, se realiza una validación cruzada exhaustiva de dejar uno fuera donde se elimina un sujeto del análisis, y el estado de HLA del sujeto se infiere basándose en la selección de rasgos y el entrenamiento de solo los sujetos restantes. El resultado es un conjunto de clasificadores (uno para cada alelo HLA) que estiman la probabilidad de un estado positivo para cada alelo HLA, tomando como entrada el vector de características para cada alelo.

Equipado con estos clasificadores, el tipo de HLA de un nuevo sujeto se puede evaluar mediante: 1. secuenciación

del repertorio inmunitario, 2. calcular los vectores de rasgos para cada alelo, y 3. definir un umbral de probabilidad para el estado positivo que se aplicará a la salida de cada clasificador. De esta forma, se puede inferir un estado de HLA para un nuevo sujeto con un estado de HLA desconocido.

En el presente documento se describe, pero no se incluye en las reivindicaciones, que, dados los resultados del esquema de selección de los rasgos, una puntuación alélica de una dimensión se determina como la fracción de las secuencias de TCRβ únicos de un individuo que aparecen en el conjunto de secuencias de TCRβ asociados con un alelo dado. Esta única cantidad, en lugar de un vector de rasgos, también se puede utilizar para entrenar a un clasificador y diagnosticar nuevos individuos con una precisión similar.

Adicionalmente, el clasificador específico descrito (regresión logística) se puede reemplazar con cualquiera de otros clasificadores binarios y da resultados sustancialmente similares. Éstos incluyen: k-vecinos más cercanos, bosques aleatorios, red neuronal artificial, naive Bayes y máquina de vectores de soporte.

Ejemplos

Los ejemplos siguientes se ofrecen únicamente con fines ilustrativos y no pretenden limitar el alcance de la presente invención de ninguna manera. Se han realizado esfuerzos para asegurar la exactitud con respecto a los números utilizados (por ejemplo, cantidades, temperaturas, etc.), pero, por supuesto, deberían permitirse algunos errores y desviaciones experimentales.

La práctica de la presente invención empleará, a menos que se indique otra cosa, métodos convencionales de química de proteínas, bioquímica, técnicas de ADN recombinante y farmacología, dentro de la habilidad de la técnica. Dichas técnicas se explican completamente en la bibliografía. Véanse, por ejemplo, T.E. Creighton, Proteins: Structures and Molecular Properties (W.H. Freeman and Company, 1993); A.L. Lehninger, Biochemistry (Worth Publishers, Inc., adición actual); Sambrook, *et al.*, Molecular Cloning: A Laboratory Manual (2ª Edición, 1989); Methods In Enzymology (S. Colowick y N. Kaplan eds., Academic Press, Inc.); Remington's Pharmaceutical Sciences, 18ª edición (Easton, Pennsylvania: Mack Publishing Company, 1990); Carey and Sundberg Advanced Organic Chemistry 3ª Ed. (Plenum Press) Vol. A y B(1992).

Ejemplo 1: Identificación de TCR asociados con citomegalovirus (CMV) y clasificación del estado de CMV a partir de datos de inmunosecuenciación

El primer objetivo del estudio fue crear un catálogo completo de secuencias de receptores de linfocitos T específicos de CMV. El objetivo era identificar secuencias de TCRβ compartidas, tanto comunes como raras, enriquecidas en individuos seropositivos para CMV en relación con individuos seronegativos para CMV. El CMV se utilizó como modelo porque proporciona un banco de pruebas ideal para el desarrollo de biomarcadores de linfocitos T. El estado serológico del CMV está ampliamente disponible y el fácil acceso a un número casi igual de individuos seropositivos (casos) y seronegativos (control) entre adultos sanos proporciona un buen poder estadístico para un análisis riguroso.

Hay muchas posibles secuencias de TCR generados durante la recombinación VDJ (~ 10^{12} fácilmente accesibles mediante recombinación VDJ). Tal como se muestra en la FIG. 1, diferentes eventos de recombinación de VDJ pueden dar como resultado la misma secuencia de nucleótidos. Además, es posible que la recombinación VDJ pueda dar como resultado diferentes secuencias de nucleótidos que se traducen en la misma secuencia de aminoácidos. Véase Venturi, Price, Douek y Davenport 2008. The molecular basis for public T-cell responses? Nature Reviews Immunology. Ciertos TCR son reordenamientos comunes y compartidos por más de un individuo. Cuando un TCR común se une a un antígeno, se llama respuesta de linfocitos T públicos. Las respuestas de las linfocitos T públicos son específicas para una enfermedad o antígeno en particular, están presentes en muchos individuos y están codificadas en un formato común (secuencias de receptores reorganizadas específicas) independientemente de la enfermedad.

En este estudio, la respuesta de los linfocitos T públicos al citomegalovirus (CMV) se examinó mediante la secuenciación de genes de receptores de linfocitos T reorganizados (TCR) obtenidos de 640 sujetos. El CMV es una infección de por vida, generalmente asintomática, que afecta a la mayoría de los adultos y provoca una sólida respuesta de linfocitos T de memoria.

Las FIG. 2A, 2B y 2C proporcionan una descripción general del método, de acuerdo con una realización de la invención. La FIG. 2A muestra un conjunto de datos de muestras de sangre periférica de 640 sujetos sanos (287 CMV- y 353 CMV+), que fueron analizados mediante perfiles de TCR de alto rendimiento. En la FIG. 2B, se identificaron secuencias de TCRβ únicos que estaban presentes en significativamente más sujetos CMV+ que sujetos CMV-, controlando la tasa de descubrimientos falsos (TDF) por permutación del estado de CMV. La presencia de estas secuencias de TCRβ asociadas al CMV se utilizó para construir un modelo de clasificación. En la FIG. 2C, el modelo de clasificación se probó utilizando una validación cruzada exhaustiva de dejar uno fuera, en la que una muestra se mantuvo fuera de los cálculos y el proceso se repitió desde el principio. El modelo de clasificación resultante se utilizó para predecir el estado del CMV del sujeto excluido.

Se obtuvieron muestras de sangre periférica humana del biorrepositorio de donantes de médula ósea sanos del Fred

Hutchinson Cancer Research Center Research Cell Bank bajo un protocolo siguiendo el consentimiento informado por escrito aprobado y supervisado por el Fred Hutchinson Cancer Research Center Institutional Review Board. Este biodepósito alberga un inventario de PBMC, B-LCL y ADN de pacientes con trasplante de células hematopoyéticas (TCH), donantes y miembros de la familia junto con líneas celulares y ADN que proviene de pacientes con trasplante de células madre, donantes y familiares seleccionados de ~ 7.800 pacientes con trasplante de células madre y 7.300 donantes (de los cuales se extrajeron muestras). El ADN para este estudio se extrajo de sangre periférica de donantes de TCH y se tipificó ampliamente para antígenos y alelos HLA junto con otros fenotipos definitorios necesarios para la compatibilidad entre donante y paciente o la inclusión/exclusión de donantes. Entre estos se encuentran el citomegalovirus (CMV), virus de Epstein-Barr (EBV), virus del herpes simple (HSV), hepatitis y diabetes mellitus.

Tal como se muestra en la FIG. 3A, 640 sujetos fueron fenotipados específicamente para el tipo de HLA y el estado de CMV para ser elegibles como donantes de TCH, y se separaron entre 287 donantes de CMV y 353 donantes de CMV. La tasa de seropositivos para el CMV fue de aproximadamente el 45 % (aproximadamente el mismo número de muestras seropositivas y seronegativas para la investigación de las respuestas de los linfocitos T públicos).

La FIG. 3B muestra las características demográficas de los sujetos incluidos en este estudio, clasificados por estado del CMV.

Secuenciación de TCRβ de alto rendimiento:

El ADN genómico se extrajo de muestras de sangre periférica utilizando el kit de extracción de sangre Qiagen DNeasy (Qiagen, Gaithersburg, MD, EE. UU.). Se secuenciaron las regiones de CDR3 de los genes de TCR-β reordenados; la región de CDR3 de TCR-β se definió de acuerdo con la colaboración IMGT[28,20]. Las regiones de CDR3 de TCRβ se amplificaron y secuenciaron utilizando los métodos descritos anteriormente y en los protocolos descritos anteriormente[5,29]. El método de PCR multiplexada utilizó una mezcla de 60 cebadores directos específicos para los segmentos del gen TCR Vβ y 13 cebadores inversos específicos para los segmentos del gen TCR Jβ. Los amplicones resultantes se secuenciaron utilizando los métodos descritos anteriormente. Se obtuvieron lecturas de 87 pb utilizando el sistema Illumina HiSeq. Los datos de secuencia de HiSeq sin procesar se preprocesaron para eliminar errores en la secuencia principal de cada lectura y para comprimir los datos. Se utilizó un algoritmo del vecino más cercano para contraer los datos en secuencias únicas fusionando secuencias estrechamente relacionadas, para eliminar errores de secuenciación y de PCR.

Con el fin de asegurar una cobertura adecuada de cada reordenamiento de linfocitos T, se generó una cobertura de secuencia de 8-10x, o ~ 6-10 millones de lecturas de secuenciación por muestra, usando aproximadamente ocho ejecuciones de secuenciación completas. Todas las lecturas de secuenciación se procesaron utilizando un modelo bioinformático estandarizado en desarrollo para 1) demultiplexar lecturas para muestras específicas, 2) eliminar secuencias de baja calidad y eliminar posibles contaminantes, 3) alinear e identificar segmentos de genes específicos de las regiones CDR3 de V y J de TCRβ, 4) agrupar altamente secuencias similares para tener en cuenta la PCR y los errores de secuenciación, 5) normalizar los datos para eliminar el sesgo de amplificación de la PCR, 6) estimar la entrada total de linfocitos T y 7) generar recuentos y distribuciones de secuencias únicas de TCRβ.

En el presente documento, se secuenciaron aproximadamente 250.000 genes de receptores de linfocitos T reordenados a partir de sangre periférica de cada sujeto.

Selección de rasgos:

En 640 sujetos, había 185.204 (+/- 84.171) TCRβ únicos por sujeto y 83.727.796 TCRβ únicos en total. En lugar de intentar una clasificación de CMV de alta dimensión utilizando todos los TCRβ únicos como posibles rasgos, se desarrolló un nuevo esquema de selección de rasgos. La selección de rasgos fue la selección de una secuencia de aminoácidos particular para identificar TCR comunes entre individuos con CMV.

Dado que muchos clones eran exclusivos de un solo sujeto (y, en consecuencia, exclusivos de las clases CMV+ o CMV-), fue vital controlar la tasa de descubrimientos falsos en la selección de características para evitar un ajuste excesivo a las muchas asociaciones espurias de TCRβ únicos con el estado de CMV.

Cada reordenamiento único de TCRβ, identificado por la asignación de genes V y J y la secuencia de aminoácidos de CDR3, fue probado para la asociación de CMV. Cada uno de estos se sometió a una prueba exacta de Fisher de una cola para determinar su incidencia en sujetos CMV- y CMV+. Específicamente, dejando n_{ij} denotar el número de sujetos con estatus de CMV j (con j - o +) y clon i presente, se calculó un p-valorp_i realizando la prueba exacta de Fisher en la tabla de contingencia.

	CMV+	**CMV-**
secuencia *i* de TCRβ presente	n_{i+}	n_{i-}
secuencia *i* de TCRβ no presente	$N_{+} - n_{i+}$	$N_{-} - n_{i-}$

en donde N_{+} y n_{-}denotan el número total de sujetos que tienen cada estado de CMV (CMV+ y CMV-, respectivamente).

Para caracterizar una región de rechazo en presencia de muchas hipótesis débilmente dependientes (una para cada TCRβ único), las asignaciones de estado de CMV se permutaron aleatoriamente 100 veces y se calcularon las estadísticas sobre el número de rechazos en el umbral teórico de p-valor. Aproximando la fracción total de hipótesis nulas verdaderas como unidad, esto permitió la estimación de la tasa de descubrimientos falsos (TDF) como la proporción entre el número medio de rechazos bajo permutación y el número real de rechazos.

Tal como se muestra por ejemplo en la FIG. 4, se seleccionó la secuencia de aminoácidos de TCR "CASSLIGVSSYNEQFF" (SEQ ID NO: 12) y se identificó en 27 sujetos CMV+ y 2 sujetos CMV- (p= 2,8E-08).

Las FIG. 5 muestra secuencias de aminoácidos de TCR ilustrativas identificadas en sujetos CMV+. Tal como se muestra en la Tabla 1, a un p-valor seleccionado de 10^{-4}, se identificaron 142 clones de linfocitos T públicos como asociados con sujetos CMV+. El p-valor se eligió en 10^{-4}, pero tal como se ve, a un p-valor más bajo, se pueden identificar un mayor número de secuencias de TCRB asociados al CMV.

Tabla 1: Se incluyó cada una de las 142 secuencias de TCRβ asociados significativamente con el estado de CMV (p≤1x10-4, TDF ~ 20%) en la cohorte. Las secuencias se definieron utilizando la secuencia de aminoácidos de la región CDR3 junto con los segmentos génicos V y J. Se proporciona el número de sujetos CMV+ y CMV- en los que se observó cada secuencia, así como el p-valor de una prueba exacta de Fisher para la asociación con el estado de CMV, y cualquier alelo HLA-A o HLA-B que estuviera asociado significativamente ($p \leq 1x10^{-3}$) con la presencia de cada secuencia.

