DESCRIPCIÓN

TÍTULO DE LA INVENCIÓN

Método *in vitro* para identificar genotípicamente ambos alelos de al menos un locus de un gen HLA en muestras de ADN genómico de un sujeto

SECTOR DE LA TÉCNICA

El sector técnico al que pertenece la presente invención es el del genotipado molecular. En particular, el de los métodos de identificación genética que utilizan amplificación de secuencias por reacción en cadena de la polimerasa (PCR, del inglés *Polymerase Chain Reaction*), secuenciación masiva y análisis bioinformático para asignar secuencias de antígenos leucocitarios humanos (HLA, del inglés *Human Leukocyte Antigens*) a una familia o subfamilia HLA conocidas o, alternativamente, para determinar secuencias *de novo*.

ANTECEDENTES

El sistema de antígenos leucocitarios humanos (HLA) está controlado por genes localizados en el brazo corto del cromosoma 6 y constituye uno de los sistemas genéticos más polimórficos del ser humano. El conjunto especifico de las moléculas HLA expresadas por un individuo determina el repertorio antigénico frente al que pueden responder las células del sistema inmune de cada individuo, incluyendo el reconocimiento de tejidos o moléculas como propias o extrañas, por lo que es el sistema responsable de la compatibilidad en el trasplante de órganos o el desarrollo y control de patologías de origen infeccioso y autoinmune. Existen 3 tipos de moléculas HLA.

Las moléculas HLA de clase I son glicoproteínas transmembrana presentes en todas las células nucleadas del organismo. Estas glicoproteínas constan de una cadena alfa codificada por cada uno de los genes HLA de clase I genes unida a una molécula de la microglobulina 2 de modo no covalente. Su función es presentar antígenos proteicos de proteínas producidas en el interior de las células a los linfocitos T citotóxicos o CD8 (de ahí que se llame a esta vía citosólica o endógena). Las células que presentan fragmentos de proteínas extrañas al organismo serán atacadas por el sistema inmune.

Las de clase II están presentes en células profesionales presentadoras de antígenos, como las células dendríticas, linfocitos B y fagocitos mononucleares. A nivel molecular,

estas proteínas consisten en 2 cadenas polipeptídicas codificadas por genes localizados en las regiones HLA-DP, -DQ o -DR del cromosoma 6. Los linfocitos T helper o CD4 reconocen en la superficie de las células presentadoras de antígenos las moléculas de tipo II, activándose de esta forma. Los antígenos HLA de tipo II presentados por las células proceden de proteínas extracelulares (por ello esta vía se denomina exógena o endocítica).

Las moléculas HLA de clase III codifican, además de otros productos, varias proteínas secretadas que desempeñan funciones inmunitarias, inclusive componentes del sistema de complemento y moléculas relacionadas con la inflamación.

Los genes de HLA están estrechamente ligados de tal forma que se heredan en bloque, como haplotipos de cada padre de una forma mendeliana. Las combinaciones aleatorias de diferentes loci de HLA en un haplotipo son muy elevadas, sin embargo, ciertos haplotipos se encuentran mucho más frecuentemente representados en la población frente a otros, en una proporción superior a la que cabría esperar por azar (desequilibrio de ligamiento). Las moléculas de HLA clase I y II son los antígenos más inmunogénicos que se reconocen durante el rechazo de un trasplante alogénico. El determinante más fuerte es HLA-DR, seguido por HLA-B y -A. En la actualidad, las guías y consorcios oficiales de referencia para el establecimiento de los criterios requeridos para proceder al tipaje HLA, establecen como requerimiento mínimo para el trasplante de Órgano Sólido (SOT) la determinación de los alelos HLA-A, HLA-B y HLA-DRB1, y para Trasplante de Progenitores Hematopoyéticos (HSCT) los alelos HLA-A, HLA-B, HLA-C, HLA-DRB1, HLA-DQB1, HLA-DPB1 a nivel de alta resolución.

La determinación precisa de los haplotipos de HLA, tanto de clase I, como de clase II, en un individuo es, pues, un hecho clave para la determinación de la compatibilidad donante-receptor, en el caso de un trasplante de órganos o tejidos.

En el estado de la técnica se han descritos métodos para genotipar alelos HLA. Los documentos más representativos del estado de la técnica aparecen citados más abajo y se discute su contenido técnico.

El documento EP3006571A1 se refiere a un método para determinar la secuencia de las moléculas HLA mediante *long range* PCR (PCR de largo alcance) utilizando cebadores diseñados para hibridar con zonas corriente arriba y corriente abajo de al menos dos genes seleccionados entre HLA-A, HLA-B, HLA-C, HLA-DQA1, HLA-DQB1, HLA-DPA1, HLA-DPB1, HLA-DRB1, HLA-DRB3, HLA-DRB4 y HLA-DRB5, amplificar al menos dos genes simultáneamente, secuenciar y buscar homologías comparando con la base de

datos IMGT/HLA. Esta base de datos es parte del proyecto internacional ImMunoGeneTics (IMGT) y está especializada en secuencias del complejo mayor de histocompatibilidad humano (MHC, del inglés Major Histocompatibility Complex) e incluye las secuencias de referencia del sistema HLA nombradas por la Organización Mundial de la Salud (OMS) para factores del sistema HLA.

El documento EP2735617A1 describe un método similar al anterior para determinar la secuencia de HLA mediante PCR de largo alcance en el que se utilizan algunos cebadores alternativos a los descritos en EP3006571A1.

El documento WO2015/200701A2 hace referencia a un método para determinar la secuencia de genes HLA mediante PCR de largo alcance utilizando cebadores que permiten la amplificación del gen completo seleccionados fuera de zonas de gran variabilidad, utilizando una mezcla de nucleótidos naturales y análogos de nucleótidos y realizando un análisis de deconvolución para determinar el genotipo de cada alelo.

En WO2014/116729A2 se describe un método para genotipar alelos de HLA que comprende las etapas de amplificación de exones e intrones de los genes en una reacción PCR de largo alcance, secuenciación profunda (*deep sequencing*) del gen amplificado y realización de un análisis de deconvolución para determinar el genotipo de cada alelo.

El documento WO2014/065410A1 se refiere a un método que comprende la amplificación por PCR utilizando cebadores que hibridan específicamente con la región corriente arriba y corriente abajo de los genes HLA-A, HLA-B, HLA-C, HLA-DRB1, HLA-DPB1 y HLA-DQB1 y la realización de una búsqueda de homología respecto a la base de datos IMGT/HLA.

El documento Hosomichi, K. et al. (2013) divulga la aplicación de PCR de largo alcance para amplificar seis genes completos HLA (HLA-A/B/C, DRB1, DQB1 y DPB1) utilizando cebadores que se unen a zonas conservadas del genoma, seguido de la construcción de librerías y secuenciación multiplex de dichas librerías.

El documento EP2599877A1 describe un método en el que se realiza amplificación de genes HLA con unos cebadores específicos, seguido de fragmentación de los amplicones obtenidos, obtención de librerías y alineamiento de las secuencias obtenidas después de secuenciación respecto a la base de datos IMGT/HLA.

El documento WO2016/054135A1 describe un método para genotipar ambos alelos de loci de HLA en el que se forman amplicones, se fragmentan, se secuencia mediante

secuenciación-por-síntesis, se alinean las secuencias nucleotídicas que solapan parcialmente para determinar una secuencia compuesta contigua que codifica para dominios de reconocimiento de antígenos (ARD, del inglés *Antigen Recognition Domains*), se comparan las secuencias nucleotídicas que solapan parcialmente con una librería de referencia de secuencias genómicas que codifican ARD de loci de HLA e identificar cada secuencia como una secuencia codificando un ARD conocido de un locus de HLA o una secuencia codificando un nuevo ARD de un locus de HLA. El método está dirigido a genotipar los genes HLA-A/B/C.

El documento WO2014/150924A2 describe un método implementado por ordenador para genotipar genes HLA. En el método, se identifican alelos relevantes mediante la descomposición de los alelos en exones, se calcula la distancia Hamming (HD, Hamming Distance) entre los exones y las lecturas que mejor se ajustan a estos exones.

El documento Shiina, T. et al. (2012) describe un método de genotipado molecular basado en secuenciación masiva en el que se determinan alelos para algunos genes HLA. En dicho método se han diseñado cebadores para amplificar los genes completos incluyendo las regiones promotoras y las regiones UTR, con el objetivo de reducir las ambigüedades en la determinación de alelos HLA.

El documento WO2014/066217A1 describe un método que comprende amplificar secuencias de ADN, secuenciarlas por secuenciación masiva y determinar la identidad de dichas secuencias amplificadas. La determinación de dicha identidad se realiza mediante un método implementado por ordenador que comprende las siguientes etapas: particionar cada lectura de secuencias en uno o dos haplotipos, basándose en la aplicación de un algoritmo *k-means* o un algoritmo de agrupación esperanza-maximización (*expectation-maximimization clustering*).

El documento EP3075863A1 describe un método en el que se amplifica y se somete a genotipado de ADN utilizando secuenciación masiva una región del exón 2 a una porción del exón 4 de genes HLA de clase II. En el método se han diseñado cebadores para amplificar específicamente los genes HLA-DRB1/3/4/5, HLA-DQB1 y HLA-DPB1.

RESUMEN DE LA INVENCIÓN

El problema del estado de la técnica consiste en mejorar la precisión, exactitud y fiabilidad en los métodos de determinación de haplotipos HLA que utilizan amplificación de secuencias por PCR, secuenciación masiva y análisis bioinformático de los resultados obtenidos en la secuenciación.

La solución a este problema técnico objetivo, según la presente invención, consiste en un método para determinar haplotipos HLA que combina las siguientes características técnicas para la determinación de los haplotipos de 11 genes HLA: utilización de combinaciones específicas de cebadores, hibridación de dichas combinaciones específicas de cebadores en regiones corriente arriba/abajo específicas de los genes y amplificación en un único amplicón de genes completos, regiones UTR completas y regiones corriente arriba/abajo de los genes y etapas de análisis bioinformático específicas que comprenden llamada de variantes (*variant calling*), alineamiento local de la secuencia de los bloques haplotípicos, pre-cálculo de distancias, selección de alelos candidatos, estudio filogenético, generación de árbol filogenético usando una estrategia *neighbor-joining*, test de contraste y asignación de los niveles y familias del haplotipo problema o, alternativamente, determinación de haplotipos HLA *de novo.* La invención se basa en esta combinación nueva e inventiva de características.

Los efectos ventajosos proporcionados por la combinación nueva e inventiva de estas características consisten en una mejora en la precisión, exactitud y fiabilidad en la determinación de haplotipos HLA. Así, en los ejemplos de la invención (ver Ejemplo 10) se describe una secuencia problema que se parece más a una familia de HLA conocida, de forma que los métodos conocidos del estado de la técnica asignarían, de forma errónea, esta familia HLA conocida a la secuencia problema. Sin embargo, el método de la invención permite determinar que esta secuencia problema no pertenece a dicha familia conocida, sino que es una secuencia *de novo***.**

La hibridación de combinaciones específicas de cebadores en regiones corriente arriba/abajo específicas de los genes y amplificación en un único amplicón de genes completos, regiones UTR completas y regiones corriente arriba/abajo de los genes realiza un tipaje alélico más preciso ya que proporciona más información relativa al perfil genético contenido en cada uno de los alelos candidatos permitiendo una asignación más exacta de alta resolución. También permite la identificación y descripción de alelos nuevos no descritos hasta la fecha al considerar un mayor contenido de información genética, al incluirse en el paso de amplificación, dichas regiones corriente arriba y debajo de los genes, conjuntamente con los genes completos, incluyendo éstos también las regiones UTR. El hecho de que se disponga de más información sobre la secuencia evita asignaciones erróneas ya que en las regiones intrónicas, UTRs o regiones corriente arriba/abajo puede estar contenida la información necesaria para discernir entre dos alelos con alta homología en la secuencia codificante pero diferente en la restantes

regiones que pueden afectar a la proteína resultante y por tanto al alelo genético al proteico y a la función de la molécula HLA en la secuencia.