Tabla 1. Ejemplos de secuencias de TCRβ asociadas al CMV

SEQ ID NO:	Secuencia de AA	Gen V	Gen J	CMV+/CMV-	p-valor de Fisher	Asociación(es) de HLA
1	CASSGQGAYEQYF	TCRBV09-01	TCRBJ02-07*03	60/11	3,3142E-13	A29,B13
2	CASSPDRVGQETQYF	TCRBV05-01*01	TCRBJ02-05*03	34/1	9,5395E-12	Ninguna
3	CASSIGPLEHNEQFF	TCRBV19-01	TCRBJ02-01*03	30/0	1,4752E-11	A1,B8
4	CASSIEGNQPQHF	TCRBV28-01*01	TCRBJ01-05*03	27/0	1,9463E-10	Ninguna
5	CASSLVAGGRETQYF	TCRBV05-06*01	TCRBJ02-05*03	51/11	2,3538E-10	B8
6	CASSLEAEYEQYF	TCRBV07-02*01	TCRBJ02-07*03	30/1	2,7085E-10	B8
7	CATSDGDEQFF	TCRBV24	TCRBJ02-01*03	41/6	4,7339E-10	A1
8	CASSLAPGATNEKLFF	TCRBV07-06*01	TCRBJ01-04*03	36/4	9,134E-10	A2
9	CASSRGRQETQYF	TCRBV07-06*01	TCRBJ02-05*03	37/5	2,1095E-09	Ninguna
10	CASSAGQGVTYEQYF	TCRBV09-01	TCRBJ02-07*03	23/0	5,8832E-09	A1,B8
11	CASSLRREKLFF	TCRBV05-06*01	TCRBJ01-04*03	28/2	1,2553E-08	A29,B50
12	CASSLIGVSSYNEQFF	TCRBV07-09	TCRBJ02-01*03	27/2	2,7588E-08	A3,B7
13	CASSFFTSGQETQYF	TCRBV07-09	TCRBJ02-05*03	27/2	2,7588E-08	A24
14	CASSPQRNTEAFF	TCRBV04-03*01	TCRBJ01-01*03	27/2	2,7588E-08	A3,B7
15	CSVRDNFNQPQHF	TCRBV29-01*01	TCRBJ01-05*03	21/0	3,1927E-08	Ninguna
16	CATSRDSQGSYGYTF	TCRBV15-01*01	TCRBJ01-02*03	21/0	3,1927E-08	Ninguna
17	CASSPGDEQYF	TCRBV25-01*01	TCRBJ02-07*03	24/1	3,7138E-08	B7
18	CASSQTGGRNQPQHF	TCRBV12	TCRBJ01-05*03	26/2	6,0357E-08	A2
19	CASSQNRGQETQYF	TCRBV14-01*01	TCRBJ02-05*03	26/2	6,0357E-08	Ninguna
20	CASSLVIGGDTEAFF	TCRBV05-01*01	TCRBJ01-01*03	26/2	6,0357E-08	B8
21	CASSFHGFNQPQHF	TCRBV05-06*01	TCRBJ01-05*03	20/0	7,4139E-08	A1,B8
22	CASSRLAGGTDTQYF	TCRBV07-03*01	TCRBJ02-03*03	50/16	7,5065E-08	A1,B8
23	CASSLPSGLTDTQYF	TCRBV28-01*01	TCRBJ02-03*03	25/2	1,3144E-07	Ninguna
24	CATSRDTQGSYGYTF	TCRBV15-01*01	TCRBJ01-02*03	19/0	1,7179E-07	Ninguna
25	CATSDGDTQYF	TCRBV24	TCRBJ02-03*03	51/18	2,2203E-07	A1,B8
26	CASSLVASGRETQYF	TCRBV05-06*01	TCRBJ02-05*03	24/2	2,8483E-07	Ninguna
27	CASSIWGLDTEAFF	TCRBV19-01	TCRBJ01-01*03	28/4	3,6207E-07	Ninguna
28	CASSPGDEQFF	TCRBV25-01*01	TCRBJ02-01*03	28/4	3,6207E-07	B7
29	CASSPSTGTEAFF	TCRBV05-06*01	TCRBJ01-01*03	18/0	3,9723E-07	Ninguna
30	CSVEEDEGIYGYTF	TCRBV29-01*01	TCRBJ01-02*03	18/0	3,9723E-07	Ninguna
31	CASSEIPNTEAFF	TCRBV06-04	TCRBJ01-01*03	18/0	3,9723E-07	A24
32	CASSQVPGQGDNEQFF	TCRBV14-01*01	TCRBJ02-01*03	21/1	4,1428E-07	A1,B8
33	CASSPAGLNTEAFF	TCRBV19-01	TCRBJ01-01*03	21/1	4,1428E-07	Ninguna
34	CASSLGLKGTQYF	TCRBV12	TCRBJ02-05*03	21/1	4,1428E-07	Ninguna
35	CASSGDRLYEQYF	TCRBV02-01*01	TCRBJ02-07*03	23/2	6,1415E-07	Ninguna
36	CSVRDNYNQPQHF	TCRBV29-01*01	TCRBJ01-05*03	23/2	6,1415E-07	Ninguna
37	CASSYGGLGSYEQYF	TCRBV06-05*01	TCRBJ02-07*03	23/2	6,1415E-07	Ninguna

(continuación)

SEQ ID NO:	Secuencia de AA	Gen V	Gen J	CMV+/CMV-	p-valor de Fisher	Asociación(es) de HLA
38	CASNRDRGRYEQYF	TCRBV06-01*01	TCRBJ02-07*03	20/1	9,1836E-07	B45
39	CASMGGASYEQYF	TCRBV27-01*01	TCRBJ02-07*03	20/1	9,1836E-07	A2
40	CASSLGVGPYNEQFF	TCRBV07-02*01	TCRBJ02-01*03	20/1	9.1836E-07	B7
41	CASSLGGAGDTQYF	TCRBV12	TCRBJ02-03*03	44/15	1,1802E-06	Ninguna
42	CATSRGTVSYEQYF	TCRBV15-01*01	TCRBJ02-07*03	30/6	1,2303E-06	Ninguna
43	CATSDGETQYF	TCRBV24	TCRBJ02-05*03	70/36	1,3031E-06	A1,B52
44	CASSEARGGVEKLFF	TCRBV06-01*01	TCRBJ01-04*03	22/2	1,3173E-06	Ninguna
45	CASSLNRGQETQYF	TCRBV14-01*01	TCRBJ02-05*03	19/1	2,0272E-06	Ninguna
46	CSVRDNHNQPQHF	TCRBV29-01*01	TCRBJ01-05*03	19/1	2,0272E-06	Ninguna
47	CASSESGHRNQPQHF	TCRBV10-02*01	TCRBJ01-05*03	16/0	2,1105E-06	Ninguna
48	CSASPGQGASYGYTF	TCRBV20	TCRBJ01-02*03	16/0	2,1105E-06	Ninguna
49	CASSEARTRAFF	TCRBV06-01*01	TCRBJ01-01*03	16/0	2,1105E-06	Ninguna
50	CASRPTGYEQYF	TCRBV06-01*01	TCRBJ02-07*03	21/2	2,8099E-06	B39
51	CASSVTGGTDTQYF	TCRBV09-01	TCRBJ02-03*03	74/41	2,8214E-06	A1,B8
52	CASSRLAASTDTQYF	TCRBV07-03*01	TCRBJ02-03*03	23/3	3,1389E-06	A1,B8
53	CATSDSVTNTGELFF	TCRBV24	TCRBJ02-02*03	18/1	4,4551E-06	B8
54	CASSRNRESNQPQHF	TCRBV06-05*01	TCRBJ01-05*03	18/1	4,4551E-06	Ninguna
55	CASSAQGAYEQYF	TCRBV09-01	TCRBJ02-07*03	28/6	4,717E-06	Ninguna
56	CASSIQGYSNQPQHF	TCRBV05-08*01	TCRBJ01-05*03	15/0	4,8494E-06	B8
57	CASSYNPYSNQPQHF	TCRBV06-06	TCRBJ01-05*03	15/0	4,8494E-06	Ninguna
58	CASSLGHRDPNTGELFF	TCRBV05-01*01	TCRBJ02-02*03	15/0	4,8494E-06	B13
59	CASSITGGDGYTF	TCRBV19-01	TCRBJ01-02*03	26/5	5,5727E-06	Ninguna
60	CASSVLAGPTDTQYF	TCRBV09-01	TCRBJ02-03*03	26/5	5,5727E-06	A1
61	CASSYRQETQYF	TCRBV06-05*01	TCRBJ02-05*03	20/2	5,9592E-06	Ninguna
62	CASSSGQVYGYTF	TCRBV05-06*01	TCRBJ01-02*03	22/3	6,4736E-06	A1,B8
63	CASGRDTYEQYF	TCRBV02-01*01	TCRBJ02-07*03	17/1	9,7453E-06	Ninguna
64	CATSDSRTGGQETQYF	TCRBV24	TCRBJ02-05*03	17/1	9,7453E-06	B8
65	CASSSPGRSGANVLTF	TCRBV28-01*01	TCRBJ02-06*03	17/1	9,7453E-06	Ninguna
66	CASSYGGEGYTF	TCRBV06-05*01	TCRBJ01-02*03	36/12	1,059E-05	Ninguna
67	CASSLAGVDYEQYF	TCRBV07-09	TCRBJ02-07*03	25/5	1,0951E-05	B8
68	CASSLQGADTQYF	TCRBV07-08*01	TCRBJ02-03*03	14/0	1,112E-05	Ninguna
69	CASSLEAENEQFF	TCRBV07-02*01	TCRBJ02-01*03	14/0	1,112E-05	A11,B8
70	CASSEAPSTSTDTQYF	TCRBV02-01*01	TCRBJ02-03*03	14/0	1,112E-05	Ninguna
71	CASSLEGQQPQHF	TCRBV28-01*01	TCRBJ01-05*03	14/0	1,112E-05	Ninguna
72	CASSLGHRDSSYEQYF	TCRBV05-01*01	TCRBJ02-07*03	14/0	1,112E-05	A29,B57
73	CASSPSRNTEAFF	TCRBV04-03*01	TCRBJ01-01*03	19/2	1,2561E-05	A3,B7
74	CSALGHSNQPQHF	TCRBV20	TCRBJ01-05*03	19/2	1,2561E-05	Ninguna
75	CASSHRDRNYEQYF	TCRBV07-09	TCRBJ02-07*03	19/2	1,2561E-05	A24
76	CASSPPGQGSDTQYF	TCRBV18-01*01	TCRBJ02-03*03	28/7	1,3913E-05	Ninguna
77	CASSLQGYSNQPQHF	TCRBV05-08*01	TCRBJ01-05*03	34/11	1,4711E-05	B8
78	CASSYVRTGGNYGYTF	TCRBV06-05*01	TCRBJ01-02*03	31/9	1,5052E-05	Ninguna
79	CASSRDRNYGYTF	TCRBV06-04	TCRBJ01-02*03	29/8	2,0024E-05	Ninguna
80	CASSTGTSGSYEQYF	TCRBV06-01*01	TCRBJ02-07*03	16/1	2,1213E-05	B54
81	CASRSDSGANVLTF	TCRBV06-04	TCRBJ02-06*03	16/1	2,1213E-05	Ninguna
82	CATSRVAGETQYF	TCRBV15-01*01	TCRBJ02-05*03	24/5	2,1345E-05	B7
83	CASSEEGIQPQHF	TCRBV02-01*01	TCRBJ01-05*03	22/4	2,4814E-05	Ninguna
84	CASSLGGPGDTQYF	TCRBV12	TCRBJ02-03*03	22/4	2,4814E-05	Ninguna
85	CASSLVAAGRETQYF	TCRBV05-06*01	TCRBJ02-05*03	13/0	2,5446E-05	B8
86	CASRGQGWDEKLFF	TCRBV06-05*01	TCRBJ01-04*03	13/0	2,5446E-05	Ninguna
87	CASSLEGQGFGYTF	TCRBV05-01*01	TCRBJ01-02*03	13/0	2,5446E-05	Ninguna
88	CASRDWDYTDTQYF	TCRBV02-01*01	TCRBJ02-03*03	13/0	2,5446E-05	Ninguna
89	CASSRSGLAGNTGELFF	TCRBV06	TCRBJ02-02*03	13/0	2,5446E-05	Ninguna
90	CASSFGQEAGANVLTF	TCRBV05-01*01	TCRBJ02-06*03	13/0	2,5446E-05	Ninguna
91	CASSLGDRPDTQYF	TCRBV11-02*02	TCRBJ02-03*03	13/0	2,5446E-05	Ninguna
92	CASSFPGGETQYF	TCRBV11-01*01	TCRBJ02-05*03	18/2	2,6306E-05	Ninguna
93	CASSLETYGYTF	TCRBV05-06*01	TCRBJ01-02*03	18/2	2,6306E-05	B78
94	CASSSGQVQETQYF	TCRBV11-02*02	TCRBJ02-05*03	18/2	2,6306E-05	B51
95	CASSFDNYGYTF	TCRBV05-04*01	TCRBJ01-02*03	18/2	2,6306E-05	Ninguna

(continuación)

SEQ ID NO:	Secuencia de AA	Gen V	Gen J	CMV+/CMV-	p-valor de Fisher	Asociación(es) de HLA
96	CASSEGARQPQHF	TCRBV10-02*01	TCRBJ01-05*03	18/2	2,6306E-05	Ninguna
97	CASSLTGGRNQPQHF	TCRBV12	TCRBJ01-05*03	33/11	2,6324E-05	A2
98	CASSLLWDQPQHF	TCRBV05-05*01	TCRBJ01-05*03	20/3	2,6954E-05	Ninguna
99	CASSLFGTGGNTEAFF	TCRBV05-06*01	TCRBJ01-01*03	20/3	2,6954E-05	B44
100	CASSISAGEAFF	TCRBV19-01	TCRBJ01-01*03	20/3	2,6954E-05	Ninguna
101	CASSPPSGLTDTQYF	TCRBV28-01*01	TCRBJ02-03*03	20/3	2,6954E-05	Ninguna
102	CASSPLSDTQYF	TCRBV07-09	TCRBJ02-03*03	30/9	2,752E-05	Ninguna
103	CASSSRGTGELFF	TCRBV28-01*01	TCRBJ02-02*03	25/6	3,3358E-05	Ninguna
104	CASSYAGDGYTF	TCRBV06-05*01	TCRBJ01-02*03	31/10	3,6414E-05	B8
105	CASSDRGNTGELFF	TCRBV04-01*01	TCRBJ02-02*03	31/10	3,6414E-05	A68,B41
106	CASSPGGTQYF	TCRBV12	TCRBJ02-05*03	28/8	3,6921E-05	Ninguna
107	CASSLGDRAYNEQFF	TCRBV05-06*01	TCRBJ02-01*03	28/8	3,6921E-05	Ninguna
108	CASSLRGSSYNEQFF	TCRBV05-08*01	TCRBJ02-01*03	23/5	4,1254E-05	Ninguna
109	CASSLTASSYEQYF	TCRBV05-01*01	TCRBJ02-07*03	23/5	4,1254E-05	Ninguna
110	CSASDHEQYF	TCRBV20-01*01	TCRBJ02-07*03	23/5	4,1254E-05	Ninguna
111	CASSQGRHTDTQYF	TCRBV14-01*01	TCRBJ02-03*03	15/1	4,5934E-05	A68
112	CASSRPGQGNTEAFF	TCRBV12	TCRBJ01-01*03	26/7	4,8714E-05	Ninguna
113	CASSLVGDGYTF	TCRBV07-08*01	TCRBJ01-02*03	36/14	4,9771E-05	Ninguna
114	CASSLGAGNQPQHF	TCRBV28-01*01	TCRBJ01-05*03	29/9	4,9861E-05	Ninguna
115	CASSLTDTGELFF	TCRBV11-02*02	TCRBJ02-02*03	29/9	4,9861E-05	Ninguna
116	CASSLTGGNSGNTIYF	TCRBV07-02*01	TCRBJ01-03*03	19/3	5,4373E-05	B14
117	CAWRGTGNSPLHF	TCRBV30-01*01	TCRBJ01-06*03	17/2	5,4712E-05	Ninguna
118	CASASANYGYTF	TCRBV12	TCRBJ01-02*03	12/0	5,8109E-05	A2
119	CASSLQAGANEQFF	TCRBV07-02*01	TCRBJ02-01*03	12/0	5,8109E-05	Ninguna
120	CASSEEAGGSGYTF	TCRBV06-01*01	TCRBJ01-02*03	12/0	5,8109E-05	Ninguna
121	CASRTGESGYTF	TCRBV06-05*01	TCRBJ01-02*03	12/0	5,8109E-05	Ninguna
122	CASSGLNEQFF	TCRBV06-01*01	TCRBJ02-01*03	12/0	5,8109E-05	Ninguna
123	CASSWDRDNSPLHF	TCRBV25-01*01	TCRBJ01-06*03	12/0	5,8109E-05	B7
124	CASSIRTNYYGYTF	TCRBV19-01	TCRBJ01-02*03	12/0	5,8109E-05	Ninguna
125	CSARSPEAFF	TCRBV20-01*01	TCRBJ01-01*03	12/0	5,8109E-05	Ninguna
126	CASSRGTGATDTQYF	TCRBV19-01	TCRBJ02-03*03	12/0	5,8109E-05	Ninguna
127	CASSPRVSNQPQHF	TCRBV12	TCRBJ01-05*03	12/0	5,8109E-05	Ninguna
128	CAISESQDRGHEQYF	TCRBV10-03*01	TCRBJ02-07*03	12/0	5,8109E-05	Ninguna
129	CASSLGRGYEKLFF	TCRBV05-06*01	TCRBJ01-04*03	12/0	5,8109E-05	B62
130	CSVEVRGTDTQYF	TCRBV29-01*01	TCRBJ02-03*03	12/0	5,8109E-05	Ninguna
131	CASRGQGAGELFF	TCRBV02-01*01	TCRBJ02-02*03	30/10	6,4924E-05	Ninguna
132	CATSREGSGYEQYF	TCRBV15-01*01	TCRBJ02-07*03	22/5	7,9023E-05	Ninguna
133	CASSLGWTEAFF	TCRBV05-01*01	TCRBJ01-01*03	22/5	7,9023E-05	Ninguna
134	CASSLGSSSYNEQFF	TCRBV07-02*01	TCRBJ02-01*03	35/14	8,4988E-05	Ninguna
135	CASSSAGADTQYF	TCRBV07-09	TCRBJ02-03*03	14/1	9,8907E-05	B5
136	CASSERKNYGYTF	TCRBV06-01*01	TCRBJ01-02*03	14/1	9,8907E-05	Ninguna
137	CASRDRDRVNTEAFF	TCRBV06-01*01	TCRBJ01-01*03	14/1	9,8907E-05	A3
138	CASSRVGEQFF	TCRBV03	TCRBJ02-01*03	14/1	9,8907E-05	Ninguna
139	CASSPRWQETQYF	TCRBV27-01*01	TCRBJ02-05*03	14/1	9,8907E-05	B14
140	CASTPGDTIYF	TCRBV25-01*01	TCRBJ01-03*03	14/1	9,8907E-05	B55
141	CASSENGGNQPQHF	TCRBV06-01*01	TCRBJ01-05*03	14/1	9,8907E-05	Ninguna
142	CASSYPGETQYF	TCRBV06-06	TCRBJ02-05*03	36/15	9,991E-05	Ninguna

Reducción de dimensionalidad y aprendizaje automático:

Se calculó la carga del CMV para cada sujeto. La carga del CMV se define como la fracción de los TCRβ únicos de un sujeto que están significativamente asociados al CMV, tal como se muestra en la FIG. 6. El gráfico de la FIG. 6 muestra la distribución de las puntuaciones del CMV (es decir, la proporción del repertorio de TCRβ de cada sujeto que coincide con el listado de 142 secuencias de TCRβ asociadas con el CMV) entre sujetos CMV+ y CMV-.

Esto permitió el entrenamiento rápido de un clasificador de regresión logística unidimensional del estado del CMV. Se

realizó una validación cruzada exhaustiva de dejar uno fuera (incluido el recálculo de los clones asociados con el CMV), que indica una alta precisión en una amplia gama de umbrales de p-valor, con un rendimiento degradado a una TDF alta.

Cada sujeto se eliminó del conjunto de datos a su vez, y la lista de TCRβ significativamente asociados con el CMV se volvió a calcular utilizando los sujetos restantes con estado conocido de CMV. A continuación, se calculó la carga del CMV de cada sujeto tal como se describe anteriormente. A continuación, se entrenó un clasificador de regresión logística unidimensional sobre la carga de CMV frente al estado del CMV, y la carga de CMV del sujeto que se mantuvo fuera se calculó y se introdujo en la regresión logística. El resultado de este modelo de regresión logística fue la probabilidad de que el sujeto de estado desconocido fuera CMV+, que luego se comparó con el estado (conocido pero no potenciado) del sujeto excluido para determinar la precisión de la clasificación de exclusión. La clasificación de un sujeto con un estado de CMV realmente desconocido procede de la misma manera: secuenciar el TCRβ, calcular una carga del CMV (como proporción de TCRβ únicos de este sujeto presentes en el listado de TCRβ asociados con CMV) e ingresar esta carga de CMV en un modelo de regresión logística entrenado en sujetos de estado conocido, con una probabilidad estimada de positividad para CMV como salida.

Tal como se muestra en la FIG. 7, el método de validación cruzada dejó fuera una de las 640 muestras iniciales, luego, la base de datos de secuencias de TCRβ específicas del CMV y las estadísticas asociadas se volvieron a entrenar de una manera no supervisada para eliminar el sesgo, y se clasificó el estado serológico del CMV de la muestra excluida.

Los resultados de validación cruzada se muestran en las FIG. 8 y 9 (mostrados para todos los sujetos y sujetos con validación cruzada (VC)). Se demostró que la precisión del modelo es del 89 %, correspondiente a una proporción de posibilidades de diagnóstico de ~ 66. Al evaluar el rendimiento del modelo en el rango de especificidad frente a sensibilidad utilizando una curva COR, se logró un área bajo la curva COR del 93 %.

La FIG. 8 (gráfico superior) muestra datos para el rendimiento de clasificación de todos y los conjuntos de datos de validación cruzada (VC) para cada umbral de p-valor, medido como el área bajo la curva COR (ABCOR). El número sobre cada conjunto de puntos de datos se corresponde con el número de TCRβ asociados al CMV identificados en ese umbral de p-valor, y el rectángulo indica el conjunto de datos seleccionado para el análisis posterior (p-valor = 10^{-4}).

La FIG. 8 (gráfico inferior) también muestra una tasa de descubrimientos falsos (TDF) estimada para cada umbral de p-valor utilizado en la identificación de secuencias de TCRβ asociadas significativamente al CMV, utilizando permutaciones del estado del CMV. El mejor rendimiento se observa con un p-valor de 10^{-4}, que se corresponde con una TDF estimada de ~ 20 %, dando como resultado la identificación de un conjunto de 142 secuencias de TCRβ que se asociaron significativamente con el estado positivo del CMV (enumeradas en la Tabla 1). El uso de estas condiciones dio como resultado una buena separación entre los sujetos CMV+ y CMV- en la cohorte medida por la puntuación del CMV (véase la FIG. 6).

La FIG. 9 muestra las curvas COR para los conjuntos de datos totales y de validación cruzada. El ABCOR para el conjunto de datos completo es 0,98, lo que indica que el enfoque de los presentes inventores dio como resultado un clasificador excelente para el estado del CMV. En el punto de mayor poder de discriminación, se observó una precisión de 0,89 y una proporción de probabilidades de diagnóstico de 66 en el conjunto de datos de validación cruzada (logrado al clasificar correctamente el 86 % de los verdaderos positivos con una tasa de falsos positivos del 8 %). En conjunto, estos datos sugieren que la presencia de respuestas de linfocitos T públicos al CMV está altamente correlacionada con el estado positivo del CMV.

Dado que los linfocitos T reconocen sus antígenos afines en el entorno de moléculas del MHC expresadas por células presentadoras de antígenos, se probó si se podía identificar la restricción de HLA de las secuencias de los presentes inventores del TCRβ asociadas al CMV. Se realizó una prueba exacta de Fisher en cada secuencia de los TCRβ asociados al CMV para determinar si su presencia estaba significativamente asociada con alguno de los alelos HLA observados en la cohorte. La asociación de 57 de las 142 secuencias de TCRβ asociados al CMV se podría asignar con confianza con al menos un alelo de HLA, con un p-valor de corte de $1x10^{-3}$. Los resultados completos se presentan en la Tabla 1 (anteriormente) y en la FIG. 10.

Las FIG. 10A y 10B muestran restricción de HLA de secuencias de TCRβ asociados al CMV. La FIG. 10A muestra la distribución de los alelos HLA-A en esta cohorte. La FIG. 10B muestra la distribución de los alelos HLA-B en esta cohorte. Se probó cada una de las 142 secuencias de los TCRβ asociados al CMV identificados en $p \leq 1x10^{-4}$ para una asociación significativa con cada alelo de HLA, con un umbral de p-valor de $1x10^{-3}$. De estas 142 secuencias de los TCRβ asociadas al CMV, 57 se asociaron significativamente con un alelo HLA-A y/o HLA-B, y ninguna secuencia se asoció significativamente con más de un alelo de cada locus. Los recuadros de colores y las secuencias de aminoácidos indican las 5 secuencias de los TCRβ identificadas en el estudio de los presentes inventores que habían sido previamente identificadas. En 4 de los casos, se recapituló la asociación correcta de HLA, y en el 5º caso, no hubo asociación de HLA estadísticamente significativa.

Se realizó una búsqueda en la bibliografía para identificar secuencias de los TCRβ reactivos al CMV previamente

documentadas, y se seleccionaron 595 secuencias de TCRβ únicos que habían sido identificadas por al menos un estudio previamente publicado[3, 10, 12-27]. De éstas, 30 secuencias únicas de los TCRβ habían sido previamente clasificadas como públicas o se documentaron en múltiples estudios. Se determinó que 5 de estas 30 secuencias de los TCRβ públicos únicas estaban contenidas en el conjunto de 142 secuencias de los TCRβ asociados al CMV. Además, se identificó una asociación de HLA para 4 de estas 5 secuencias de los TCRβ, y en los 4 casos, la asociación de HLA de los presentes inventores estuvo de acuerdo con la bibliografía. Las FIG. 11 A, 11B y 12 proporcionan información más completa sobre la prevalencia en el conjunto de datos de los presentes inventores de secuencias de TCRβ reactivos al CMV previamente identificadas, tal como se describe a continuación.

Las FIG. 11A y 11B muestran la incidencia de secuencias de los TCRβ reactivos al CMV previamente documentados en esta cohorte. Después de una búsqueda de la bibliografía, se identificaron 565 secuencias de TCRβ reactivos al CMV 'privadas' (documentadas en un individuo de un solo estudio) y 30 secuencias de TCRβ reactivos al CMV 'públicas' (documentadas en múltiples estudios o en múltiples individuos dentro de un solo estudio). La FIG. 11A muestra la incidencia de cada secuencia de los TCRβ en la cohorte de 640 sujetos representada gráficamente a lo largo del eje horizontal mediante la disminución de la incidencia total, con la incidencia en sujetos CMV+ por encima de la horizontal y la incidencia en sujetos CMV- por debajo de la horizontal. Se observaron muchas secuencias documentadas anteriormente en el conjunto de datos, pero la mayoría se observó en aproximadamente el mismo número de sujetos CMV+ y CMV-, lo que podría explicarse por secuencias de receptores con una frecuencia excepcionalmente alta en el repertorio sin exposición previa, o podría reflejar receptores de reacción cruzada que se unen a antígenos del CMV, pero también a otros antígenos comunes. La FIG. 11B muestra un histograma de incidencia de estas secuencias de los TCRβ en la cohorte de 640 sujetos representados gráficamente para cada grupo de secuencias. La mayoría de las secuencias del TCRβ reactivos al CMV documentados anteriormente se encontraron en el conjunto de datos de los presentes inventores en niveles apreciables, aunque solo un puñado se encontró de manera desproporcionada en sujetos CMV+ en $p \leq 1x10^{-4}$. Las secuencias de TrRβ "públicos" descritas en la bibliografía fueron considerablemente más comunes en la cohorte de los presentes inventores.

La FIG. 12 muestra la concordancia de las secuencias de los TCRB en la cohorte en comparación con las de la bibliografía. De las 142 secuencias de TCRβ asociadas significativamente con el estado de CMV en la cohorte con un p-valor de menos de $1x10^{-4}$, cinco secuencias de TCRβ (definidas por la coincidencia de la secuencia de aminoácidos de CDR3) se han documentado previamente en la bibliografía como clones públicos. La FIG. 12 proporciona la secuencia de aminoácidos de CDR3, los genes V y J, y la asociación de HLA para cada secuencia y la compara con informes anteriores. Tal como se esperaba, las dos secuencias previamente identificadas como públicas se vieron en más sujetos en la cohorte de los presentes inventores que las tres secuencias previamente documentadas solo una vez. La concordancia del gen V, del gen J y el alelo HLA con las descritas en la bibliografía fue muy buena. Cinco de las nueve comparaciones totales tenían el mismo gen V y las otras 4 comparaciones tenían la misma subfamilia del gen V. Había genes J concordantes en 9 de 9 comparaciones, y una asociación de HLA idéntica en 8 de 9 comparaciones, con una secuencia no significativamente asociada a HLA en este estudio.