Estas regiones corriente arriba/abajo presentan un menor contenido en variantes polimórficas permitiendo de esta forma realizar diseños de oligonucleótidos más específicos para el gen diana, evitando amplificaciones inespecíficas, amplificaciones de regiones pseudogénicas y posibles fenómenos de pérdida alélica (*allele dropout*), que podrían implicar asignaciones alélicas erróneas. Los fenómenos de pérdida alélica se producen cuando se pierde un alelo durante la amplificación de ADN por PCR. A los efectos de la presente memoria de patente, se utilizan indistintamente, como sinónimos, los términos cebador/es y oligonucleótido/s.

El método de la invención permite realizar un test genético a pacientes con el fin de conocer los haplotipos (combinación de alelos de un mismo loci de un cromosoma que son transmitidos juntos) de sus antígenos leucocitarios humanos (HLA en inglés) con una precisión, exactitud y fiabilidad mejorados respecto a los métodos conocidos en el estado de la técnica. La invención establece, mediante análisis bioinformático, la determinación precisa de estos haplotipos y permite saber qué individuos son compatibles para trasplantes; diagnóstico/pronóstico de enfermedades autoinmunes; tests de paternidad; reacciones adversas a fármacos en medicina personalizada; vacunación, etc....

DEFINICIONES

A los efectos de la presente memoria de patente, los siguientes términos técnicos se definen como sigue:

Neighbor-joining: Método descrito en Felsenstein, J. (1981), Saitou, N. et al. (1987) y Studier, J. A. et al. (1988). Método de reconstrucción filogenética que agrupa las secuencias basándose en distancias genéticas según el criterio de mínima evolución (BME, del inglés *Balanced Minimum Evolution*), en el que el mejor árbol filogenético es aquel que minimiza la longitud de las ramas internas. Con este método se crea un nuevo nodo en cada paso, y finalmente se obtiene un árbol filogenético único, que se considera el más ajustado a los datos. A partir de un árbol filogenético en estrella, se determina la pareja de secuencias más cercanas y se unen mediante un nodo interno. Este proceso se repite con el resto de secuencias hasta que quedan todas unidas por nodos internos que minimizan la longitud de cada una de las ramas internas.

Llamada de variantes (*variant calling*): La llamada de variantes a partir de los datos de secuenciación masiva se refiere a un grupo de técnicas computacionales para identificar la existencia de variantes de nucleótido único (SNV, *Single Nucleotide Variants*) a partir de los resultados de los experimentos de secuenciación masiva. Estas técnicas computacionales son alternativas a métodos experimentales basados en polimorfismos de nucleótido único (SNP, *Single Nucleotide Polymorphisms*) conocidos en toda la población. La llamada de variantes se utiliza en la identificación genotípica de SNP, con una amplia variedad de algoritmos diseñados para aplicaciones y diseños experimentales específicos. Estas técnicas se han aplicado de forma exitosa en la identificación de SNP raros dentro de una población y para detectar SNV somáticos dentro de un sujeto utilizando múltiples muestras de tejidos.

Silueta (Silhouette): Silueta se refiere a un método de interpretación y validación de consistencia dentro de grupos de datos. La técnica proporciona una representación gráfica de la situación de un objeto dentro de su grupo. El valor de silueta es una medida de la semejanza de un objeto respecto a su propio grupo (cohesión) en comparación con otros grupos (separación). El valor de silueta varía de -1 a +1, donde un valor alto indica que el objeto está bien emparejado en su propio grupo y está poco relacionado con los grupos vecinos. Si la mayoría de los objetos tienen un valor alto, entonces la configuración de agrupamiento es apropiada. Si muchos puntos tienen un valor bajo o negativo, entonces la configuración de agrupamiento puede tener demasiados o demasiado pocos grupos. La silueta se puede calcular con cualquier medida de distancia, como la distancia euclidiana o la distancia de Manhattan.

Contraste Wilcoxon: La prueba de los rangos con signo de Wilcoxon es una prueba no paramétrica para comparar el rango medio de dos muestras relacionadas y determinar si existen diferencias entre ellas. Se utiliza como alternativa a la prueba t de Student cuando no se puede suponer la normalidad de dichas muestras. Debe su nombre a Frank Wilcoxon, que la publicó en 1945. Es una prueba no paramétrica de comparación de dos muestras relacionadas y por lo tanto no necesita una distribución específica. Usa más bien el nivel ordinal de la variable dependiente. Se utiliza para comparar dos mediciones relacionadas y determinar si la diferencia entre ellas se debe al azar o no (en este último caso, que la diferencia sea estadísticamente significativa). Se utiliza cuando la variable subyacente es continua pero no se presupone ningún tipo de distribución particular.

PCR de largo alcance se refiere a la amplificación de longitudes de ADN que normalmente no pueden amplificarse utilizando métodos o reactivos rutinarios de PCR. Para moldes de de ADN simples, las polimerasas optimizadas para PCR de largo alcance

pueden amplificar hasta 30 kb y más. Para genomas complejos, un objetivo típico es 20 kb.

Test de contraste: Prueba para comparar el rango medio de dos muestras relacionadas y determinar si existen diferencias entre ellas.

Escala Phred: Un valor Phred es una medida de la calidad de la identificación de los nucleótidos generada por secuenciación automática de ADN. Originalmente fue desarrollado para ayudar en la automatización de la secuenciación de ADN en el Proyecto Genoma Humano. Los valores de calidad Phred se asignan a cada llamada a una base nucleotídica en secuencias obtenidas de forma automatizada. Los valores de calidad Phred son ampliamente aceptados para caracterizar la calidad de las secuencias de ADN y se pueden usar para comparar la eficacia de diferentes métodos de secuenciación. Quizás el uso más importante de los valores de calidad Phred es la determinación automática basada en calidad de secuencias de consenso.

Algoritmo BWA: *Burrows-Wheeler Aligner*, una herramienta bioinformática (algoritmo) para mapear secuencias cortas de nucleótidos a un genoma de referencia.

Algoritmo mem: Algoritmo de coincidencia de búsqueda de cadenas. Dadas dos cadenas, los mem son subcadenas comunes que no pueden extenderse hacia la izquierda o hacia la derecha sin causar un desajuste.

Binary Alignment Map (BAM) is the comprehensive raw data of genome sequencing; it consists of the lossless, compressed binary representation of the Sequence Alignment Map.

Archivo BAM: El mapa de alineamiento binario (*Binary Alignment Map*, BAM) es la información completa sin procesar de la secuenciación del genoma; consiste en la representación binaria sin pérdidas y comprimida del mapa de alineamiento de secuencia (*Sequence Alignment Map*, SAM). BAM es la representación binaria comprimida de SAM.

Software WhatsHap: WhatsHap es un programa informático para determinar la fase de las variantes genómicas usando lecturas de secuenciación de ADN, proceso llamado ensamblaje de haplotipos. Es especialmente adecuado para lecturas largas, pero también funciona bien con lecturas cortas.

SAMtools: Conjunto de herramientas para la selección de regiones concretas dentro de un genoma y para el el indexado de los archivos .bam. En la presente memoria se utilizó la versión 1.2 de SAMtools. El conjunto de herramientas SAMtools interactúa con el procesamiento posterior de alineaciones cortas de lectura de secuencia de ADN en los

formatos SAM (mapa de alineamiento de secuencia), BAM (mapa de alineamiento binario) y CRAM. Estos archivos se generan como salida por herramientas de alineamiento de lecturas cortas, como BWA. El conjunto contiene herramientas para llamada de variantes (*variant calling*), visualización del alineamiento, clasificación, indexación, extracción de datos y conversión de formatos. Los archivos SAM pueden ser muy grandes, por lo que la compresión se usa para ahorrar espacio. Los archivos CRAM tienen un formato de contenedor binario reestructurado orientado a columnas. SAMtools hace posible trabajar directamente con un archivo comprimido BAM, sin tener que descomprimir todo el archivo.

Software bcftools: El comando mpileup de SAMtools produce un archivo de formato pileup (o BCF) que proporciona, para cada coordenada genómica, las bases de lectura superpuestas e indels en esa posición en los archivos BAM de entrada. Esto se puede usar para llamadas de polimorfismos de nucleótido único (SNP), por ejemplo.

Electroferograma: Un electroferograma es un gráfico realizado con los resultados de un análisis por electroforesis. Se pueden realizar electroferogramas con resultados derivados de: pruebas genealógicas de ADN, pruebas de paternidad, secuenciación de ADN, huella genética. El electroferograma muestra la secuencia de datos producida por una máquina automática de secuenciación de ADN.

PCR Master Mix (PCRMM): PCRMM es una solución concentrada premezclada que tiene todos los componentes para una reacción de PCR en tiempo real que no son específicos de la muestra. La mezcla maestra PCRMM generalmente contiene una ADN polimerasa termoestable, dNTPs, $MgCl_2$ y aditivos en un tampón optimizado para PCR. Solo es necesario añadir el molde de ADN, los cebadores, las sondas (si se usan) y agua para alcanzar el volumen de reacción deseado. Se pueden incluir uno o más sondas para la detección fluorescente de la formación del producto en PCR cuantitativa en tiempo real (qPCR).

Secuenciación masiva: A efectos de la presente memoria de patente este término técnico es sinónimo y se utiliza indistintamente al término “Next Generation Sequencing (NGS)”.

Tagmentación (del inglés *tagmentation*). Proceso en el que las enzimas transposasas cortan aleatoriamente el ADN en fragmentos cortos (etiquetas, o *tags*). Simultáneamente la enzima añade en los extremos del ADN fragmentado unas secuencias específicas que permitirán la incorporación de los adaptadores en el siguiente proceso de amplificación por PCR.

Alelo: Cada una de las formas alternativas que puede tener un mismo gen que se diferencian en su secuencia y que se puede manifestar en modificaciones concretas de la función de ese gen (producen variaciones en características heredadas como, por ejemplo, el color de ojos o el grupo sanguíneo). Dado que la mayoría de los mamíferos son diploides, poseen dos juegos de cromosomas, uno de ellos procedente del padre y el otro de la madre. Cada par de alelos se ubica en igual locus o lugar del cromosoma.

Locus: Es una posición fija en un cromosoma, que determina la posición de un gen o de un marcador (marcador genético). En biología, y, por extensión, en computación evolutiva, se le usa para identificar posiciones de interés sobre determinadas secuencias. Una variante de la secuencia del ADN en un determinado locus se llama alelo. La lista ordenada de locus conocidos para un genoma particular se denomina mapa genético, mientras que se denomina cartografía genética al proceso de determinación del locus de un determinado carácter biológico. Las células diploides y poliploides cuyos cromosomas tienen el mismo alelo en algún locus se llaman homocigotos, mientras que los que tienen diferentes alelos en un locus, heterocigotos.

Haplotipo: Es una combinación de alelos de diferentes locus de un cromosoma que son transmitidos juntos. Un haplotipo puede ser un locus, varios loci, o un cromosoma entero dependiendo del número de eventos de recombinación que han ocurrido entre un conjunto dado de loci. En un segundo significado, un haplotipo es un conjunto de polimorfismos de un solo nucleótido (SNPs) en un cromosoma particular que están estadísticamente asociados. A los efectos de la presente memoria de patente los términos “haplotipo” y “bloque haplotípico” se consideran sinónimos y se utilizan indistintamente.

Haplotipo HLA *de novo*: Haplotipo HLA que pertenece a una familia que no está contenida en la versión de la base de datos de referencia que contiene secuencias genómicas de los alelos HLA que se haya utilizado (por ejemplo, la base de datos la base de datos de referencia IPD-IMGT/HLA (https://www.ebi.ac.uk/ipd/imgt/hla/)).