Por tanto, la inmunosecuenciación se puede usar para determinar secuencias de linfocitos T específicos de CMV en una gran cohorte y predecir el estado del CMV en nuevos sujetos con alta precisión. En resumen, se ha demostrado que la información obtenida de los receptores de linfocitos T reordenados se puede utilizar para inferir el estado de la enfermedad basándose en la presencia de respuestas de linfocitos T públicos; el único requisito es una muestra grande de muestras positivas y negativas para patógenos con las que identificar estas respuestas de linfocitos T públicos. Debido a que la secuenciación de alto rendimiento de los receptores de linfocitos T captura todas las respuestas de los linfocitos T por igual, y estas almacenan la memoria inmunológica para todos los patógenos en un formato común, la lectura de la memoria de los linfocitos T buscando respuestas públicas conocidas puede ser una estrategia viable para diagnosticar simultáneamente una amplia gama de agentes infecciosos utilizando una sola muestra de sangre periférica y un ensayo simple unificado. Estos métodos se pueden aplicar de forma necesaria a las infecciones agudas, dado que la memoria de los linfocitos T persiste durante años, lo que explica el hecho de que no se sabe cómo los clones públicos decaen con el tiempo después de una infección aguda. Por consiguiente, el presente método se puede utilizar para evaluar múltiples infecciones simultáneamente, tal como HPV, EBV, CMV y otras. El método también se puede usar para predecir o diagnosticar una enfermedad no infecciosa (por ejemplo, autoinmune).

Los métodos de la invención también se pueden usar para cuantificar la carga de la enfermedad (por ejemplo, reactivación del CMV después del trasplante).

Ejemplo 2: Identificación del tipo de HLA

El primer objetivo del estudio fue crear un catálogo completo de secuencias de receptores de linfocitos T y tipos de HLA asociados.

Fenotipado de HLA:

640 sujetos fueron fenotipados para el tipo HLA de acuerdo con el protocolo estándar del Fred Hutchinson Cancer Research Center.

Secuenciación del repertorio inmunitario:

A continuación, las secuencias de TCRB para cada sujeto se determinaron mediante amplificación y secuenciación.

El ADN genómico se extrajo de muestras de sangre periférica utilizando el kit de extracción de sangre Qiagen DNeasy (Qiagen, Gaithersburg, MD, EE. UU.). Se secuenciaron las regiones de CDR3 de los genes de TCRβ reordenados; la región CDR3 de TCRβ se definió de acuerdo con la colaboración del IMGT[28, 20]. Las regiones de CDR3 de TCRβ se amplificaron y secuenciaron utilizando los métodos descritos anteriormente y en los protocolos descritos anteriormente[5, 29]. El método de PCR multiplexada utilizó una mezcla de 60 cebadores directos específicos para los segmentos del gen TCR Vβ y 13 cebadores inversos específicos para los segmentos del gen TCR Jβ. Los amplicones resultantes se secuenciaron utilizando los métodos descritos anteriormente. Se obtuvieron lecturas de 87 pb utilizando el sistema Illumina HiSeq. Los datos de secuencia de HiSeq sin procesar se preprocesaron para eliminar errores en la secuencia principal de cada lectura y para comprimir los datos. Se utilizó un algoritmo del vecino más cercano para contraer los datos en secuencias únicas fusionando secuencias estrechamente relacionadas, para eliminar errores de secuenciación y de PCR.

Para garantizar una cobertura adecuada de cada reordenamiento de linfocitos T, la cobertura de secuencia de 8-10x, o ~ 6-10 millones de lecturas de secuenciación por muestra se generó utilizando aproximadamente ocho ejecuciones de secuenciación completas. Todas las lecturas de secuenciación se procesaron utilizando un modelo bioinformático estandarizado en desarrollo para 1) demultiplexar lecturas para muestras específicas, 2) eliminar secuencias de baja calidad y eliminar posibles contaminantes, 3) alinear e identificar segmentos de genes específicos de las regiones CDR3 de V y J de TCRβ, 4) agrupar altamente secuencias similares para tener en cuenta la PCR y los errores de secuenciación, 5) normalizar los datos para eliminar el sesgo de amplificación de la PCR, 6) estimar la entrada total de linfocitos T y 7) generar recuentos y distribuciones de secuencias únicas de TCRβ.

En el presente documento, se secuenciaron aproximadamente 250.000 genes de receptores de linfocitos T reordenados a partir de sangre periférica de cada sujeto. Para cada sujeto, se obtuvieron aproximadamente 10^5 secuencias de TCRB y se cuantificaron las abundancias de cada secuencia única.

Selección de rasgos: Para definir la asociación entre un alelo HLA y secuencias de TCRB

Para cada sujeto, se determina el número total de secuencias de TCRB único y la frecuencia de cada secuencia de TCRB único. Para una secuencia de TCRB único, se determina cuántos sujetos que son positivos para un alelo HLA tienen la secuencia de TCRB y cuántos sujetos que son negativos para el alelo HLA tienen la secuencia de TCRB. Además, se puede determinar el número de sujetos que son positivos para un alelo HLA y negativos para la secuencia de TCRB y el número de sujetos que son negativos para el alelo HLA y positivos para la secuencia de TCRB. La siguiente tabla muestra la categorización de los sujetos según la presencia o ausencia de una secuencia de TCRB y la presencia o ausencia de un alelo de HLA, HLA-A2.

	HLA-A2+	**HLA-A2-**
secuencia *i* de TCRβ presente	n_{i+}	n_{i-}
secuencia *i* de TCRβ no presente	$N_+ - n_{i+}$	$N_- - n_{i-}$

La FIG. 14 muestra un listado de secuencias de TCRB únicos ilustrativos y el número de sujetos que son positivos o negativos para un alelo HLA-A2 que tienen una secuencia de TCRB particular.

Se determina un p-valor para la asociación de cada secuencia de TCRB con un estado de HLA utilizando una prueba exacta de Fisher (de dos colas). El p-valor para la asociación de cada TCR con el estado del alelo usando una prueba exacta de Fisher (dos colas) se calcula de la siguiente manera:

$$\Pr(\text{tabla}) = \frac{\binom{n_+ + n_-}{n_+}\binom{N_+ + N_- - n_+ - n_-}{n_+ + n_-}}{\binom{N_+ + N_-}{N_+}}$$

La FIG. 14 muestra p-valores ilustrativos que se calcularon para la asociación de una secuencia de TCRB particular con un tipo de HLA (HLA-A2).

Tal como se muestra en la FIG. 15, se selecciona un p-valor como un punto de corte para identificar un conjunto de "rasgos de los TCR" del listado completo de las posibles secuencias de los TCR. La definición de un umbral de p-valor y la permutación del estado del alelo entre individuos proporciona una estimación de la tasa de descubrimientos falsos.

Esto se realiza para cada alelo HLA, dando como resultado un conjunto de secuencias de TCRβ asociadas a los alelos para cada alelo HLA. En la FIG. 15, se usa un valor de corte del p-valor de $p \leq 10^{-4}$ y una TDF de 0,1 para identificar 288 secuencias de TCRB que están asociadas positivamente con HLA-A2. Para cada una de las secuencias de TCRB asociadas a los alelos, la frecuencia de la secuencia también se determina en cada sujeto.

Tal como se muestra en la FIG. 15, se selecciona un p-valor como un punto de corte para identificar un conjunto de "rasgos de los TCR" del listado completo de las posibles secuencias de los TCR. La definición de un umbral de p-valor y la permutación del estado del alelo entre individuos proporciona una estimación de la tasa de descubrimientos falsos. Esto se realiza para cada alelo HLA, dando como resultado un conjunto de secuencias de TCRβ asociadas a los alelos para cada alelo HLA. En la FIG. 15, se usa un valor de corte del p-valor de $p \leq 10^{-4}$ y una TDF de 0,1 para identificar 288 secuencias de TCRB que están asociadas positivamente con HLA-A2. Para cada una de las secuencias de TCRB asociadas a los alelos, la frecuencia de la secuencia también se determina en cada sujeto.

La tasa de descubrimientos falsos (TDF) se determinó mediante la permutación del estado del alelo. Véase Storey *et al.* Statistical significance for genomewide studies. PNAS, 100(6), págs. 9440-9445.

La etapa de selección de rasgos va seguida de un proceso de aprendizaje automático. Tal como se muestra en la FIG. 16, para cada alelo HLA, se entrena un modelo de regresión logística utilizando el conjunto de vectores de características sobre todos los sujetos, junto con el estado conocido de la presencia de ese alelo.

Tal como se muestra en la FIG. 17, se realiza una validación cruzada exhaustiva de dejar uno fuera donde se elimina un sujeto del análisis, y el estado de HLA del sujeto se infiere basándose en la selección de rasgos y el entrenamiento de solo los sujetos restantes. El resultado es un conjunto de clasificadores (uno para cada alelo HLA) que estiman la probabilidad de un estado positivo para cada alelo HLA, tomando como entrada el vector de características para cada alelo.

Equipado con estos clasificadores, el tipo de HLA de un nuevo sujeto se puede evaluar mediante: 1. secuenciación del repertorio inmunitario, 2. calcular los vectores de rasgos para cada alelo, y 3. definir un umbral de probabilidad para el estado positivo que se aplicará a la salida de cada clasificador. De esta forma, se puede inferir un estado de HLA para un nuevo sujeto con un estado de HLA desconocido.

La FIG. 18 muestra los resultados de un estudio de validación cruzada. La validación cruzada de dejar uno fuera se realizó durante 78 rondas. Para el alelo HLA-A2, el método descrito anteriormente dio como resultado la identificación de 288 rasgos de TCR y una predicción precisa de 41 sujetos de 43 como positivos para HLA-A2. Como se ha descrito anteriormente, el tipo de HLA de los sujetos se conocía antes de realizar la validación cruzada. Solo 2/43 sujetos fueron falsos positivos para HLA-A2. Por tanto, el método tuvo una precisión del 96 % para predecir la presencia de HLA-A2 en un sujeto basándose en el perfil de secuencia de TCRB del sujeto.

La FIG. 18 también muestra los resultados de la validación cruzada del alelo HLA-A24, que también tuvo una precisión del 96 %. Se predijo con precisión que 10 de 13 sujetos poseían el alelo HLA-A24, y se predijo con precisión que 65 de 65 sujetos no tenían el alelo HLA-A24 según los rasgos identificados de los TCRB.

Ejemplo 3: Estudio de HLA

En otro ejemplo, la respuesta de los linfocitos T públicos al citomegalovirus (CMV) se investigó mediante la secuenciación de los receptores de linfocitos T (TCR) reordenados en 650 sujetos (294 con y 356 sin CMV). Se evaluó la concordancia entre ~ 90 millones de TCR únicos y el estado serológico del CMV, centrándose en la identificación de asociaciones significativas. En este estudio, se identificaron 157 TCR asociados al CMV a $p \leq 10^{-3}$, TDF ≈ 0,15). Entrenando un clasificador binario en estas rasgos, se predijo que el estado serológico del CMV en un procedimiento de validación cruzada de dejar uno fuera tenía una razón de probabilidades de diagnóstico de 44. El clasificador también se probó en una segunda cohorte independiente de 120 sujetos con estado serológico del CMV conocido, generando una razón de probabilidades de diagnóstico de 49.

A continuación, la restricción de HLA de cada TCR asociado al CMV se investigó evaluando la sobrerrepresentación de tipos de HLA particulares entre los sujetos que portan cada TCR asociado al CMV. De 157 TCR asociados a CMV, 61 estaban asociados a HLA a $p \leq 10^{-3}$. Ninguno de estos se asoció significativamente con múltiples alelos HLA-A o HLA-B.

Hubo una concordancia sustancial entre los datos de los presentes inventores y los TCR asociados con el CMV y HLA publicados anteriormente. La mayoría de los TCR específicos del CMV públicos documentados anteriormente se observaron en los datos, aunque solo se han documentado previamente 5/157 TCR asociados con el CMV identificados en este estudio. De éstos, 4 tenían una restricción significativa de HLA y los cuatro confirmaron hallazgos anteriores.

Además, se investigó la asociación de los TCR con cada alelo HLA-A presente en la cohorte, con resultados significativos ($p \leq 10^{-4}$) para muchos alelos de mayor frecuencia. El entrenamiento del clasificador binario dio como

resultado una predicción de alta precisión (-96%) para estos alelos, lo que indica que el tipo de HLA se puede inferir a partir de los datos de inmunosecuenciación.

En resumen, el presente estudio demostró la validez de los estudios de asociación que utilizan la inmunosecuenciación para la detección y la asociación HLA de las respuestas de los linfocitos T públicos a la infección, y mostró que evaluar la presencia de respuestas de los linfocitos T asociadas puede servir como un poderoso clasificador de diagnóstico.