Índices P7 y P5: Secuencia de nucleótidos incluida en los oligonucleótidos de amplificación que permiten la identificación inequívoca de una muestra cuando se secuencian varias muestras a la vez.

DESCRIPCIÓN DETALLADA DE LA INVENCIÓN

La presente invención proporciona un método para identificar genotípicamente ambos alelos de al menos un locus de un gen HLA de un sujeto, que comprende:

(a) amplificar por PCR de largo alcance muestras de ADN genómico de dicho sujeto utilizando al menos un par de cebadores sentido y antisentido, para cada muestra, donde dicho par de cebadores, sentido y antisentido, está identificado por las secuencias seleccionadas del grupo que consiste en SEQ ID NO: 1-2, SEQ ID NO: 3-4, SEQ ID NO: 5-6, SEQ ID NO: 7-8, SEQ ID NO: 9-10, SEQ ID NO: 11-12, SEQ ID NO: 13-14, SEQ ID NO: 15-16, SEQ ID NO: 17-18, SEQ ID NO: 19-20 y SEQ ID NO: 21-22, donde dicho par de cebadores amplifica en un único amplicón al menos un gen HLA completo junto con regiones 5'-UTR y 3'-UTR de dicho gen HLA, una región corriente arriba y una región corriente abajo de dicho gen HLA, de ambos alelos, de al menos un locus de un gen HLA, formando así amplicones, donde dicho gen HLA está seleccionado del grupo que consiste en HLA-A, HLA-B, HLA-C, HLA-DRB1, HLA-DRB3, HLA-DRB4, HLA-DRB5, HLA-DPB1, HLA-DQB1, HLA-DPA1 y HLA-DQA1,

(b) combinar los amplicones de cada una de las muestras,

(c) fragmentar los amplicones, obteniendo así fragmentos y etiquetar dichos fragmentos, formando así librerías de fragmentos etiquetados,

(d) combinar dichas librerías de fragmentos etiquetados,

(e) secuenciar dichas librerías de fragmentos etiquetados por secuenciación masiva, generando así lecturas de secuencias que solapan parcialmente,

(f) alinear dichas lecturas de secuencias contra un genoma de referencia,

(g) identificar variantes en base a una llamada de variantes,

(h) determinar bloques haplotípicos en base al alineamiento de las lecturas de secuencias contra un genoma de referencia y añadir regiones flanqueantes de dichos bloques haplotípicos a dichos bloques haplotípicos,

(i) alinear localmente las lecturas de secuencias de los bloques haplotípicos obtenidos en la etapa anterior contra todos los alelos de una base de datos de referencia de secuencias genómicas que codifican genes HLA para cada gen HLA, obteniendo así unos primeros alelos candidatos,

(j) pre-calcular las distancias genéticas y determinar la agrupación filogenética de las familias de HLA a diferentes niveles en la base de datos de referencia de secuencias genómicas que codifican genes HLA, obteniendo así una matriz de distancia que contiene los valores de distancia de cada alelo consigo mismo y con los demás,

(k) llevar a cabo una selección primaria de dichos primeros alelos candidatos en base al resultado de contraste de homología global de las lecturas de secuencias a los haplotipos, obteniendo así unos segundos alelos candidatos,

(l) llevar a cabo un estudio de filogenia evolutiva de dichos segundos alelos candidatos, obteniendo así un árbol filogenético, acceder a la matriz de distancia obtenida en la etapa (j) y seleccionar alelos HLA de la base de datos de referencia de secuencias genómicas que codifican genes HLA más cercanos filogenéticamente, obteniendo así alelos HLA contrastados y

(m) llevar a cabo un test de contraste con alelos HLA contrastados, calcular las distancias entre las lecturas de secuencias y los alelos HLA contrastados en cada nivel y asignar cada lectura de secuencias a una familia HLA conocida a diferentes niveles, donde se asigna una lectura de secuencia a una familia HLA conocida a diferentes niveles cuando se obtiene una significancia estadística en dicho test de contraste por encima de un valor de corte definido y donde se asigna una lectura de secuencia como secuencia *de novo* que codifica para un nuevo locus del gen HLA cuando se obtiene una significancia estadística en dicho test de contraste por debajo de un valor de corte definido.

En la Tabla 1 se especifican los genes HLA que amplifican los pares de cebadores sentido y antisentido identificados por SEQ ID NO: 1-22 y se proporciona información sobre dichos cebadores.

Tabla 1. Cebadores, identificados por SEQ ID NO: 1-22, que amplifican los genes HLA

		SEQ ID NO	Longitud	% guanina-citosina (GC)	UTR	Corriente arriba (pb)	Corriente abajo (pb)
HLA-A	Sentido	1	26	46	Completa	774	
	Antisentido	2	28	43	Completa		1291
HLA-B	Sentido	3	21	47.62	Completa		1827
	Antisentido	4	21	47.62	Completa	1027	
HLA-C	Sentido	5	22	45.5	Completa		638
	Antisentido	6	22	50	Completa	510	
HLA-DRB1	Sentido	7	23	35	Completa		868
	Antisentido	8	29	38	Completa	505	
HLA-DRB3	Sentido	9	22	50	Completa	1982	
	Antisentido	10	23	39	Completa		1122
HLA-DRB4	Sentido	11	20	55	Completa	1320	
	Antisentido	12	22	41	Completa		872
HLA-DRB5	Sentido	13	26	38	Completa	591	
	Antisentido	14	24	42	Completa		297
HLA-DPB1	Sentido	15	24	58	Completa		535
	Antisentido	16	26	50	Completa	1434	
HLA-DQB1	Sentido	17	23	43	Completa		699
	Antisentido	18	21	48	Completa	1204	
HLA-DPA1	Sentido	19	25	44	Completa		1350
	Antisentido	20	20	55	Completa	1779	
HLA-DQA1	Sentido	21	22	50	Completa		2811
	Antisentido	22	20	55	Completa	3188	

En una realización del método, después de secuenciar dichas librerías de fragmentos etiquetados por secuenciación masiva en la etapa (e), se determinan los valores de calidad de los datos en bruto de la secuenciación, donde dicha determinación es una determinación de la calidad de los datos en bruto. Los valores de calidad son los valores

de probabilidad de error por base (*quality values*). Estos valores de calidad pueden presentarse bajo una escala tipo Phred donde un valor de 20 significa que existe la probabilidad de 1/100 de que la base haya sido producto de error.

En una realización del método, después alinear dichas lecturas de secuencias contra un genoma de referencia en la etapa (f), se lleva a cabo un filtrado, en el que se eliminan secuencias que introducen sesgos y ruido.

En una realización del método, después alinear dichas lecturas de secuencias contra un genoma de referencia en la etapa (f), se eliminan secuencias de baja calidad y secuencias que corresponden a duplicados de PCR.

En una realización del método, la identificación de variantes en base a una llamada de variantes de la etapa (g) se lleva a cabo en base al análisis de resultados de los alineamientos de las lecturas de secuencias. En dicha idenfificación de variantes, los desajustes (*mismatches*) identificados entre las lecturas de secuencias y el genoma de referencia se analizan para identificar variantes reales de la muestra. En la identificación de variantes se puede utilizar el algoritmo GATK (*Genome Analysis Tool Kit*) del Broad Institute (https://software.broadinstitute.org/gatk/) (McKenna, A. et al. (2010)).

En una realización del método, en la determinación de bloques haplotípicos en base al alineamiento de las lecturas de secuencias contra un genoma de referencia de la etapa (h) se determinan los fragmentos de las secuencias de los alelos a determinar que tengan unas variantes concretas, discerniendo entre las distintas secuencias que tienen cada uno de los alelos en caso de ser heterocigoto o, en el caso de homocigosis, una sola secuencia con sus variantes con respecto a la referencia. La determinación de bloques haplotípicos está basada en el alineamiento de las lecturas contra los genes de referencia y la identificación de variantes sobre esta secuencia, determinadas previamente.

En una realización del método, en la adición de regiones flanqueantes de dichos bloques haplotípicos a dichos bloques haplotípicos de la etapa (h), dichas regiones flanqueantes son homocigotas. La adición de regiones flanqueantes homocigotas facilita el proceso de identificación del alelo. Al aumentar el tamaño de la secuencia se reduce la ambigüedad.

En una realización del método, el pre-cálculo de las distancias genéticas y determinación de la agrupación filogenética de las familias de HLA a diferentes niveles en la base de datos de referencia de secuencias genómicas que codifican genes HLA de la etapa (j) se lleva a cabo calculando distancias genéticas y su relación evolutiva usando el método *neihgbor-joining* (Saitou, N. et al. (1987)). El resultado final de esta etapa es una matriz

de distancia de tamaño NxN, donde se establecen las distancias filogenéticas de todas las secuencias frente a todas y donde N es el número de alelos que tiene la base de datos de alelos HLA de referencia (p. ej. la base de datos IPD-IMGT/HLA ((https://www.ebi.ac.uk/ipd/imgt/hla/)). La matriz de distancia es una matriz que contiene los valores de distancia de cada alelo consigo mismo y con los demás.

En una realización del método, después de la etapa (j) de pre-cálculo de las distancias genéticas y determinación de la agrupación filogenética de las familias de HLA a diferentes niveles en la base de datos de referencia de secuencias genómicas que codifican genes HLA, se evalúa la calidad de las familias desde el punto de vista filogenético usando el método silueta (Rousseeuw, P. J. (1987)). El método silueta permite evaluar si las distancias o relaciones filogenéticas entre los miembros de una misma familia son estadísticamente más estrechas respecto al resto de familias de la base de datos. El valor de silueta oscila entre -1 y 1, cuanto más próximo es al 1 más definido es el grupo. Normalmente, valores por encima de 0.5 se consideran grupos aceptables.

En una realización del método, en la etapa (k) se lleva a cabo una selección primaria de dichos primeros alelos candidatos en base al resultado de contraste de homología global de las lecturas de secuencias a los haplotipos, obteniendo así unos segundos alelos candidatos y en dicho contraste de homología global se utiliza el algoritmo fasta36 (Pearson, W. R. (2016)). El contraste global permite evaluar la cercanía del haplotipo/s identificado/s usando toda la longitud como unidad, evitando la similitud por dominios o perfiles menores. La selección de un pequeño grupo de posibles HLA, los segundos alelos candidatos, se basa en una combinación de parámetros, como son los que definen la homología, el porcentaje de identidad y la longitud del alineamiento.

En una realización del método, en la etapa (l) se lleva a cabo un estudio de filogenia evolutiva de dichos segundos alelos candidatos, obteniendo así un árbol filogenético, se accede a la matriz de distancia obtenida en la etapa (j) y se seleccionan alelos HLA de la base de datos de referencia de secuencias genómicas que codifican genes HLA más cercanos filogenéticamente, obteniendo así alelos HLA contrastados. En dicha etapa (l), en el estudio de filogenia evolutiva de dichos segundos alelos candidatos se utiliza la plataforma R y la librería phangorn (Schliep K. P. (2011); Schliep, K. P. et al. (2017)). En dicha etapa (l), se parte de los segundos alelos candidatos y el haplotipo o haplotipos asociados a las muestras problema, se realiza un cálculo de distancia genética en un contexto “todos contra todos”, usando como modelo para describir la distancia genética la estrategia F81 (Li, H. et al. (2009)), se genera un árbol filogenético usando la estrategia

neighbor-joining (Felsenstein, J. (1981); Saitou, N. et al. (1987); Studier, J. A. et al. (1988)) y se accede a la matriz de distancias utilizando el método cophenetic (Gascuel, O. (1997)).