REFERENCIAS

1. Neller, M.A., Burrows, J.M., Rist, M.J., Miles, J.J. & Burrows, S.R. High frequency of herpesvirus-specific clonotypes in the human T cell repertoire can remain stable over decades with minimal turnover. J Virol 87, 697-700 (2013).
2. Li, H., Ye, C., Ji, G. & Han, J. Determinants of public T cell responses. Cell Res 22, 33-42 (2012).
3. Venturi, V., Price, D.A., Douek, D.C. & Davenport, M.P. The molecular basis for public T-cell responses? Nat Rev Immunol 8, 231-238 (2008).
4. Arstila, T.P. *et al.* A direct estimate of the human alphabeta T cell receptor diversity. Science 286, 958-961 (1999).
5. Robins, H.S. *et al.* Comprehensive assessment of T-cell receptor beta-chain diversity in alphabeta T cells. Blood 114, 4099-4107 (2009).
6. Cabaniols, J.P., Fazilleau, N., Casrouge, A., Kourilsky, P. & Kanellopoulos, J.M. Most alpha/beta T cell receptor diversity is due to terminal deoxynucleotidyl transferase. J Exp Med 194, 1385-1390 (2001).
7. Davis, M.M. & Bjorkman, P.J. T-cell antigen receptor genes and T-cell recognition. Nature 334 (1988).
8. Robins, H.S. *et al.* Overlap and effective size of the human CD8+ T cell receptor repertoire. Sci Transl Med 2, 47ra64 (2010).
9. Li, H. *et al.* Recombinatorial biases and convergent recombination determine interindividual TCRbeta sharing in murine thymocytes. J Immunol 189, 2404-2413 (2012).
10. Venturi, V. *et al.* TCR beta-chain sharing in human CD8+ T cell responses to cytomegalovirus and EBV. J Immunol 181, 7853-7862 (2008).
11. Gandhi, M.K. & Khanna, R. Human cytomegalovirus: clinical aspects, immune regulation, and emerging treatments. Lancet Infect Dis 4, 725-738 (2004).
12. Babel, N. *et al.* Clonotype analysis of cytomegalovirus-specific cytotoxic T lymphocytes. J Am Soc Nephrol 20, 344-352 (2009).
13. Iancu, E.M. *et al.* Clonotype selection and composition of human CD8 T cells specific for persistent herpes viruses varies with differentiation but is stable over time. J Immunol 183, 319-331 (2009).
14. Khan, N., Cobbold, M., Keenan, R. & Moss, P.A. Comparative analysis of CD8+ T cell responses against human cytomegalovirus proteins pp65 and immediate early 1 shows similarities in precursor frequency, oligoclonality, and phenotype. J Infect Dis 185, 1025-1034 (2002).
15. Klarenbeek, P.L. *et al.* Deep sequencing of antiviral T-cell responses to HCMV and EBV in humans reveals a stable repertoire that is maintained for many years. PLoSPathog 8, e1002889 (2012).
16. Klinger, M. *et al.* Combining next-generation sequencing and immune assays: a novel method for identification of antigen-specific T cells. PLoS One 8, e74231 (2013).
17. Koning, D. *et al. In vitro* expansion of antigen-specific CD8(+) T cells distorts the T-cell repertoire. J Immunol Methods 405, 199-203 (2014).
18. Miconnet, I. *et al.* Large TCR diversity of virus-specific CD8 T cells provides the mechanistic basis for massive TCR renewal after antigen exposure. J Immunol 186, 7039-7049 (2011).
19. Price, D.A. *et al.* Avidity for antigen shapes clonal dominance in CD8+ T cell populations specific for persistent DNA viruses. J Exp Med 202, 1349-1361 (2005).
20. Retiere, C. *et al.* Generation of cytomegalovirus-specific human T-lymphocyte clones by using autologous B-lymphoblastoid cells with stable expression of pp65 or IE1 proteins: a tool to study the fine specificity of the antiviral response. J Virol 74, 3948-3952 (2000).
21. Scheinberg, P. *et al.* The transfer of adaptive immunity to CMV during hematopoietic stem cell transplantation is dependent on the specificity and phenotype of CMV-specific T cells in the donor. Blood 114, 5071-5080 (2009).
22. Schub, A., Schuster, I.G., Hammerschmidt, W. & Moosmann, A. CMV-specific TCR-transgenic T cells for immunotherapy. J Immunol 183, 6819-6830 (2009).
23. Trautmann, L. *et al.* Selection of T cell clones expressing high-affinity public TCRs within Human cytomegalovirus-specific CD8 T cell responses. J Immunol 175, 6123-6132 (2005).
24. Wynn, K. K. *et al.* Impact of clonal competition for peptide-MHC complexes on the CD8+ T-cell repertoire selection in a persistent viral infection. Blood 111, 4283-4292 (2008).
25. Peggs, K. *et al.* Characterization of human cytomegalovirus peptide-specific CD8(+) T-cell repertoire diversity following *in vitro* restimulation by antigen-pulsed dendritic cells. Blood 99, 213-223 (2002).
26. Weekes, M.P., Wills, M.R., Mynard, K., Carmichael, A.J. & Sissons, J.G. The memory cytotoxic T-lymphocyte (CTL) response to human cytomegalovirus infection contains individual peptide-specific CTL clones that have undergone extensive expansion *in vivo*. J Virol 73, 2099-2108 (1999).
27. Lim, A. *et al.* Frequent contribution of T cell clonotypes with public TCR features to the chronic response against a dominant EBV-derived epitope: application to direct detection of their molecular imprint on the human peripheral T cell repertoire. J Immunol 165, 2001-2011 (2000).
28. Yousfi Monod, M., Giudicelli, V., Chaume, D. & Lefranc, M.P. IMGT/JunctionAnalysis: the first tool for the analysis of the immunoglobulin and T cell receptor complex V-J and V-D-J JUNCTIONs. Bioinformatics 20 Suppl 1, i379-385 (2004).
29. Carlson, C.S. *et al.* Using synthetic templates to design an unbiased multiplex PCR assay. Nat Commun 4, 2680 (2013).
30. Storey, J.D. & Tibshirani, R. Statistical significance for genomewide studies. Proc Natl Acad Sci USA 100, 9440-9445 (2003).

REIVINDICACIONES

1. Un método para predecir la presencia o ausencia de una enfermedad infecciosa determinada en un sujeto de estado de infeccioso desconocido, comprendiendo el método:

a) realizar amplificación y secuenciación de alto rendimiento de ADN genómico obtenido de una muestra que comprende linfocitos T obtenida del sujeto para determinar un perfil de TCR que comprende secuencias de aminoácidos de CDR3 únicos;
b) comparar el perfil de TCR con una base de datos de secuencias de receptores de linfocitos T públicos de diagnóstico previamente identificado que de manera estadísticamente significativa están asociadas a la enfermedad infecciosa determinada;
c) generar una primera puntuación para el sujeto, en donde la primera puntuación es la proporción de secuencias de TCR únicas en el perfil del sujeto que coincide con las secuencias de TCR públicas en la base de datos;
d) introducir la primera puntuación de c) en un algoritmo, en donde el algoritmo compara la primera puntuación del sujeto y el estado infeccioso de una pluralidad de sujetos de estado infeccioso desconocido, en donde el algoritmo es un modelo de regresión logística; y
e) determinar una probabilidad estimada de estado infeccioso del sujeto como resultado del algoritmo

2. El método de la reivindicación 1, en donde la infección es por un citomegalovirus (CMV), un virus de Epstein-Barr (EBV) o un virus del herpes simple (HSV).

3. El método de la reivindicación 1, en donde la infección es una infección por CMV y la base de datos de las secuencias de linfocitos T públicos se determina que de manera estadísticamente significativa está asociada con infección por CMV obteniendo secuencias de TCR únicas de un grupo de sujetos con infección por CMV y de un grupo de sujetos sin infección por CMV.

4. El método de la reivindicación 3, en donde el método comprende someter cada secuencia de TCR única a una prueba exacta de Fisher de una cola basándose en la presencia o en la ausencia de cada secuencia de TCR en los sujetos de estado infeccioso de CMV desconocido.

5. El método de la reivindicación 4, en donde la hipótesis nula es que cada secuencia de TCR no es más común en sujetos con una infección por CMV que en sujetos sin una infección por CMV.

6. El método de la reivindicación 4, en donde se establece un umbral teórico de p-valor y la tasa de descubrimientos falsos (FDR) se controla por permutación del estado infeccioso de CMV en cada sujeto para generar una distribución empírica de p-valores nulos.

7. El método de la reivindicación 6, en donde el umbral teórico de p-valor es menor o igual a $1{,}0*10^{-4}$.

8. El método de la reivindicación 3, em donde las secuencias de TCR únicos identificadas estadísticamente que se correlacionan con la presencia o ausencia de infección por CMV comprenden una o más de las SEQ ID NO: 1 a 142.

9. El método de una cualquiera de las reivindicaciones 1-8, en donde la amplificación comprende una PCR multiplexada utilizando un conjunto de cebadores directos que se hibridan específicamente con segmentos V y un conjunto de cebadores inversos que se hibridan específicamente con segmentos J de un locus de TCR, en donde la PCR multiplexada permite la amplificación de todas las combinaciones posibles de VJ (y VDJ) dentro de una población dada de linfocitos T o B.

10. El método de una cualquiera de las reivindicaciones 1-9, los datos de secuencia sin procesar de la secuenciación de alto rendimiento se preprocesan para eliminar errores en la secuencia primaria de cada lectura y comprimir los datos.

11. El método de una cualquiera de las reivindicaciones 1-10, en donde se utiliza un algoritmo del vecino más cercano para colapsar los datos de la secuenciación de alto rendimiento en secuencias únicas fusionando secuencias estrechamente relacionadas para eliminar errores tanto de PCR como de secuenciación.

FIG. 1

Respuestas de linfocitos T públicos

FIG. 2

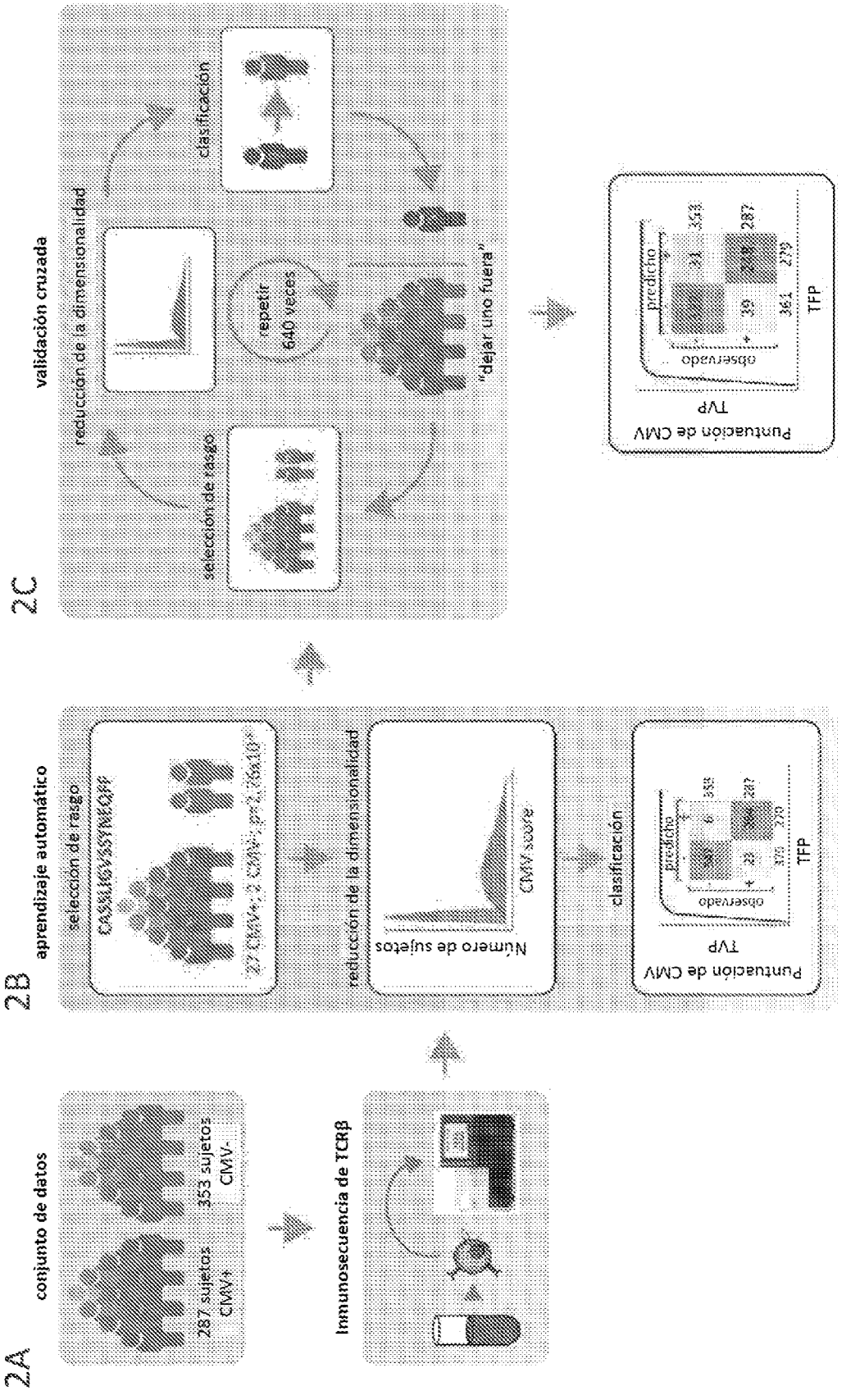
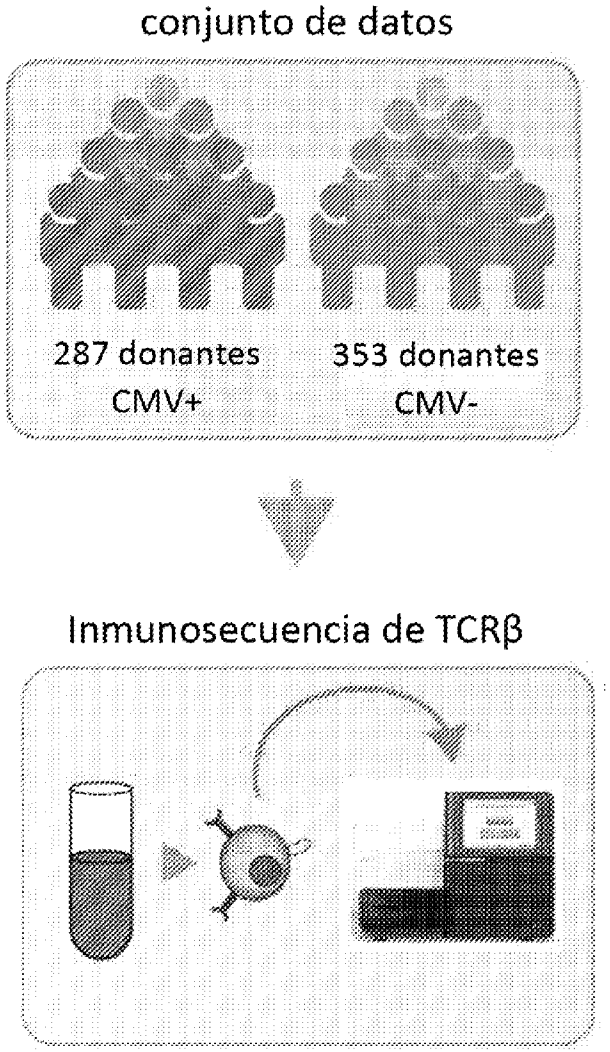

2A
conjunto de datos
287 sujetos CMV+
353 sujetos CMV-
Inmunosecuencia de TCRβ
2B
aprendizaje automático
selección de rasgo
reducción de la dimensionalidad
Número de sujetos
CMV score
clasificación
Puntuación de CMV
TVP
TFP
predicho
observado
2C
validación cruzada
reducción de la dimensionalidad
selección de rasgo
repetir 640 veces
clasificación
"dejar uno fuera"
Puntuación de CMV
TVP
TFP
predicho
observado

FIG. 3A

- 640 sujetos sanos
  - Donantes de médula ósea en el FHCRC
  - Principalmente sanos
- Serotipo de CMV determinado (rutinario para donantes de MO)
  - 287 CMV+, 353 CMV-
- Tipificación de HLA realizada (rutinario para donantes de MO)
- Secuenciados 250.000 TCR de sangre periférica



ES 3 002 665 T3

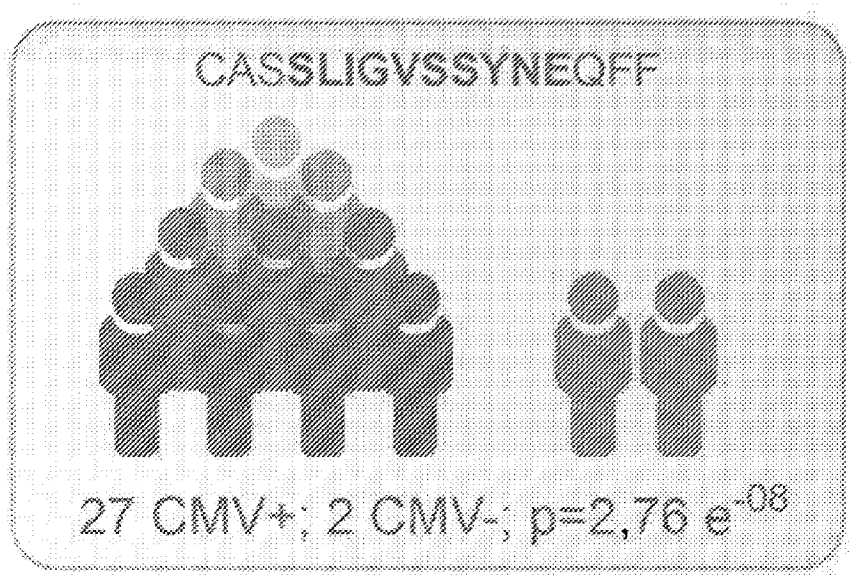
FIG. 3B

Sexo	CMV+	CMV-	Total
Femenino	150	147	297
Masculino	137	206	343
Total	287	353	640
Raza/etnia*	**CMV+**	**CMV-**	**Total**
Blanca	164	212	376
Negra o afroamericana	8	0	8
Asiática	15	2	17
Amerindio o nativo de Alaska	7	2	9
Hawaiano native u otro isleño del Pacífico	3	0	3
Hispano o latino	20	6	26
Desconocida	70	131	201
Total	287	353	640
Edad (en años)	**CMV+**	**CMV-**	**Total**
Media	42	37	40
Mediana	44	38	41
Intervalo	5-74	1-70	1-74
n.º de individuos de edad desconocida	31	56	87

*Se usaron las categorías de raza y etnia tal como se definen por el NIH. Todos los individuos hispanos/latinos se clasificaron como Raza=Desconocida pero Etnia=Hispana o latina. Los 26 individuos clasificados como Desconocida no proporcionaron información sobre su raza o etnia.

FIG. 4

Ensayo para la asociación de los TCR al CMV



Rasgo
CASSLIGVSSYNEQFF

	CMV+	CMV-	
Clon presente	27	2	29
Clon ausente	260	351	611
	287	353	N = 640 p = 2,8E-08*

* TDF estimada por permutación de marcadores del CMV

ES 3 002 665 T3

FIG. 5

TCR asociados al CMV

Secuencia de AA	Gen V	Gen J	CMV+	CMV-	P-valor
CASSQGAYEQYF	TCRBV09-01	TCRBJ02-07*01	60	11	3,31E-13
CASSPDRVGQETQYF	TCRBV05-01*01	TCRBJ02-05*01	34	1	9,54E-12
CASSIGPLEHNEQFF	TCRBV19-01	TCRBJ02-01*01	30	0	1,48E-11
CASSIEGNQPQHF	TCRBV28-01*01	TCRBJ01-05*01	27	0	1,95E-10
CASSLVAGGRETQYF	TCRBV05-06*01	TCRBJ02-05*01	51	11	2,35E-10
...					

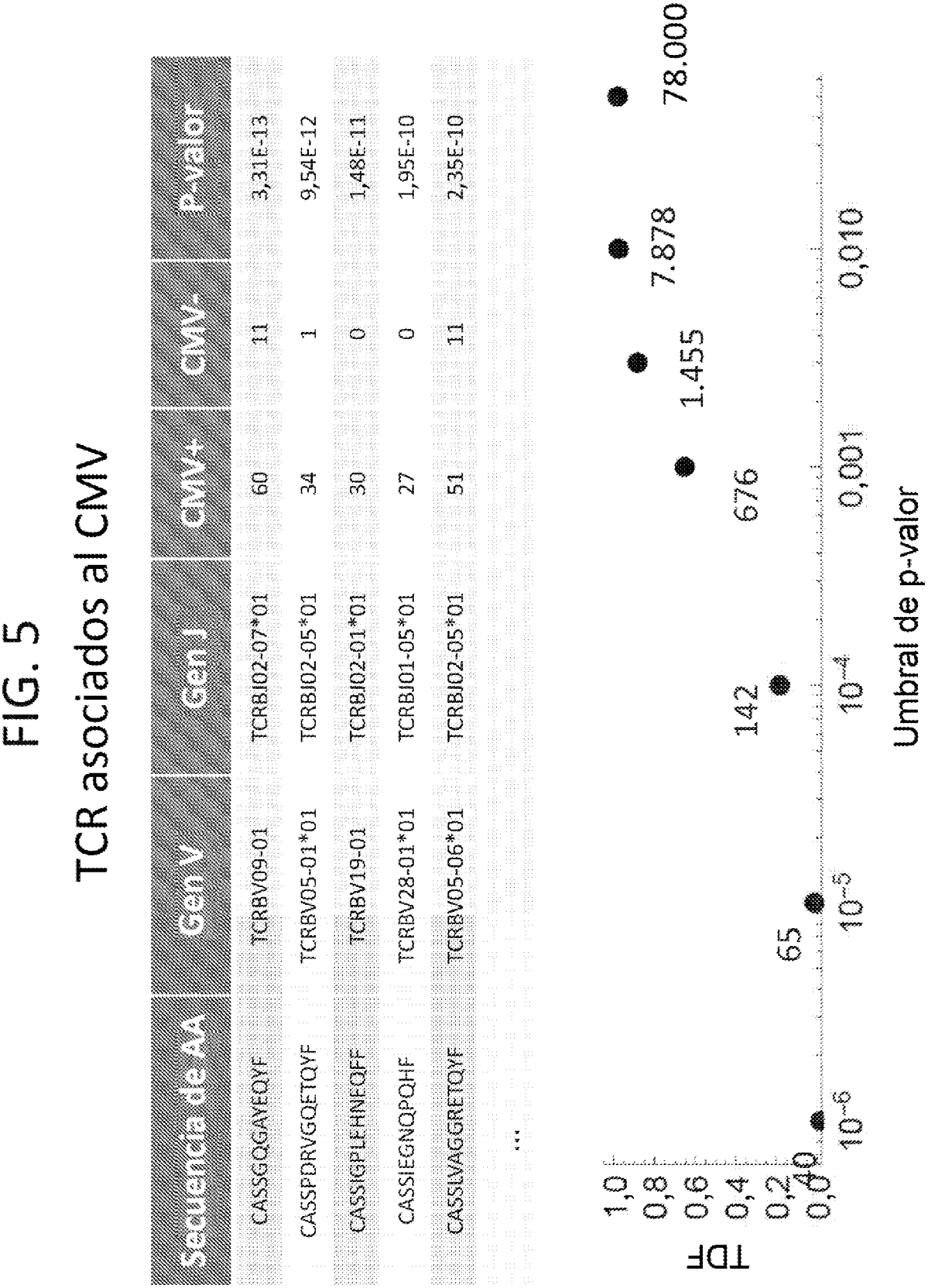
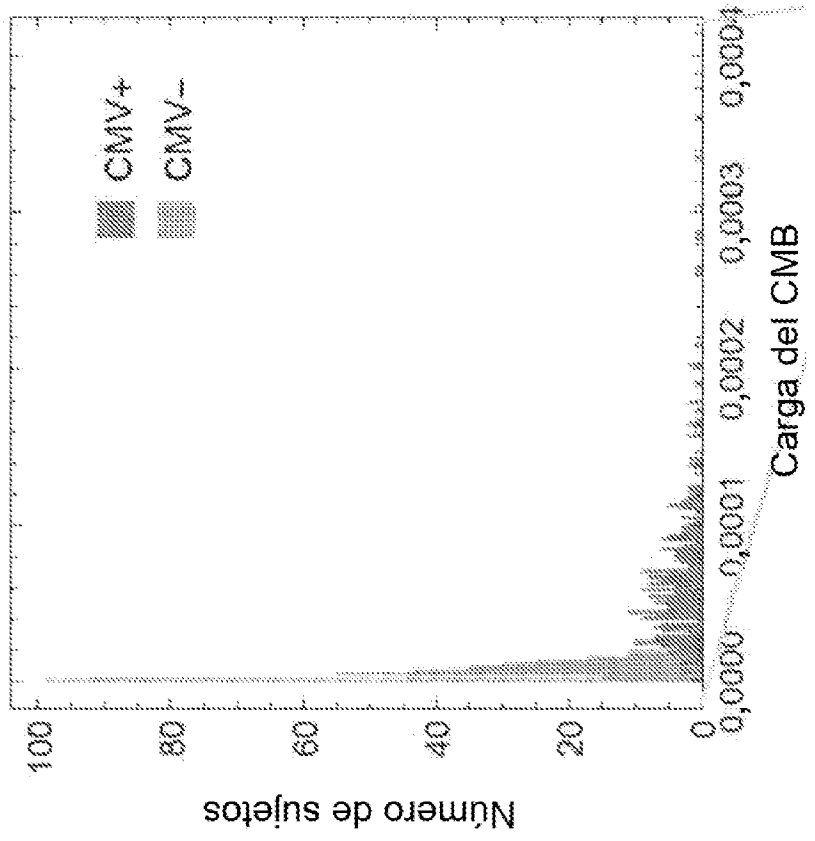
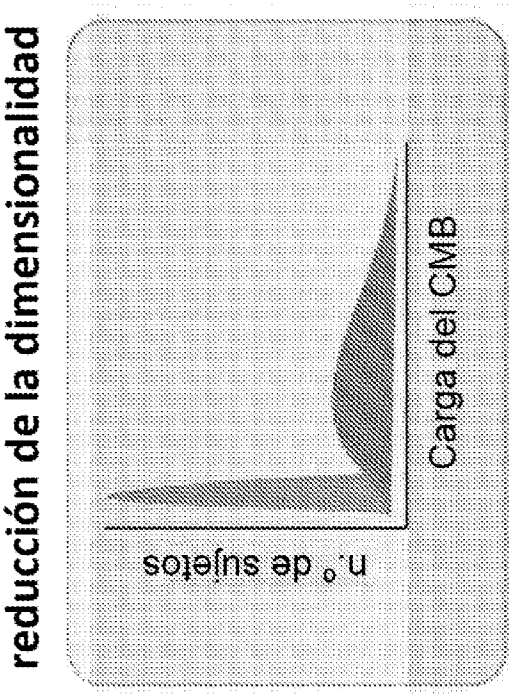


FIG. 6

Reducción de la dimensionalidad





* Carga del CMV: proporción de los únicos TCR en cada sujeto que están significativamente asociados al CMV

FIG. 7

Clasificación

- Regresión logística de la carga del CMV frente al estado del CMV
- Validación cruzada dejando uno fuera