En una realización del método, en la etapa (m) se lleva a cabo un test de contraste con alelos HLA contrastados, se calculan las distancias entre las lecturas de secuencias y los alelos HLA contrastados en cada nivel y se asigna cada lectura de secuencias a una familia HLA conocida a diferentes niveles, donde se asigna una lectura de secuencia a una familia HLA conocida a diferentes niveles cuando se obtiene una significancia estadística en dicho test de contraste por encima de un valor de corte definido y donde se asigna una lectura de secuencia como secuencia *de novo* que codifica para un nuevo locus del gen HLA cuando se obtiene una significancia estadística en dicho test de contraste por debajo de un valor de corte definido. Dicho test de contraste comienza con el alelo HLA contrastado, que es el alelo HLA más cercano filogenéticamente que se obtiene de la base de datos IPD-IMGT/HLA (https://www.ebi.ac.uk/ipd/imgt/hla/) y que se obtiene a partir del estudio de filogenia evolutiva y el haplotipo problema. Posteriormente, la asignación de la familia y los diferentes niveles del haplotipo problema se igualan al HLA contrastado. Posteriormente, se recorre desde el nivel superior al inferior de la secuencia del HLA contrastado. En cada nivel se evalua la distancia media, mediana y los intervalos de confianza de cada distribución de distancia de los miembros de la familia con nivel idéntico. Si la distancia entre el haplotipo problema y HLA contrastado más cercano se encuentra dentro de la distribución teórica del nivel estudiado con una significancia de 95%, se considera que pertenece a ese nivel de manera estadísticamente significativa (Sneath, P.H.A. et al. (1973); Bauer, D. F. (1972)). El test termina con una identificación estadística del haplotipo problema con los diferentes niveles. Se considera que el haplotipo problema era el alelo contrastado si los 3 niveles superiores son estadísticamente significativos. Si esa premisa no se cumple, el proceso se repite con la segunda HLA de mayor homología, y así recurrentemente hasta terminar todos los alelos HLA obtenidos en la etapa (i).

En una realización del método, dicha significancia estadística se calcula en base al valor estadístico p.

En una realización del método, dicho valor de corte definido se calcula en base al valor estadístico p y tiene un valor de 1^{-6}.

En una realización del método, la longitud de los fragmentos generados en la etapa (c) es entre 1200 y 1500 pb.

En la presente memoria, el término “pb” hace referencia a “pares de bases” y ambos términos son intercambiables.

En una realización del método, las librerías se secuencian en la etapa (e) por secuenciación masiva con una longitud de lectura entre 200 y 300 pb.

La invención también proporciona un par de cebadores identificado por las secuencias seleccionadas del grupo que consiste en SEQ ID NO: 1-2, SEQ ID NO: 3-4, SEQ ID NO: 5-6, SEQ ID NO: 7-8, SEQ ID NO: 9-10, SEQ ID NO: 11-12, SEQ ID NO: 13-14, SEQ ID NO: 15-16, SEQ ID NO: 17-18, SEQ ID NO: 19-20 y SEQ ID NO: 21-22, para uso en la identificación genotípica de ambos alelos de al menos un locus de un gen HLA de un sujeto, donde dicho par de cebadores amplifica en un único amplicón un gen HLA completo junto con regiones 5’-UTR y 3’-UTR de dicho gen HLA, una región corriente arriba y una región corriente abajo de dicho gen HLA, de ambos alelos de al menos un locus de un gen HLA, donde dicho gen HLA está seleccionado, respectivamente, del grupo que consiste en HLA-A, HLA-B, HLA-C, HLA-DRB1, HLA-DRB3, HLA-DRB4, HLA-DRB5, HLA-DPB1, HLA-DQB1, HLA-DPA1 y HLA-DQA1.

La invención también proporciona un kit para identificar genotípicamente ambos alelos de al menos un locus de un gen HLA de un sujeto, que comprende:

(a) al menos un par de cebadores, sentido y antisentido, identificado por las secuencias seleccionadas del grupo que consiste en SEQ ID NO: 1-2, SEQ ID NO: 3-4, SEQ ID NO: 5-6, SEQ ID NO: 7-8, SEQ ID NO: 9-10, SEQ ID NO: 11-12, SEQ ID NO: 13-14, SEQ ID NO: 15-16, SEQ ID NO: 17-18, SEQ ID NO: 19-20 y SEQ ID NO: 21-22, donde dicho par de cebadores amplifica en un único amplicón un gen HLA completo junto con regiones 5’-UTR y 3’-UTR de dicho gen HLA, una región corriente arriba y una región corriente abajo de dicho gen HLA, de ambos alelos de al menos un locus del gen HLA, donde dicho gen HLA está seleccionado del grupo que consiste en HLA-A, HLA-B, HLA-C, HLA-DRB1, HLA-DRB3, HLA-DRB4, HLA-DRB5, HLA-DPB1, HLA-DQB1, HLA-DPA1 y HLA-DQA1,

(b) una ADN polimerasa e

(c) instrucciones de uso.

En una realización, el kit comprende adicionalmente reactivos para amplificación y secuenciación.

En una realización, el kit es para uso en la identificación genotípica de ambos alelos de al menos un locus de un gen HLA, donde dicho gen HLA está seleccionado del grupo que

consiste en HLA-A, HLA-B, HLA-C, HLA-DRB1, HLA-DRB3, HLA-DRB4, HLA-DRB5, HLA-DPB1, HLA-DQB1, HLA-DPA1 y HLA-DQA1.

La invención también proporciona un producto informático que comprende un medio legible por ordenador, en el que hay codificadas instrucciones para controlar un sistema informático que comprende los medios siguientes para llevar a cabo una identificación genotípica de ambos alelos de al menos un locus de un gen HLA de un sujeto, donde dicha identificación genotípica comprende amplificar por PCR de largo alcance muestras de ADN genómico de dicho sujeto utilizando al menos un par de cebadores sentido y antisentido, para cada muestra, donde dicho par de cebadores, sentido y antisentido, está identificado por las secuencias seleccionadas del grupo que consiste en SEQ ID NO: 1-2, SEQ ID NO: 3-4, SEQ ID NO: 5-6, SEQ ID NO: 7-8, SEQ ID NO: 9-10, SEQ ID NO: 11-12, SEQ ID NO: 13-14, SEQ ID NO: 15-16, SEQ ID NO: 17-18, SEQ ID NO: 19-20 y SEQ ID NO: 21-22, donde dicho par de cebadores amplifica en un único amplicón al menos un gen HLA completo junto con regiones 5'-UTR y 3'-UTR de dicho gen HLA, una región corriente arriba y una región corriente abajo de dicho gen HLA, de ambos alelos, de al menos un locus de un gen HLA, formando así amplicones; donde dicho gen HLA está seleccionado del grupo que consiste en HLA-A, HLA-B, HLA-C, HLA-DRB1, HLA-DRB3, HLA-DRB4, HLA-DRB5, HLA-DPB1, HLA-DQB1, HLA-DPA1 y HLA-DQA1, combinar los amplicones de cada una de las muestras, fragmentar los amplicones, obteniendo así fragmentos y etiquetar dichos fragmentos, formando así librerías de fragmentos etiquetados, combinar dichas librerías de fragmentos etiquetados, secuenciar dichas librerías de fragmentos etiquetados por secuenciación masiva, generando así lecturas de secuencias que solapan parcialmente y donde dicho sistema comprende los medios siguientes:

(a) medios para llevar a cabo un alineamiento de dichas lecturas de secuencias contra un genoma de referencia,

(c) medios para identificar variantes en base a una llamada de variantes,

(d) medios para determinar bloques haplotípicos en base al alineamiento de las lecturas de secuencias contra un genoma de referencia y medios para añadir regiones flanqueantes de dichos bloques haplotípicos a dichos bloques haplotípicos,

(e) medios para alinear localmente las lecturas de secuencias de los bloques haplotípicos obtenidos en la etapa anterior contra todos los alelos de una base de datos de referencia de secuencias genómicas que codifican genes HLA para cada gen HLA, obteniendo así unos primeros alelos candidatos,

(f) medios para pre-calcular las distancias genéticas y determinar la agrupación filogenética de las familias de HLA a diferentes niveles en la base de datos de referencia de secuencias genómicas que codifican genes HLA, obteniendo así una matriz de distancia que contiene los valores de distancia de cada alelo consigo mismo y con los demás,

(g) medios para llevar a cabo una selección primaria de dichos primeros alelos candidatos en base al resultado de contraste de homología global de las lecturas de secuencias a los haplotipos, obteniendo así unos segundos alelos candidatos,

(h) medios para llevar a cabo un estudio de filogenia evolutiva de dichos segundos alelos candidatos, obteniendo así un árbol filogenético, medios para acceder a la matriz de distancia obtenida en la etapa (f) y medios para seleccionar alelos HLA de la base de datos de referencia de secuencias genómicas que codifican genes HLA más cercanos filogenéticamente, obteniendo así alelos HLA contrastados,

(i) medios para llevar a cabo un test de contraste con alelos HLA contrastados y

(j) medios para calcular las distancias entre las lecturas de secuencias y los alelos HLA contrastados en cada nivel y medios para asignar cada lectura de secuencias a una familia HLA conocida a diferentes niveles, donde se asigna una lectura de secuencia a una familia HLA conocida a diferentes niveles cuando se obtiene una significancia estadística en dicho test de contraste por encima de un valor de corte definido y donde se asigna una lectura de secuencia como secuencia *de novo* que codifica para un nuevo locus del gen HLA cuando se obtiene una significancia estadística en dicho test de contraste por debajo de un valor de corte definido.

En una realización del producto informático, dicha significancia estadística se calcula en base al valor estadístico p.

En una realización del producto informático, dicho valor de corte definido se calcula en base al valor estadístico p y tiene un valor de 1^{-6}.

Secuenciación masiva

Es posible utilizar cualquier método de secuenciación adecuado para llevar a cabo los métodos descritos en este documento. En algunas realizaciones, se usa un método de secuenciación de alto rendimiento. En los métodos de secuenciación de alto rendimiento generalmente las secuencias de ADN se secuencian de forma masiva y paralela dentro de una celda de flujo. El carril (lane) es la unidad básica física en una plataforma de

secuenciación masiva y se refiere a la ejecución independiente básica en dicha plataforma de secuenciación masiva. Las tecnologías de secuenciación de alto rendimiento la secuenciación por síntesis.

Los sistemas utilizados para métodos de secuenciación de alto rendimiento están disponibles comercialmente e incluyen, por ejemplo, las plataformas MiniSeq®, MiSeq®, NextSeq® 550, HiSeq® 2000, HiSeq® 2500, HiSeq® 3000, HiSeq® 4000 y NovaSeq® de Illumina, Inc y las plataformas de secuenciación ION Proton y ION PGM de ThermoFisher entro otros.

En la secuenciación por síntesis, millones de fragmentos de ácido nucleico (por ejemplo, ADN) pueden secuenciarse en paralelo. En un ejemplo de este tipo de tecnología de secuenciación, se usa una celda de flujo que contiene un portaobjetos ópticamente transparente con 2 carriles individuales sobre cuyas superficies se unen oligonucleótidos (por ejemplo, oligonucleótidos adaptadores). Una celda de flujo a menudo es un soporte sólido que se puede configurar para retener y/o permitir el paso ordenado de las soluciones de reactivo sobre los analitos unidos. Las celdas de flujo con frecuencia tienen una forma plana, ópticamente transparente, generalmente en escala milimétrica o submilimétrica, y a menudo tienen canales o carriles en los que se produce la interacción analito/reactivo. En ciertos procedimientos de secuenciación por síntesis, por ejemplo, el ADN molde se fragmenta en longitudes de varios cientos de pares de bases en preparación para la generación de la librería. En ciertas realizaciones, el aislamiento de la muestra y la generación de la librería se realizan usando métodos y aparatos automatizados.