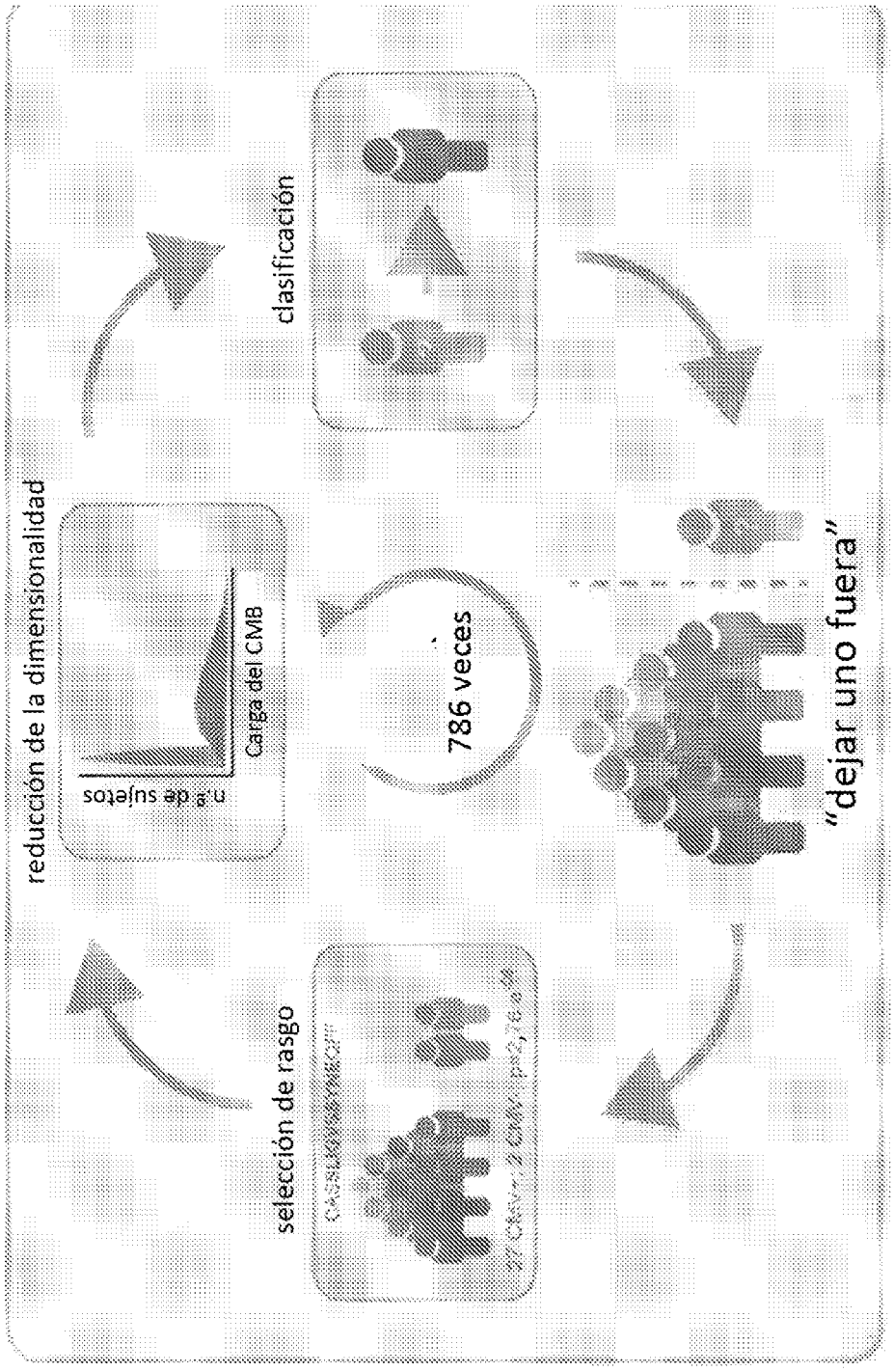


FIG. 8

Resultados de validación cruzada

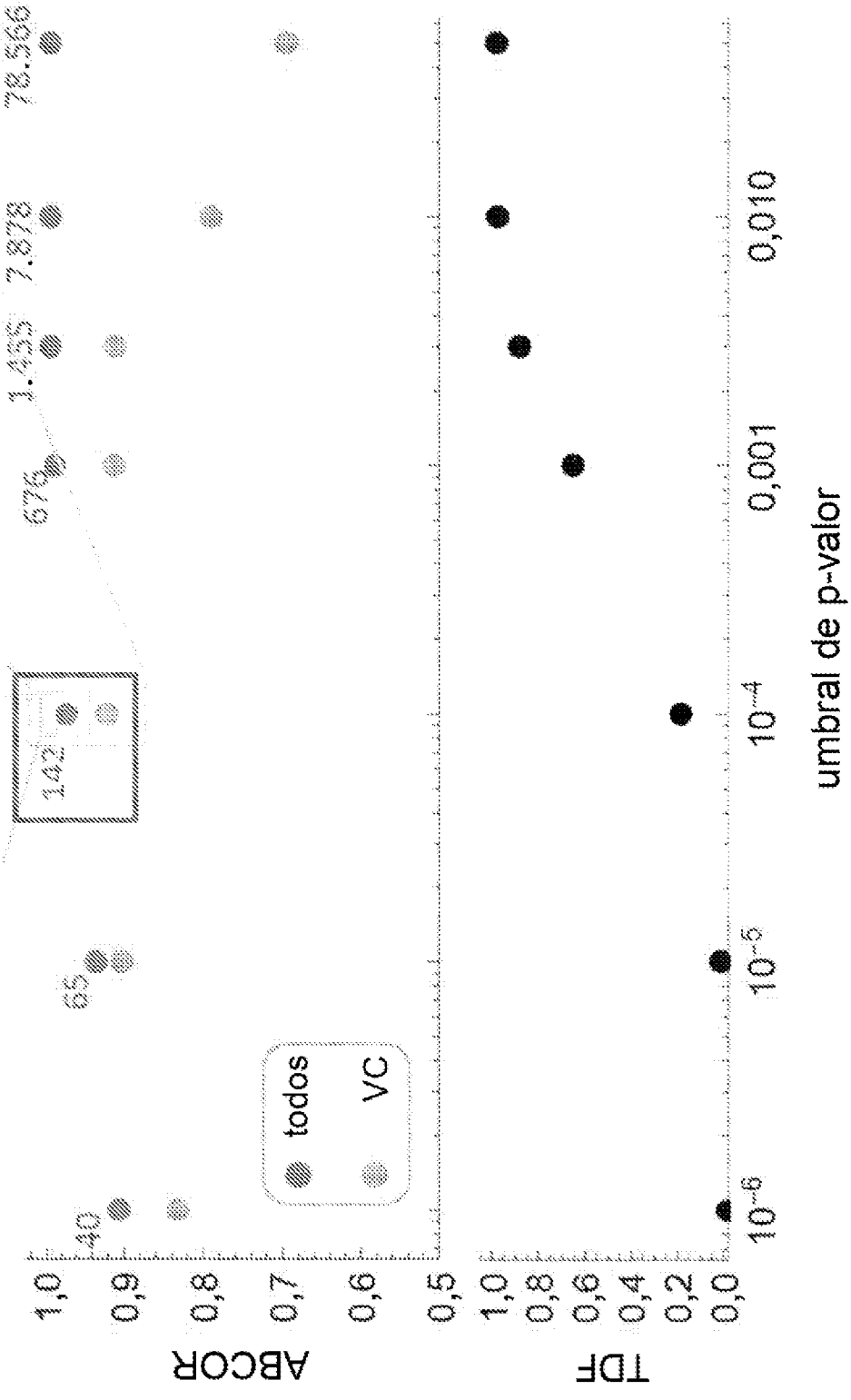


FIG. 9

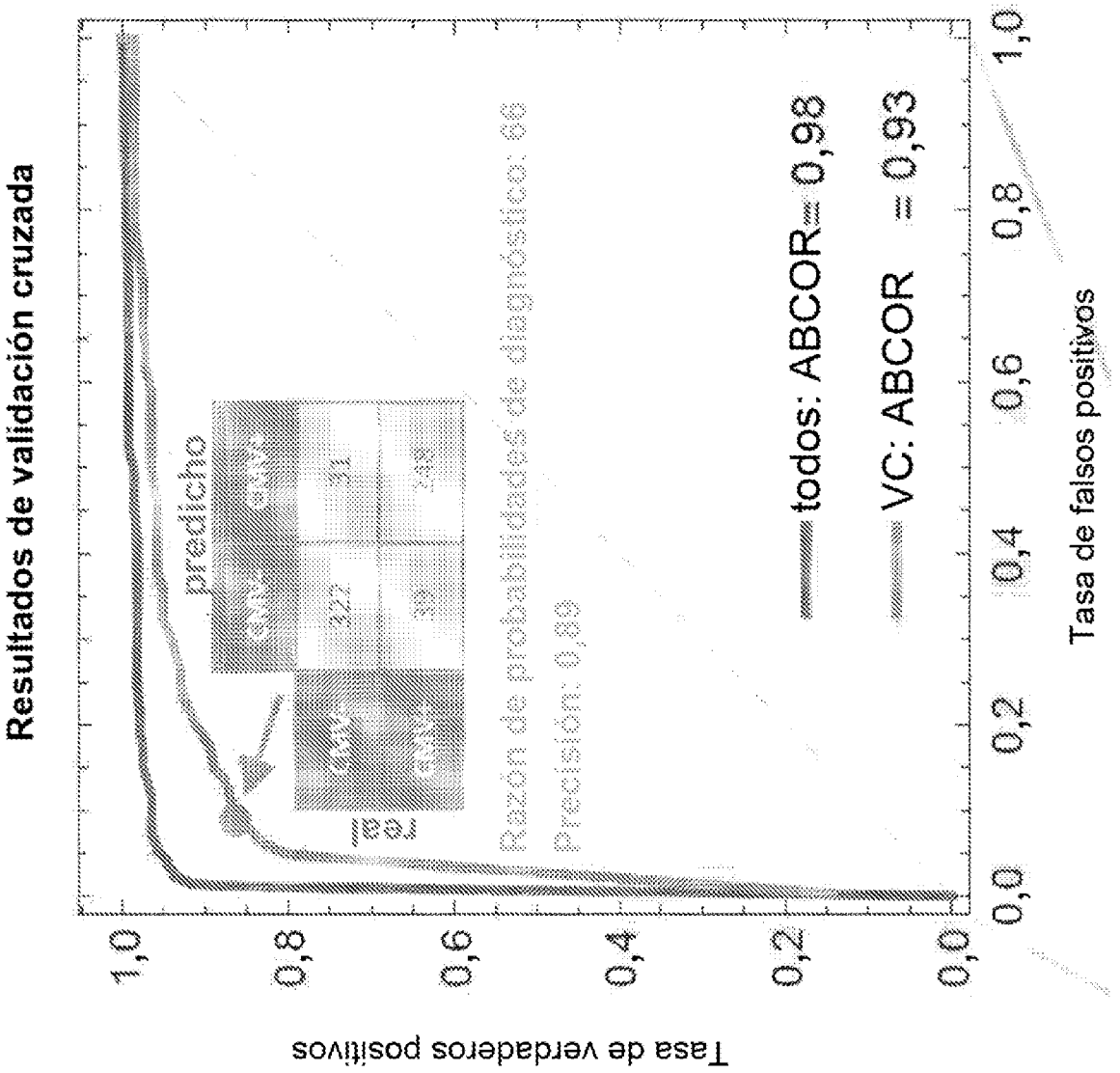

Resultados de validación cruzada
predicho
real
227
31
30
Razon de probabilidades de diagnóstico: 66
Precisión: 0,89
todos: ABCOR= 0,98
VC: ABCOR = 0,93
Tasa de verdaderos positivos
Tasa de falsos positivos
0,0
0,2
0,4
0,6
0,8
1,0

FIG. 10

Restricción de HLA de secuencias de TRCβ asociados al CMV

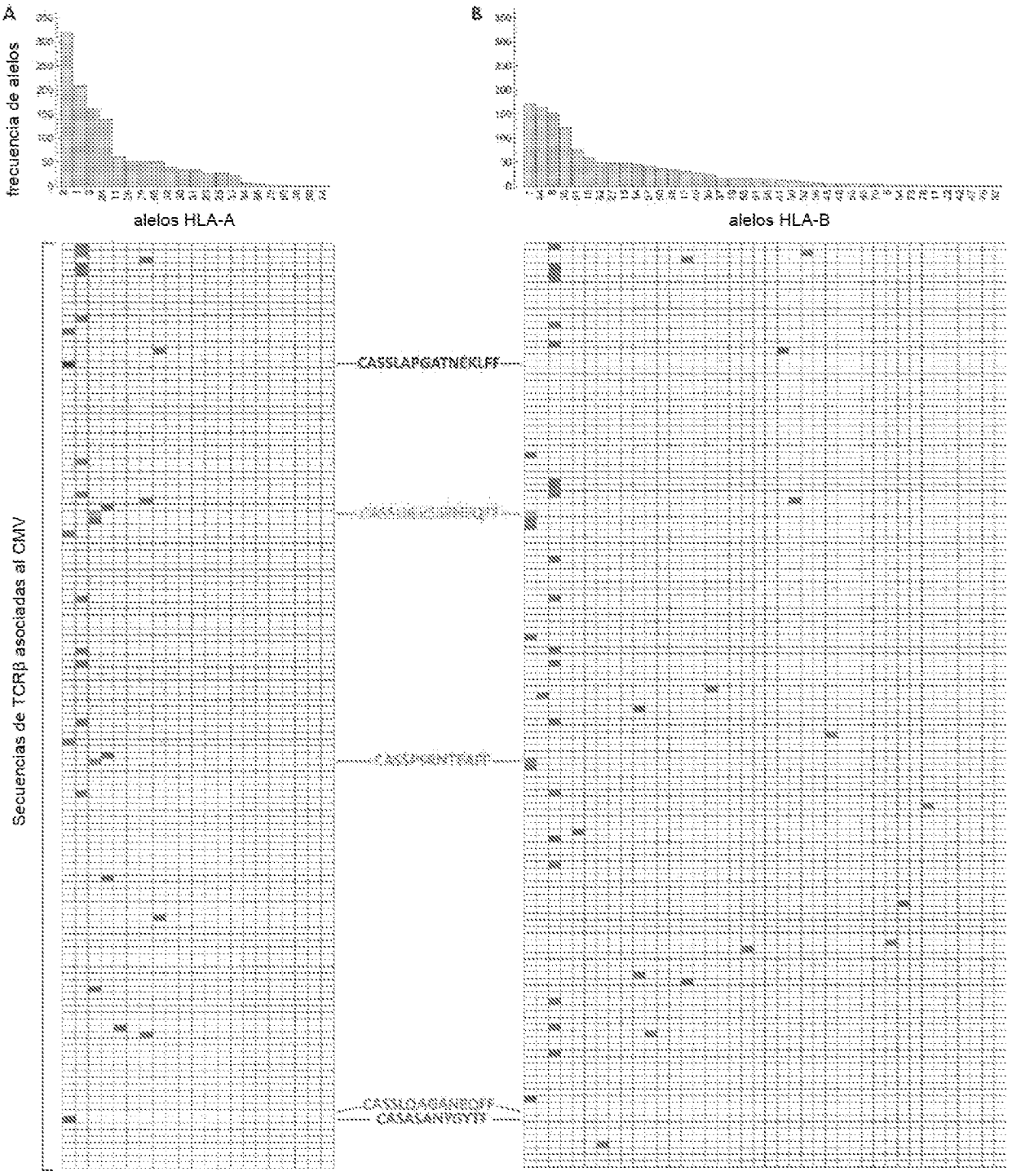


FIG. 11

Incidencia de secuencias de TCRβ reactivos al CMV en esta cohorte previamente documentadas

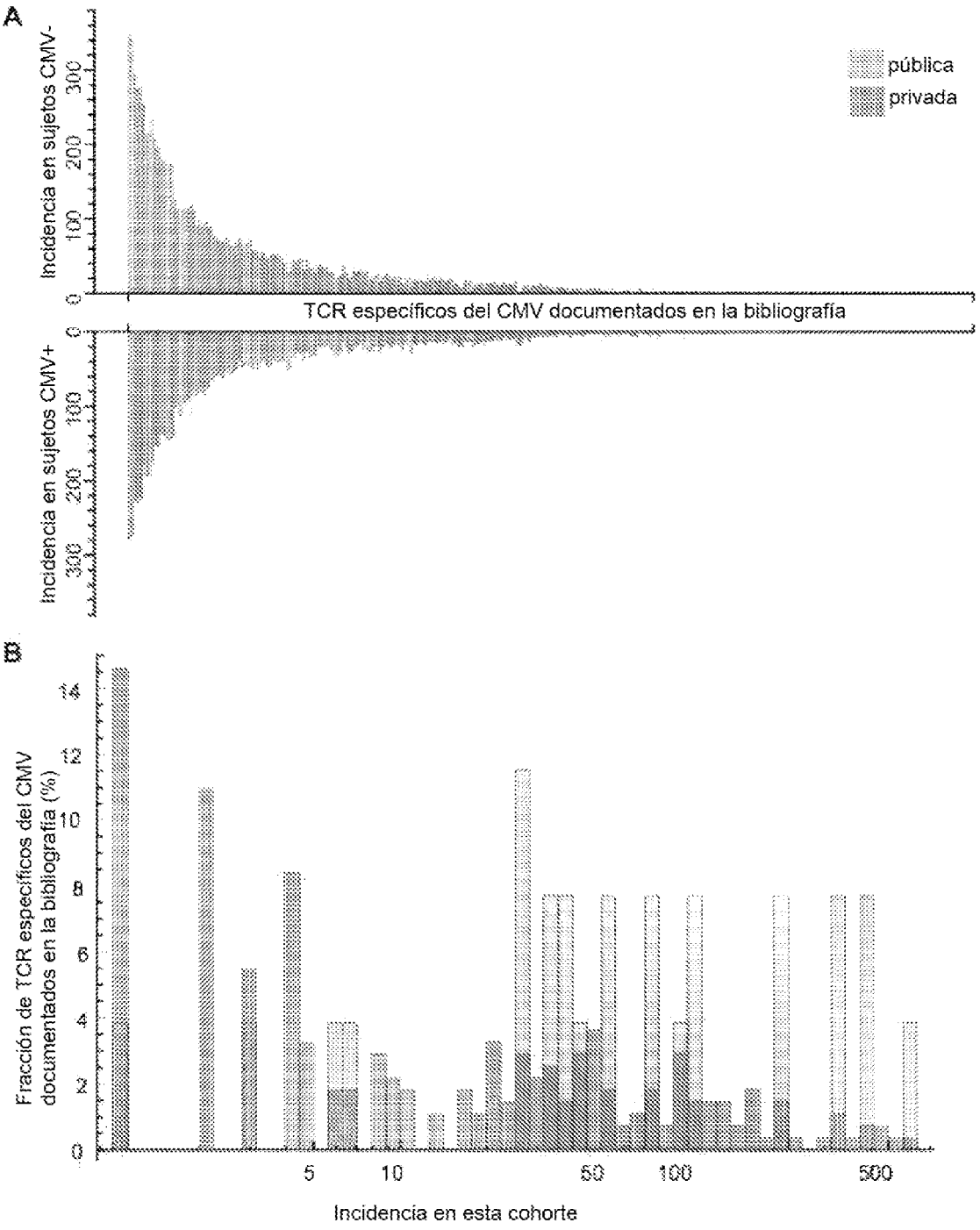


FIG. 12

Concordancia con la bibliografía anterior

este estudio					publicaciones anteriores						
secuencia de aa	gen V	gen J	asociación de HLA	pública	V gen*	¿coincidente?	J gen*	¿coincidente?	asociación de HLA	¿coincidente?	Ref.
CASSLAPGATNEKLFF	TCRBV07-06*01	TCRBJ01-04*03	HLA-A2	sí	6S3 (07-06)	√	1S4 (01-04)	√	HLA-A2	√	Trautmann et al., 2005
					7-6 (07-06)	√	1-4 (01-04)	√	HLA-A2	√	Price et al., 2005
					7-6 (07-06)	√	1-4 (01-04)	√	HLA-A2	√	Venturi et al., 2008
					6.2/7-9 (07-08 O-09)	X	1-4 (01-04)	√	HLA-A2	√	Miconnet et al. 2011
[illegible]	TCRBV07-09	TCRBJ02-01*03	HLA-B7, HLA-A3	sí	6.4 (07-06)	X	2-1 (02-01)	√	HLA-B7	√	Weekes et al., 1999
					6.2/7-9 (07-08 O-09)	√	2.1 (02-01)	√	HLA-B7	√	Miconnet et al., 2011
[illegible]	TCRBV04-03*01	TCRBJ01-01*03	HLA-B7, HLA-A3	no	7.2 (04-03)	√	1.1 (01-01)	√	HLA-B7	√	Weekes et al., 1999
[illegible]	TCRBV07-02*01	TCRBJ02-01*03	-	no	6.1 (07-03)	X	2.1 (02-01)	√	HLA-A2	NA	Peggs et al., 2003
[illegible]	TCRBV12	TCRBJ01-02*03	HLA-A2	no	12.4 (12-04)	√	1-2 (01-02)	√	HLA-A2	√	Venturi et al., 2008

(*) Gen documentado, seguido por el nombre en la nomenclatura actual (según IMGT); para Peggs et al., la identidad del gen J se dedujo a partir de la secuencia de aminoácidos.

FIG. 13

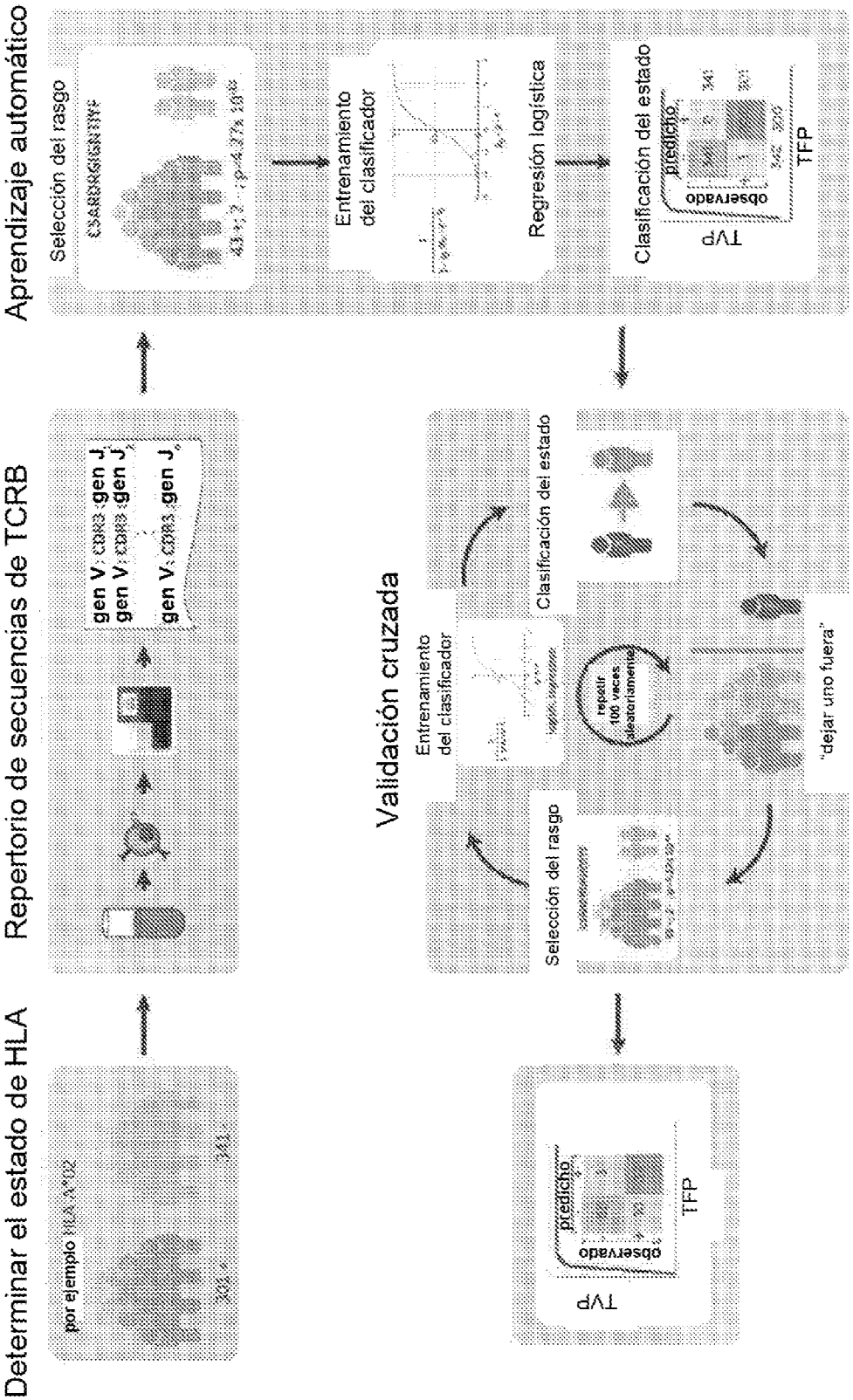

Determinar el estado de HLA
por ejemplo HLA-A*02
Repertorio de secuencias de TCRB
gen V; CDR3 gen J
gen V; CDR3 gen J
gen V; CDR3 gen J
Aprendizaje automático
Selección del rasgo
Entrenamiento del clasificador
Regresión logística
Clasificación del estado
observado
predicho
VP
FP
Validación cruzada
Entrenamiento del clasificador
Clasificación del estado
Selección del rasgo
repetir 100 veces, aleatoriamente
"dejar uno fuera"

FIG. 14

Selección de rasgo

P-valor para la asociación de cada TCR con el estado del alelo usando la prueba exacta de Fisher (dos colas)

	HLA-A2+	HLA-A2-
Secuencia de TCRβ / presente	n_{i+}	n_{i-}
Secuencia de TCRβ / no presente	$N_+ - n_{i+}$	$N_- - n_{i-}$

$$\Pr(\text{tabla}) = \frac{\binom{n_+ + n_-}{n_+}\binom{N_+ + N_- - n_+ - n_-}{n_+ + n_-}}{\binom{N_+ + N_-}{N_+}}$$

TCR	Datos de asociación			
	Sujetos HLA-A2+	Sujetos HLA-A2+	p-valor	selección
TCRBV07-02*01|CASSLAQAQFF|TCRBJ01-05*01	17	0	2,99e-08	+
CRBV19-01|CASSIRSSYEQY|TCRBJ02-07*01	15	0	9,43e-08	+
CRBV20|CSARDRGGNTIY|TCRBJ01-03*02	43	2	5,25e-13	+
CRBV27-02*01|CASSCPGGTGFY|TCRBJ02-03*02	49	2	2,12e-16	+
CRBV28-01*01|CASVPLTGGAF|TCRBJ01-01*02	5	36	7,90e-08	-
CRBV28-01|CASVPRSGTSGF|TCRBJ02-01*02	30	4	5,41e-06	+
CRBV28-01|CASVPSGQGAYEQY|TCRBJ02-07*01	20	0	2,43e-06	+
CRBV28-01*01|CASVPTGGTGELF|TCRBJ02-02*01	16	0	6,47e-07	+
CRBV28-01|CASVRLGQGPQH|TCRBJ01-05*01	14	46	4,41e-01	-
CRBV28-01*01|CASVSPSGQHTDTQY|TCRBJ02-03*01	15	1	3,47e-05	+
CRBV28-01|CASVSRGGGYEQY|TCRBJ02-07*01	14	1	7,27e-05	+
TCRBV28-01*01|CASVSTGGTGELF|TCRBJ02-02*01	20	4	4,67e-05	+

ES 3 002 665 T3

FIG. 15

Vector de rasgo para cada sujeto

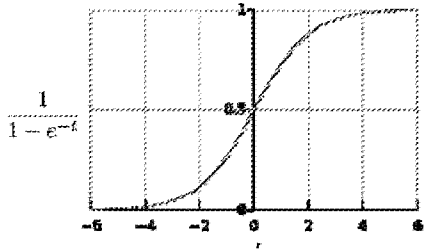
- Aplicar un corte de p-valor para identificar un conjunto de rasgos de TCR del listado anterior
- Comprobar la TDF por permutación del estado de los alelos
- Para HLA-A2, 288 rasgos a $p \leq 10^{-4}$, TDF = 0,1
- Calcular la abundancia de cada rasgo de TCR en cada sujeto

		Rasgos	
Sujeto	Estado de HLA-A2	TCRBV02-01*01|CASSRGGSSYNEQFF|TCRBJ02-01*01	TCRBV02-01*01|CASRVGTEAFF|TCRBJ01-03*01
1		0	0
2	+	1	10
3	+	1	1
4	+	3	0
5	-	0	0
6	-	0	2
7	-	1	0
8	+	12	100
9	-	0	0

FIG. 16
Aprendizaje automático

- Ajustar un modelo de regresión logística para los datos de los rasgos

$$\Pr(\text{HLA-A2+} \mid \mathbf{x}) = \frac{1}{1 - e^{-(\beta_0 + \boldsymbol{\beta} \cdot \boldsymbol{x})}}$$



- $\boldsymbol{x}$ → vector de abundancias para los clones de los rasgos
- $\boldsymbol{\beta}$ → coeficientes para cada rasgo, ajustados para maximizar la probabilidad de los datos de entrenamiento

ES 3 002 665 T3

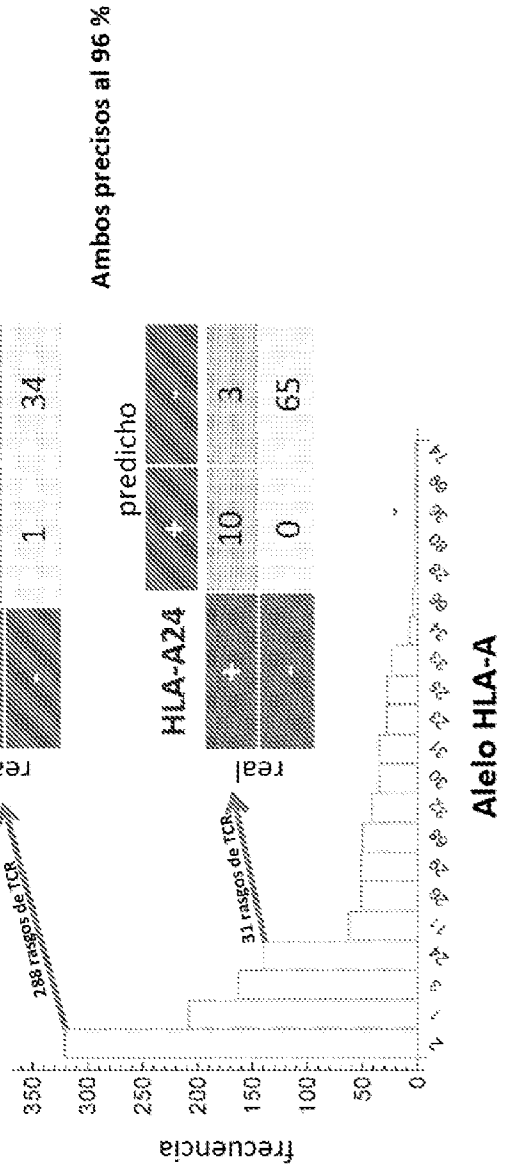
FIG. 17

Inferencia del estado de los alelos

- Validación cruzada:
  - Dejar un sujeto fuera del conjunto de datos
  - Realizar la identificación de rasgos de TCR
  - Ajustar el modelo de regresión logística
  - Usar el modelo para inferir el estado de los alelos para el sujeto que ha quedado fuera
- Realizar la validación cruzada para cada sujeto
- El entrenamiento sobre el conjunto de datos completo permite la clasificación de nuevos sujetos

ES 3 002 665 T3

FIG. 18

Precisión de la validación cruzada

- Basado en 78 rondas de validación cruzada+