En ciertos procedimientos de secuenciación por síntesis, los oligonucleótidos adaptadores son complementarios a oligonucleótidos presentes en las células de flujo, y algunas veces se utilizan para asociar el ADN con un soporte sólido, la superficie interna de una celda de flujo, por ejemplo. En algunas realizaciones, el oligonucleótido adaptador incluye oligonucleótidos de indexación (índices) (por ejemplo, una secuencia única de nucleótidos utilizable como índices para permitir la identificación inequívoca de una muestra), uno o más sitios de hibridación de cebadores de secuenciación (p. ej. cebadores de secuenciación universal, cebadores de secuenciación de extremo único, cebadores de secuenciación de extremos emparejados, cebadores de secuencia multiplexados, y similares), o combinaciones de los mismos (por ejemplo, adaptador/secuenciación, adaptador/índice, adaptador/índice/secuenciación). Los cebadores de indexación a menudo tienen seis o más nucleótidos de longitud. En ciertas realizaciones, los índices están asociados con una muestra, pero son secuenciados en

una reacción de secuenciación separada para evitar comprometer la calidad de las lecturas de secuencia. Posteriormente, las lecturas de la secuencia de los índices y la secuencia de las muestras se vinculan entre sí y las lecturas se desmultiplexan.

En ciertos procedimientos de secuenciación por síntesis, la utilización de oligonuclétidos adaptadores permite la multiplexación de reacciones de secuencia en un carril de una celda de flujo, lo que permite el análisis de múltiples muestras por carril de la celda de flujo. El número de muestras que se pueden analizar en un carril de celda de flujo dado a menudo depende del número de índices únicos utilizados durante la preparación de la librería. El número de índices utilizados no está limitado en los métodos descritos en este documento y dichos métodos pueden realizarse usando cualquier número de índices (p. ej., 4, 8, 12, 20, 24, 48 o 96). Cuanto mayor es el número de índices, mayor es el número de muestras que se pueden multiplexar en un solo carril de celda de flujo. La multiplexación usando 12 índices P7 y 8 índices P5 permite analizar 96 muestras (p. ej., igual a la cantidad de pocillos en una placa de 96 pocillos) en una celda de flujo de 2 carriles.

En ciertos procedimientos de secuenciación por síntesis, se añade ADN molde modificado con adaptadores a la celda de flujo y se inmoviliza mediante hibridación con los oligonucleótidos de anclaje presentes en la celda de flujo en condiciones de dilución limitante. Múltiples ciclos de amplificación convierten el molde de ADN de una sola molécula en un grupo (*cluster*) amplificado clonalmente. Para la secuenciación, los grupos se desnaturalizan. La secuenciación se inicia hibridando un cebador complementario a las secuencias adaptadoras, seguido de la adición de polimerasa y una mezcla de cuatro terminadores de sondas reversibles fluorescentes de diferentes colores. Los terminadores se incorporan de acuerdo con la complementariedad de secuencia en cada cadena en un clúster clonal. Después de la incorporación, los reactivos en exceso se eliminan por lavado, los grupos se analizan ópticamente y se registra la señal de fluorescencia. Con pasos químicos sucesivos, los terminadores de sondas reversibles se desbloquean, las etiquetas fluorescentes se escinden y eliminan por lavado, y se realiza el siguiente ciclo de secuenciación.

BREVE DESCRIPCIÓN DE LOS DIBUJOS

Figura 1. Perfil de fragmentos de la librería indexada. UF son unidades de fluorescencia.

Figura 2. Diagrama de flujo de las etapas del análisis bioinformático de acuerdo con la invención.

Figura 3. Distribución de valores de Silueta (*Silhouette*) para diferentes genes de histocompatibilidad.

Figura 4. Perfil de tamaños obtenido para las librerías de las muestras VP346 (curva 3), VP347 (curva 2), VP348 (curvas 4 y 5) y VP349 (curva 1). UF son unidades de fluorescencia.

Figura 5. Gráfico de calidad obtenido para las lecturas de secuencias obtenidas con la combinación de las librerías de las muestras VP346, VP347, VP348 y VP349 secuenciadas en Hiseq 2500 Rapid Mode® utilizando una longitud de lectura de 100 pb. El 93.8% de las lecturas de secuencias tiene un valor Q (Q *score*) superior a 30. Se muestra una línea vertical a un valor Q igual a 30.

Figura 6. Gráfico de calidad obtenido para las lecturas de secuencias obtenidas con la combinación de las librerías de las muestras VP346, VP347, VP348 y VP349 secuenciadas en Miseq® con longitud de lectura de 250 pb. El 89.4% de las lecturas de secuencias tiene un valor Q (Q *score*) superior a 30. Se muestra una línea vertical a un valor Q igual a 30.

Figura 7. Ensayo de selección de tamaño de librería con las librerías. de las muestras VP347 y VP348. Se muestran los resultados obtenidos con el aparato TapeStation® 4200 de Agilent®. Se muestran los resultados para una relación de Ampure XP beads® sobre el resultado de PCR en el proceso de obtención de librerías de 0.7X de la librería de la muestra VP347 (A7) y de la muestra VP348 (B7), los resultados para una relación de Ampure XP beads® sobre el resultado de PCR en el proceso de obtención de librerías de 0.6X de la librería de la muestra VP347 (C7) y de la muestra VP348 (D7) y los resultados para una relación Ampure XP beads® sobre el resultado de PCR en el proceso de obtención de librerías de 0.5X de la librería de la muestra VP347 (E7) y de la muestra VP348 (F7).

Figura 8. Gráfico de calidad obtenido para para las lecturas de secuencias obtenidas con la combinación de las librerías ensayadas en la Figura 7 secuenciadas en MiSeq® con una longitud de lectura de 250 pb. El 91.2% de las lecturas de secuencias tiene un valor Q (Q *score*) superior a 30. Se muestra una línea vertical a un valor Q igual a 30.

Figura 9. Resultado estadístico del contraste entre la secuencia problema (correspondiente a la muestra VP367) y la familia A*25 (A). Se observa que la secuencia problema está dentro del grupo de la familia A*25. El contraste realizado por el Wilcoxon

indica que son las mismas distribuciones al tener un valor p >= que el valor de corte p 1^{-6}. Árbol filogenético de la secuencia problema con la familia A*25 (B)

Figura 10. Resultado estadístico del contraste entre la secuencia problema (correspondiente a la muestra VP367) y la subfamilia A*25:01 (A). Se observa que la secuencia problema está dentro del grupo de la subfamilia A*25:01. El contraste realizado por el Wilcoxon indica que son las mismas distribuciones al tener un valor p (pValue) >= que el valor de corte p 1^{-6}. Árbol filogenético de la secuencia problema con la subfamilia A*25:01 (B).

Figura 11. Resultado estadístico del contraste entre la secuencia problema (correspondiente a la muestra VP367) y la subfamilia A*25:01:01 (A). Se observa que la secuencia problema está dentro del grupo de la subfamilia A*25:01:01. El contraste realizado por el Wilcoxon indica que son las mismas distribuciones al tener un valor p (pValue) >= que el valor de corte p 1^{-6}. Árbol filogenético de la secuencia problema con la subfamilia A*25:01:01 (B).

Figura 12. Resultado estadístico del contraste entre la secuencia problema (correspondiente a la muestra VP367) y la subfamilia A*26. Se observa que la secuencia problema está fuera del grupo de la familia A*26. El contraste de distribuciones indica que no pertenece a la familia con un valor p (pValue) de 2^{-16}, inferior al valor de corte p 1^{-6}, por tanto, según las distancias, la secuencia problema no pertenece a la familia A*26.

Figura 13. Resultado estadístico del contraste entre la secuencia problema (correspondiente a la muestra VP371) y la familia C*12. Se puede observar como la secuencia problema no está dentro del grupo de la familia C*12. El contraste realizado por el Wilcoxon indica que no son las mismas distribuciones al tener un valor p (pValue) de 2^{-16} inferior al valor de corte p 1^{-6}.

Figura 14. Resultado estadístico del contraste entre la secuencia problema (correspondiente a la muestra VP371) y la familia C*12:03. Se observa que la secuencia problema no está dentro del grupo de la subfamilia C*12:03. El contraste realizado por el Wilcoxon indica que no son las mismas distribuciones al tener un valor p (pValue) de 2^{-16} inferior al valor de corte p 1^{-6}.

DESCRIPCIÓN DE MODOS DE REALIZACIÓN

Ejemplo 1. Muestras analizadas

Se analizaron 4 muestras de referencia, proporcionadas por parte del grupo internacional denominado International Histocompatibility Working Group: IHW09028, IHW01038, IHW09273 e IHW09014 que corresponden a las muestras VP346, VP347, VP348 y VP349, respectivamente. Los tipajes HLA de todas estas muestras eran conocidos antes de la realización del ensayo.

También se analizaron muestras muestra clínica de pacientes reales anonimizadas cuyo tipaje HLA era conocido previamente a la realización del ensayo (muestras VP367, VP371 y VP373).

Se analizaron muestras de ADN genómico. El ADN genómico estaba libre de ARN y contaminantes como fenol, alcohol, EDTA o exceso de sales. Se comprobó la relación de absorbancias 260/280 nm y 260/230 nm y la absorbancia a 340 nm.

Ejemplo 2. Amplificación por PCR de largo alcance (*Long Range PCR*)

Se amplificaron las regiones de interés mediante PCR de largo alcance a partir del ADN genómico (gADN). Posteriormente, se verificaron los productos de amplificación mediante electroforesis y se purificaron los amplicones.

Se prepararon las reacciones de PCR de largo alcance de los 11 amplicones de los genes HLA-A, HLA-B, HLA-C, HLA-DRB1, HLA-DRB3, HLA-DRB4, HLA-DRB5, HLA-DPB1, HLA-DQB1, HLA-DPA1 y HLA-DQA1 siguiendo las condiciones descritas en las Tablas 2 y 3. La Tabla 2 indica si el tipo de mezcla de reacción es “A” o “B” y la Tabla 3 especifica las mezclas de reacción para estos dos tipos de mezcla de reacción.

Tabla 2. Condiciones de amplificación para los 11 amplicones

Mezcla de cebadores	Tamaño del amplicón	Tipo de mezcla de reacción	Programa PCR	ng de partida de gADN
I	5478	A	1	100
II	6234	A	3	100
III	4536	B	1	100
IV	14267	B	3	100
V	12286	B	1	100
VI	11373	B	1	100
VII	9300	A	3	100
VIII	12437	A	3	100
IX	16360	B	3	100
X	15075	B	2	100
XI	13744	A	3	100

Se prepararon las 11 reacciones de PCR de largo alcance en hielo en tubos, utilizando los volúmenes y cantidades descritas en la Tabla 3 y se mezcló.

Tabla 3. Mezclas de reacción para PCR de largo alcance

	Volumen para 1 reacción	
Reactivos	Mezcla de reacción A	Mezcla de reacción B
Amplicon PCR Buffer 5X (PrimeSTAR® GXL DNA Polymerase (Takara Bio Inc).	10 µl	10 µl
dNTPs-AMP	4 µl	4 µl
DNA Polimerasa	1 µl	1 µl
I-XI (20 µM)	0,5 µl	0,5 µl
Betaina	-	13 µl
gADN	100 ng	100 ng
Agua libre de nucleasas	Hasta 50 µl	Hasta 50 µl

Se pusieron los tubos en el termociclador y se ejecutaron los programas indicados en la Tabla 4 para cada amplicón.

Tabla 4. Programa para la amplificación de los amplicones

Programa para la amplificación de los amplicones I, III, V, VI y IX				
Programa	Paso	Ciclos #	Temperatura	Tiempo
3	1	30	98°C	10 s
			68°C	10 min
	2	1	4°C	*Mantener*
Programa para la amplificación del amplicón X				
Programa	Paso	Ciclos #	Temperatura	Tiempo
3	1	30	98°C	10 s
			60°C	15 s
			68°C	10 min
	2	1	4°C	*Mantener*
Programa para la amplificación de los amplicones II, IV, VII y XI				
Programa	Paso	Ciclos #	Temperatura	Tiempo
3	1	30	98°C	10 s
			55°C	15 s
			68°C	10 min
	2	1	4°C	*Mantener*

Una vez finalizado el programa del termociclador, se dejaron las muestras en hielo.

Control de amplificación y purificación de los amplicones

Los productos de amplificación se comprobaron mediante una electroforesis en gel de agarosa a 0,8%. Para realizar la electroforesis, se utilizaron marcadores de peso

molecular Low DNA Mass Ladder (Invitrogen) Low-Mass y Lambda DNA/HindIII Marker (Thermo Fisher), y se utilizó un tampón de carga BPB (Bromophenol Blue sodium salt (Sigma-Aldrich)). Se purificó cada uno de los amplicones de manera independiente. Se verificó nuevamente mediante electroforesis en gel de agarosa el resultado de la purificación y se comprobó que los cebadores de la muestra se habían eliminado completamente. Se cuantificó la concentración de cada uno de los amplicones.

Ejemplo 3. Combinación de los amplicones

Se preparó una combinación equimolar de 1 µg por muestra. Según el tamaño de cada amplicón, se utilizó una cantidad determinada de cada purificación de amplificación, según las cantidades indicadas en la Tabla 5. La combinación se preparó con los amplicones de cada una de las muestras de manera independiente.

Tabla 5. Preparación de una combinación equimolar de amplicones purificados

Amplicón	Tamaño (pb)	ng para 1 µg de pool equimolar
I	5478	45,2
II	6234	51,5
III	4536	37,5
IV	14267	117,8
V	12286	101,5
VI	11373	93,9
VII	9300	76,8
VIII	12437	102,7
IX	16360	135,1
X	15075	124,5
XI	13744	113,5
Total	121090	1000

Se comprobó mediante electroforesis que no existían restos de cebadores en el producto final y se purificó el producto final mediante columnas Zymo-Spin® IC-XL (Zymo-Research) siguiendo las instrucciones del fabricante para purificaciones de productos de PCR. Se comprobó la relación de absorbancias 260/280 nm y 260/230 nm y la absorbancia a 340 nm. Se cuantificó la concentración final de la combinación.

Ejemplo 4. Tagmentación de los amplicones

El ADN amplificado se fragmentó enzimáticamente con la enzima tagmentasa (una especie de trasposasa), que fragmenta aleatoriamente el DNA en fragmentos de tamaño

deseado mediante incubación a 45ºC durante 10 minutos. Simultáneamente la enzima añade en los extremos del DNA fragmentado unas secuencias específicas que permitirán la incorporación de los adaptadores en el siguiente proceso de amplificación por PCR. Este proceso se conoce como tagmentación, que da lugar a ADN tagmentado. Posteriormente, el ADN tagmentado se purificó utilizando perlas magnéticas Agencourt AMPure XP beads® (Beckman Coulter).

Se prepararon las reacciones de fragmentación en hielo utilizando tubos. Se añadieron 17 µl de tampón TB (GeneSGKit® Tagmentation Buffer (Sistemas Genómicos)) a cada tubo. Se añadieron 2 µl de ADN amplificado a su tubo correspondiente, con la punta de la pipeta en el fondo del tubo. Se añadieron 0,5 µl de tampón TE (GeneSGKit® Tagmentation Enzyme (Sistemas Genómicos)) a cada tubo, con la punta de la pipeta en el fondo del tubo y se mezcló. Se colocaron los tubos en el termociclador y se siguió el programa de termociclado de la Tabla 6.

Tabla 6. Programa de termociclador para la fragmentación del ADN.

Paso	Temperatura	Tiempo
1	45°C	10 minutos
2	4°C	1 minuto
3	4°C	*Mantener*

Se incubó 1 minuto a 4ºC en el termociclador y se transfirieron las muestras al hielo. Se añadieron 32 µl de tampón SS (GeneSGKit® Stop Solution (Sistemas Genómicos) a cada reacción y se mezcló. Se añadieron 52 µl de la solución de Agencourt AMPure XP beads® (Beckman Coulter) a cada tubo que contenía la muestra y se mezcló. Se colocaron los tubos en el soporte magnético y se esperó 5 minutos. Con los tubos en el soporte magnético, se eliminó el sobrenadante con pipeta, evitando tocar las perlas magnéticas Agencourt AMPure XP beads® (Beckman Coulter). Los tubos se lavaron 2 veces con 200 µl de etanol 70% y se eliminó todo el etanol con una pipeta. Se mantuvieron los tubos en el soporte magnético y se dejaron secar con la tapa abierta a temperatura ambiente durante 5 minutos o hasta que todo el etanol residual se evaporó. Se añadieron 11 µl de agua libre de nucleasas a cada tubo, se mezcló y se incubó a temperatura ambiente durante 2 minutos. Se colocaron los tubos en el soporte magnético durante 2-3 minutos, hasta que la solución no contenía perlas magnéticas Agencourt AMPure XP beads® (Beckman Coulter). Se transfirió el sobrenadante (~10 µl) a un nuevo tubo de 0.2 ml y se eliminaron las perlas magnéticas Agencourt AMPure XP beads® (Beckman Coulter).

Ejemplo 5.1. Amplificación e indexado de las librerías. Combinación de las librerías

En este paso, el ADN tagmentado y purificado se amplificó mediante un número reducido de ciclos de PCR. Se utilizó todo el ADN tagmentado. Tras la amplificación, la librería se purificó utilizando perlas magnéticas.

A cada librería individual se le añadió una secuencia de nucleótidos específica (índice) mediante amplificación por PCR, permitiendo mezclar varias librerías para secuenciarlas juntas en un único carril (*lane*). El carril (*lane*) es la unidad básica física en una máquina de secuenciación masiva y se refiere a la ejecución independiente básica en una máquina de secuenciación masiva.

Se mezclaron librerías de forma que cada una tenía un único índice P7. Con esta estrategia de combinación, solo fue necesaria la secuenciación de un índice (P7). Se utilizaron las combinaciones de la Tabla 7. Cada posición produjo señal de fluorescencia en los 2 canales durante la secuenciación del índice. La secuencia de nucleótidos para los índices se muestra en la Tabla 8.

Ejemplo 5.2. Amplificación e indexado de las librerías. Combinación de las librerías.

Se mezclaron librerías de forma que los índices P7 no eran únicos. En este caso, se incluyeron diferentes índices P5. Se utilizaron las combinaciones de la Tabla 6. Cada posición produjo señal de fluorescencia en los 2 canales durante la secuenciación del índice. La secuencia de nucleótidos para los índices se muestra en la Tabla 7.

Tabla 7. Combinaciones de librerías

<table>
<tr><th>Complejidad del pool</th><th>Índices P7 utilizados</th><th>Índices P5 utilizados</th><th>Secuenciación</th></tr>
<tr><td>1-plex (sin pooling)</td><td>Cualquier índice P7 (1-12)</td><td rowspan="6">Cualquier índice P5 (13-20)</td><td>Sin secuenciación del índice</td></tr>
<tr><td>2-plex</td><td>Combinación A: 1 y 2
Combinación B: 2 y 4</td><td rowspan="5">Secuenciación de un único índice (P7)</td></tr>
<tr><td>3-plex</td><td>Combinación A: 1, 2 y 4
Combinación B: 3, 12 y 6</td></tr>
<tr><td>4- o 5-plex</td><td>Combinación A: 1, 2 y 4 y cualquier otro índice P7
Combinación B: 3,12 y 6 y cualquier otro índice P7</td></tr>
<tr><td>6-plex</td><td>1, 2, 3, 4, 6 y 12</td></tr>
<tr><td>7- a 12 -plex</td><td>1, 2, 3, 4, 6, 12 y cualquier otro índice P7</td></tr>
<tr><td>> 12-plex</td><td>1,2,3,4,6,12 y cualquier otro índice P7</td><td>Combinación A: 13, 14 y cualquier otro índice P5
Combinación B: 15, 16 y cualquier otro índice P5
Combinación C: 17, 18 y cualquier otro índice P5</td><td>Secuenciación de 2 índices (P7 y P5)</td></tr>
</table>

Tabla 8. Secuencia de nucleótidos de los índices P7 y P5

Índice P7	Secuencia	Índice P5	Secuencia
1	TAAGGCGA	13	TAGATCGC
2	CGTACTAG	14	CTCTCTAT
3	AGGCAGAA	15	TATCCTCT
4	TCCTGAGC	16	AGAGTAGA
5	GTAGAGGA	17	GTAAGGAG
6	TAGGCATG	18	ACTGCATA
7	CTCTCTAC	19	AAGGAGTA
8	CAGAGAGG	20	CTAAGCCT
9	GCTACGCT		
10	CGAGGCTG		
11	AAGAGGCA		
12	GGACTCCT		

Se utilizaron las combinaciones descritas en la Tabla 9.

Tabla 9. Combinaciones de índices utilizadas

Índice 1	Índice 2
TAAGGCGA	TAGATCGC
TCCTGAGC	CTCTCTAT
GTAGAGGA	TAGATCGC
TAGGCATG	CTCTCTAT
CGTACTAG	CTCTCTAT
AGGCAGAA	TAGATCGC
CTCTCTAC	TAGATCGC
CAGAGAGG	CTCTCTAT
TAAGGCGA	CTCTCTAT

5

Se preparó en hielo el volumen necesario de PCR Master Mix (PCRMM) como se describe en la Tabla 10 y se mezcló.

Tabla 10. PCR Master Mix para la amplificación del ADN tagmentado

Reactivos	Volumen para 1 reacción	Volumen para 12 reacciones
Agua libre de nucleasas	22 µl	277,2 µl
GeneSGKit® 5X PCR Buffer	10 µl	126 µl
GeneSGKit® dNTPs	0,5 µl	6,3 µl
DMSO	2,5 µl	31,5 µl
GeneSGKit® PCR Enzyme	1 µl	12,6 µl
Total	36 µl	453,6 µl

Se añadieron 36 µl de PCRMM a cada tubo que contiene 10 µl de ADN tagmentado y purificado obtenido en los pasos previos. Se añadieron 2 µl de uno de los cebadores GeneSGKit® P7 indexing primers (1-12) a cada mezcla de PCR. Se añadieron 2 µl de uno de los cebadores GeneSGKit® P5 indexing primers (13-20) a cada mezcla de PCR y se mezcló. Se colocaron los tubos en el termociclador y se ejecutó el programa de la Tabla 11.

Tabla 11. Programa de termociclador para amplificar el ADN tagmentado

Paso	Nº ciclos	Temperatura	Tiempo
1	1	68°C	2 minutos
2	1	98°C	30 segundos
3	5	98°C	30 segundos
		56°C	30 segundos
		72°C	1 minuto
4	1	4°C	*Mantener*

Se incubó la solución de perlas magnéticas Agencourt AMPure XP beads® (Beckman Coulter) a temperatura ambiente durante 30 minutos. La solución de perlas magnéticas Agencourt AMPure XP beads® (Beckman Coulter) se mezcló enérgicamente hasta obtener una suspensión homogénea. Se añadieron 30 µl de la solución de Agencourt AMPure XP beads® (Beckman Coulter) a cada tubo que contiene la muestra, se mezcló y se incubó 5 minutos a temperatura ambiente. Se colocaron los tubos en el soporte magnético y se esperó 5 minutos. Se mantuvieron los tubos en el soporte magnético mientras se eliminó el sobrenadante con pipeta, evitando tocar las perlas. Con los tubos

en el soporte magnético, se lavó 2 veces añadiendo 200 µl de etanol 70% a cada tubo. Se dejaron los tubos en el soporte magnético 1 minuto y se eliminó el etanol con ayuda de una pipeta. Se mantuvieron los tubos en el soporte magnético y se secaron con la tapa abierta a temperatura ambiente durante 5 minutos. Se añadieron 30 µl de agua libre de nucleasas a cada tubo, se mezcló y se incubó durante 2 minutos a temperatura ambiente. Se colocaron los tubos en el soporte magnético durante 2-3 minutos, hasta que la solución era clara, sin perlas. Se transfirió el sobrenadante (~30 µl) a un nuevo tubo y se eliminaron las perlas magnéticas Agencourt AMPure XP beads® (Beckman Coulter).

Ejemplo 6. Validación de las librerías

La calidad de la librería obtenida se determinó analizando una alícuota de la librería en un equipo Agilent 2100 Bioanalyzer Platform® (Agilent), utilizando el kit Agilent High Sensitivity DNA assay® (Agilent) siguiendo las instrucciones y protocolos del fabricante para llevar a cabo este análisis.

Se obtuvo un electroferograma, que muestra un pico de ADN entre 1200-1500 pb, mostrado en la Figura 1.

Ejemplo 7. Combinación de librerías para secuenciación

Se calculó la concentración de cada librería. Se determinó la cantidad de cada una de las librerías que debe usarse para mezclar y hacer el pool utilizando la siguiente fórmula:

Volumen de librería = (Vf x Cf)/(nl x Ci)

donde

Vf es el volumen final del pool

Cf es la concentración final de librería en el pool (4 nM)

nl es el número de librerías

Ci es la concentración inicial de cada librería (nM)

Se combinaron las librerías utilizando el volumen determinado con la fórmula anterior. Se ajustó el volumen final del pool de librerías a la concentración final con Tris-Cl 10 mM, pH 8.5, Tween 20 0.1% v/v.

Ejemplo 8. Secuenciación de las librerías por secuenciación masiva

La concentración de librería o pool de librerías secuenciadas fue de 17 pM. Las librerías se secuenciaron en las plataformas de secuenciación masiva MiSeq® o HiSeq® 2500 de Illumina.

En secuenciaciones en las plataformas de secuenciación masiva MiSeq® de Illumina se añadió una cantidad de los cebadores GeneSGKit® Read Primer 1 (RP1) (Sistemas Genómicos), GeneSGKit® Index Read Primer (iRP) (Sistemas Genómicos) y GeneSGKit® Read Primer 2 (RP2) al cartucho (*cartridge*) de la plataforma de secuencación. El cartucho contiene además cebadores de secuenciación *Illumina Sequencing Primers*® (Illumina, Inc.). La cantidad añadida de los cebadores se indica en la Tabla 12.

Tabla 12. Preparación de cebadores de secuenciación

Reactivo	Pocillo	Cantidad a añadir
GeneSGKit® Read Primer 1	12	3,4 µl
GeneSGKit® Index Read Primer	13	3,4 µl
GeneSGKit® Read Primer 2	14	3,4 µl

En secuenciaciones en las plataformas de secuenciación masiva para HiSeq® 2000 y HiSeq® 2500 de Illumina en modo *High Output,* se combinaron los cebadores RP1, iRP y RP2 con los cebadores de secuenciación *Illumina Sequencing Primers*® de Illumina indicados en la Tabla 13 y en las cantidades indicadas en dicha tabla.

Tabla 13. Preparación de cebadores de secuenciación

Reactivo	Volumen de GeneSGKit® primer	Volumen de Illumina® Primer
GeneSGKit® Read Primer 1	5 µl	995 µl HP6 ó HP10
GeneSGKit® Index Read Primer	15 µl	2985 µl HP8 ó HP12
GeneSGKit® Read Primer 2	15 µl	2985 µl HP7 ó HP11

En secuenciaciones en las plataformas de secuenciación masiva HiSeq® 2500 de Illumina en modo *Rapid Mode* se combinaron los cebadores RP1, iRP y RP2 con los cebadores de secuenciación *Illumina Sequencing Primers*® de Illumina indicados en la Tabla 14 y en las cantidades indicadas en dicha tabla.

Tabla 14: Preparación de cebadores de secuenciación

Reactivo	Abreviatura	Volumen de GeneSGKit® primer	Volumen de Illumina® Primer
GeneSGKit® Read Primer 1	RP1	8,8 µl	1741,2 µl HP10
GeneSGKit® Index Read Primer	iRP	8,8 µl	1741,2 µl HP12
GeneSGKit® Read Primer 2	RP2	8,8 µl	1741,2 µl HP11

Se añadieron los cebadores en el pocillo correspondiente del *cartucho* con ayuda de una punta de pipeta (o pipeta Pasteur). Se mezclaron bien los cebadores evitando la formación de burbujas. Se observó el cartucho para asegurarse de que no hay burbujas. Se eliminaron golpeando suavemente el cartucho sobre la mesa.

Se procedió con la preparación del carril en la plataforma de secuenciación masiva MiSeq® o HiSeq® de Illumina, siguiendo las instrucciones del fabricante.

Ejemplo 9. Identificación de cada secuencia

La Figura 2 muestra un diagrama de flujo de las etapas del análisis bioinformático, que se describen detalladamente más abajo.

Ejemplo 9.1. Determinación de la calidad de los datos en bruto

La secuenciación masiva (NGS, del inglés *Next Generation Sequencing*) generó valores de probabilidad de error por base denominados valores de calidad (*quality values*). Estos valores de calidad se presentaron bajo una escala tipo Phred donde un valor de 20 significa que existe la probabilidad de 1/100 de que la base haya sido producto de error. El estudio global de calidad de datos proporcionó información sobre la calidad de la secuenciación.

Ejemplo 9.2. Alineamiento de secuencias y evaluación del diseño del enriquecimiento de captura

Las lecturas obtenidas se alinearon contra el genoma de referencia versión GRCh38/hg38 mediante BWA (algoritmo mem). Tras el mapeo de las lecturas, se filtraron aquellas secuencias que introducían mayores sesgos y ruido que afectarían a los siguientes pasos de análisis. A partir del archivo BAM formateado obtenido tras el filtrado,

se eliminaron las secuencias de baja calidad y aquellas etiquetadas como duplicados de PCR. Además, se evaluó en este punto la cobertura global de la muestra y la eficiencia de la estrategia diseñada.

Ejemplo 9.3. Llamada de variantes

Se identificaron variantes mediante la llamada de variantes (*variant calling)*. Esta identificación se realizó mediante el análisis de resultados de los alineamientos de las lecturas procedentes de la secuenciación. Los desajustes (*mismatches*) identificados entre la lectura obtenida y el genoma de referencia fueron estudiados con mayor profundidad con el fin de identificar variantes reales de la muestra. Para la ejecución de esta fase se utilizó el algoritmo GATK (*Genome Analysis Tool Kit*) del Broad Institute (https://software.broadinstitute.org/gatk/) (McKenna, A. et al. (2010)).

Ejemplo 9.4. Determinación de bloques haplotípicos y ampliación de los límites de estos bloques

La identificación de bloques haplotípicos implica la determinación de fragmentos de las secuencias de los alelos a determinar que tengan unas variantes concretas y que sean lo más largos posibles, discerniendo así entre las distintas secuencias que tienen cada uno de los alelos en caso de ser heterocigoto o, en el caso de homocigosis, una sola secuencia con sus variantes con respecto a la referencia. La determinación de bloques haplotípicos está basada en el alineamiento de las lecturas contra los genes de referencia y la identificación de variantes sobre esta secuencia, determinadas previamente.

El elevado solapamiento de las lecturas alineadas sobre el genoma de referencia (debido a la gran profundidad de lectura utilizada en el análisis) así como el gran tamaño de longitud de lectura utilizado (2 x 250 nucleótidos) y del tamaño del inserto, permitió discernir, mediante el software WhatsHap (https://whatshap.readthedocs.io/en/latest/) (Patterson, M. et al. (2015)) bloques haplotípicos que en muchos casos cubren todo el gen de estudio. Se obtuvo la secuencia de estos bloques con la herramienta bcftools (https://www.sanger.ac.uk/science/tools/samtools-bcftools-htslib). Una vez obtenidas estas secuencias se procedió al primer proceso de identificación de los alelos.

Además, en muchas ocasiones se dio el caso de que un determinado bloque haplotípico estaba flanqueado por regiones homocigóticas que no quedan registradas en dicho bloque, la adición de estas regiones flanqueantes homocigotas a los bloques facilitó el

proceso de identificación del alelo ya que al aumentar el tamaño de la secuencia se reduce la ambigüedad.

Ejemplo 9.5. Alineamiento local de la secuencia de los bloques haplotípicos obtenidos contra todos los alelos de IPD-IMGT/HLA para cada gen

El primer proceso de identificación de los alelos se realizó mediante alineamiento local de las secuencias obtenidas a partir de los bloques haplotípicos en cada caso contra todas las secuencias genómicas de todos los alelos presentes en la base de datos IPD-IMGT/HLA (https://www.ebi.ac.uk/ipd/imgt/hla/) para cada gen. Esto se realizó con el programa fasta36 (Pearson, W. R. (2016)).

La salida derivada de esta operación proporcionó una lista de posibles candidatos al alelo a identificar. La salida de este proceso priorizada en función del porcentaje de identidad de su secuencia con el alelo en cuestión se muestra en las Tablas 15A y 15B.

Tablas 15A y 15B. Diez primeros candidatos a ser el haplotipo correcto de la muestra VP347 priorizados por el porcentaje de similitud de su secuencia

Alelo	% identidad	Longitud alineamiento	Desajustes (*mismatches*)	Deleción/inserción (*gap opens*)	Comienzo secuencia problema
A*32:04	99.80	3503	5	2	785
A*03:01:01:01	99.80	3503	5	2	785
A*03:284N	99.77	3503	5	3	785
A*03:272	99.77	3503	6	2	785
A*03:271	99.77	3503	6	2	785
A*03:269N	99.77	3503	6	2	785
A*03:268	99.77	3503	6	2	785
A*03:267	99.77	3503	6	2	785
A*03:265	99.77	3503	6	2	785
A*03:264	99.77	3503	6	2	785

Tabla 15B

Alelo	Final secuencia problema	Comienzo secuencia destino	Final secuencia destino	Valor e (*e value*)	Valor *bit score*
A*32:04	4286	1	3502	0	4707.8
A*03:01:01:01	4286	1	3502	0	4707.8
A*03:284N	4286	1	3501	0	4702.1
A*03:272	4286	1	3502	0	4705.3
A*03:271	4286	1	3502	0	4705.3
A*03:269N	4286	1	3502	0	4705.3
A*03:268	4286	1	3502	0	4705.3
A*03:267	4286	1	3502	0	4705.3
A*03:265	4286	1	3502	0	4705.3
A*03:264	4286	1	3502	0	4705.3

El valor e (*e value*) es un parámetro matemático que indica el número de bases dentro de la secuencia respecto a los cuales se puede esperar que sean modificados por azar.

El valor *bit score* es un valor en escala logarítmica. El *score* es un valor matemático que indica la calidad del alineamiento basándose en varios parámetros, entre ellos, longitud de la secuencia, matriz de sustitución y valor de penalización.

En las Tablas 15A y 15B, los primeros dos candidatos tienen idénticos parámetros de salida y son, por lo tanto, indistinguibles por ello. Esto plantea un problema ya que la solución es única. Para solucionar este problema se realizó un segundo criterio de clasificación basado en las distancias filogenéticas de los candidatos a las familias de todos los alelos posibles de cada gen.

Ejemplo 9.6. Pre-cálculo de distancias de la base de datos IPD-IMGT/HLA

Diferentes estudios anteriores han evaluado la diversidad de las secuencias de histocompatibilidad desde el punto de vista evolutivo y jerárquico (Gu, X. et al. (1999); McKenzie, L. M. et al. (1999)). Estos estudios han podido identificar grupos de alelos, pertenecientes a diferentes secuencias de HLA, describiendo una relación filogenética y estableciendo relación con el comportamiento serológico.

Se evaluó la calidad de la agrupación filogenética de las diferentes familias, a diferentes niveles, de HLA contenida en la base de datos IPD-IMGT/HLA (https://www.ebi.ac.uk/ipd/imgt/hla/). Para ello, se tomó el alineamiento múltiple de las familias usando el método mafft (Katoh, K. et al. (2013)). Se calcularon las distancias genéticas y su relación evolutiva usando el método *neihgbor-joining* (Saitou, N. et al. (1987)). El resultado final es una matriz de tamaño NxN, donde se establecen las distancias filogenéticas de todas las secuencias frente a todas y donde N es el número de alelos que tiene la base de datos de alelos HLA de referencia (p. ej. la base de datos IPD-IMGT/HLA ((https://www.ebi.ac.uk/ipd/imgt/hla/)). El número de alelos de la base de datos de alelos HLA puede variar con el tiempo. La matriz de distancia es una matriz que contiene los valores de distancia de cada alelo consigo mismo y con los demás.

Se evaluó la calidad de las familias desde el punto de vista filogenético usando el método silueta (Rousseeuw, P. J. (1987)). Este método es un evaluador de clusters. Esta metodología ha permitido evaluar si las distancias o relaciones filogenéticas entre los miembros de una misma familia son estadísticamente más estrechas respecto al resto de familias de la base de datos. El valor de silueta oscila entre -1 y 1, cuanto más próximo es

al 1 más definido es el grupo. Normalmente, valores por encima de 0.5 se consideran grupos aceptables.

Los resultados obtenidos para la base de datos IPD-IMGT/HLA (https://www.ebi.ac.uk/ipd/imgt/hla/), versión 3.32, se muestran en la Figura 3. Se observa que las diferentes familias de genes de histocompatibilidad, en diferentes niveles, tienen valores de silueta cercanos a 1, lo que indica una relación de cercanía filogenética respecto al resto de secuencias.

Ejemplo 9.7. Selección primaria de alelos candidatos

En un primer paso, para una selección global de posibles HLA candidatos, se realizó un contraste de homología global usando el algoritmo fasta36 (Pearson, W. R. (2016)). El contraste global permitió evaluar la cercanía del haplotipo/s identificado/s usando toda la longitud como unidad, evitando la similitud por dominios o perfiles menores. La selección de un pequeño grupo de posibles HLA se basó en una combinación de parámetros, como son los que definen la homología, el porcentaje de identidad y la longitud del alineamiento.

Ejemplo 9.8. Estudio de filogenia

Una vez identificado el subgrupo de posibles HLA candidatos con mejores parámetros de homología global se realizó un estudio de filogenia evolutiva. Este estudio filogenético se realizó usando la plataforma R y la librería phangorn (Schliep K. P. (2011); Schliep, K. P. et al. (2017)). Se partió de las secuencias de ADN génicas de las HLA pre-seleccionadas y el haplotipo o haplotipos asociados a las muestras problema. Se realizó un cálculo de distancia genética en un contexto “todos contra todos”, usando como modelo para describir la distancia genética la estrategia F81 (Li, H. et al. (2009)). Posteriormente se generó el árbol filogenético usando la estrategia *neighbor-joining* (Felsenstein, J. (1981); Saitou, N. et al. (1987); Studier, J. A. et al. (1988)). Finalmente se accedió a la matriz de distancias utilizando el método cophenetic (Gascuel, O. (1997)).

Ejemplo 9.9. Test de contraste

El test de contraste comenzó con el alelo HLA más cercano filogenéticamente obtenido de la base de datos IPD-IMGT/HLA (https://www.ebi.ac.uk/ipd/imgt/hla/) y obtenido a partir del estudio de filogenia (denominado HLA contrastado) y el haplotipo problema. La

asignación de la familia y los diferentes niveles del haplotipo problema se igualaron al HLA contrastado. Posteriormente se recorrió desde el nivel superior al inferior de la secuencia del HLA contrastado. En cada nivel se evaluó la distancia media, mediana y los intervalos de confianza de cada distribución de distancia de los miembros de la familia con nivel idéntico, obtenida en el Ejemplo 9.4. Si la distancia entre el haplotipo problema y HLA contrastado más cercano se encuentra dentro de la distribución teórica del nivel estudiado con una significancia de 95%, se consideró que pertenecía a ese nivel de manera estadísticamente significativa (Sneath, P.H.A. et al. (1973); Bauer, D. F. (1972)). El test terminó con una identificación estadística del haplotipo problema con los diferentes niveles. Se consideró que el haplotipo problema era el alelo contrastado si eran significativos los 3 niveles superiores. Si esa premisa no se cumplía, el proceso se repitió con la segunda HLA de mayor homología, y así recurrentemente hasta terminar todos los HLA obtenidos en el ejemplo 9.5.

Ejemplo 9.10. Asignación de los niveles y familias del haplotipo problema

Cada haplotipo problema obtuvo en el Ejemplo 9.9 los niveles estadísticamente significativos. Los haplotipos problema que no obtuvieron esa significancia fueron asignados como un alelo de “novo” en el nivel correspondiente.

Ejemplo 10. Análisis de resultados

Se analizaron las librerías generadas de las 4 muestras de referencia VP346, VP347, VP348 y VP349.

Se realizó tagmentación y secuenciación utilizando el kit de Agilent SureSelectQXT Whole Genome Library Prep for Illumina Multiplexed Sequencing. El tamaño de inserto aproximado era 400 pb. El perfil de tamaños obtenidos de las librerías VP346, VP347, VP348 y VP349 se muestra en la Figura 4. Las 4 librerías se secuenciaron en la plataforma de secuenciación Hiseq® 2500 de Illumina en el modo *Rapid Mode* utilizando una longitud de lectura de 100 pb. El gráfico de calidad se muestra en la Figura 5.

Posteriormente, se secuenciaron las mismas librerías en Miseq® con longitud de lectura de 250 pb. El gráfico de calidad se muestra en la Figura 6.

Se realizaron pruebas de tagmentación sobre las muestras VP347 y VP348. También se realizó un ensayo de selección de tamaño de librería utilizando las relaciones 0.7X, 0.6X

y 0.5X de Ampure XP beads® sobre el resultado de PCR en el proceso de obtención de librerías. Los resultados obtenidos se muestran en la Figura 7.

Se secuenciaron las 6 librerías en MiSeq® con una longitud de lectura de 250 pb. El gráfico de calidad se muestra en la Figura 8.

Muestra VP367

Los resultados finales de la comparativa de la muestra VP367 para el alelo HLA*A se muestran en las Tablas 16A y 16B.

Tablas 16A y 16B. Comparativa de la muestra VP367 para el alelo HLA*A

Tabla 16A

Alelo	% identidad	Longitud alineamiento	Desajustes (*mismatches*)	Deleción/inserción (*gap opens*)	Comienzo secuencia problema
A*25:01:01:01	99.49	3517	1	17	785
A*25:42N	99.46	3517	2	17	785
A*25:41	99.46	3517	2	17	785
A*25:40	99.46	3517	2	17	785
A*25:39	99.46	3517	2	17	785
A*25:38	99.46	3517	2	17	785
A*25:29	99.46	3517	2	17	785
A*25:02	99.46	3517	2	17	785
A*25:01:11	99.46	3517	2	17	785
A*25:01:10	99.46	3517	2	17	785
A*25:27:02	99.43	3517	3	17	785
A*25:44	99.37	3517	5	17	785
A*25:43	99.35	3517	6	17	785
A*26:01:01:01	99.26	3517	9	17	785
A*26:17	99.23	3517	10	17	785
A*26:148	99.23	3517	10	17	785
A*26:137	99.23	3517	10	17	785

Tabla 16B

Alelo	Final secuencia problema	Comienzo secuencia destino	Final secuencia destino	Valor e (*e value*)	Valor *bit score*
A*25:01:01:01	4284	1	3517	0	8846.2
A*25:42N	4284	1	3517	0	8841.6
A*25:41	4284	1	3517	0	8841.6
A*25:40	4284	1	3517	0	8841.6
A*25:39	4284	1	3517	0	8841.6
A*25:38	4284	1	3517	0	8841.6
A*25:29	4284	1	3517	0	8841.6
A*25:02	4284	1	3517	0	8841.6
A*25:01:11	4284	1	3517	0	8841.6
A*25:01:10	4284	1	3517	0	8841.6
A*25:27:02	4284	1	3517	0	8837.0
A*25:44	4284	1	3517	0	8827.8
A*25:43	4284	1	3517	0	8823.2
A*26:01:01:01	4284	1	3517	0	8809.4
A*26:17	4284	1	3517	0	8804.8
A*26:148	4284	1	3517	0	8804.8
A*26:137	4284	1	3517	0	8804.8

En las Tablas 16A y 16B se observa que hay un candidato, la secuencia HLA A*25:01:01:01, que tiene mejores estadísticas de similitud, lo que a priori, puede indicar que es la que más se parece. Los métodos del estado de la técnica ofrecen únicamente este resultado, la secuencia de HLA que más se parece a la secuencia problema. En los ejemplos previos de la invención, se determinó que las familias de los genes HLA*A en los diferentes niveles tienen una agrupación o cluster filogenéticos buenos. A continuación, se determinó si la distribución de distancias filogenéticas de la secuencia problema con la familia A*25 y la familia A*25 sola eran estadísticamente idénticas o compatibles. Si la distribución de distancias filogenéticas de la secuencia problema con la familia A*25 y la familia A*25 sola son estadísticamente idénticas o compatibles, el método de la invención determina que la secuencia problema pertenece a la familia A*25 y pasa al siguiente nivel, en este caso A*25:01. Los resultados estadísticos para la familia A*25 se pueden observar en la Figura 9, para la subfamilia A*25:01 en la Figura 10 y para la subfamilia A*25:01:01 en la Figura 11.

El método de la presente invención indica no solo la secuencia más cercana, como los métodos del estado de la técnica, sino que, con el contraste de distribuciones filogenéticas entre la familia y las diferentes subfamilias con la secuencia problema, el método determina, estadísticamente, que pertenecen al grupo A*25:01:01.

Se evaluó la especificidad del método realizando el mismo cálculo para la siguiente familia más cercana, la familia A*26. Los resultados se muestran en la Figura 12.

Muestra VP371

Los resultados finales de la comparativa de la muestra VP371 para el alelo HLA*C se muestran en la Tablas 17A y 17B.

Tablas 17A y 17B. Comparativa de la muestra VP367 para el alelo HLA*C

Tabla 17A

Alelo	% identidad	Longitud alineamiento	Desajustes (*mismatches*)	Deleció/inserción (*gap opens*)
C*12:03:01:01	99.77	4303	10	0
C*12:20	99.72	4303	12	0
C*12:18:01	99.67	4303	14	0
C*06:02:01:02	99.51	4303	21	0
C*06:02:01:01	99.51	4303	21	0
C*06:58	99.49	4303	22	0
C*06:07	99.49	4303	22	0
C*06:06	99.49	4303	22	0
C*06:02:20	99.49	4303	22	0
C*06:02:01:12	99.49	4303	22	0

Tabla 17B

Alelo	Comienzo secuencia problema	Final secuencia problema	Sentido
C*12:03:01:01	4808	506	antisentido
C*12:20	4808	506	antisentido
C*12:18:01	4808	506	antisentido
C*06:02:01:02	4808	506	antisentido
C*06:02:01:01	4808	506	antisentido
C*06:58	4808	506	antisentido
C*06:07	4808	506	antisentido
C*06:06	4808	506	antisentido
C*06:02:20	4808	506	antisentido
C*06:02:01:12	4808	506	antisentido

En las Tablas 17A y 17B se observa que hay un candidato, la secuencia HLA C*12:01:01:01, que tiene mejores estadísticas de similitud, lo que a priori, puede indicar que es la que más se parece. Los métodos del estado de la técnica ofrecen únicamente este resultado, la secuencia de HLA que más se parece a la secuencia problema. En los ejemplos previos de la invención se determinó que las familias de los genes HLA*C en los diferentes niveles tienen una agrupación o cluster filogenéticos buenos. A continuación, se determinó si la distribución de distancias filogenéticas de la secuencia problema con la familia C*12 y la familia C*12 sola eran estadísticamente idénticas o compatibles.

El ejemplo del contraste entre la secuencia problema y la familia C*12 se muestra en la Figura 13.

Este ejemplo indica que, a pesar de que la secuencia problema se parece más a la familia C*12, tal y como indicarían los métodos del estado de la técnica, el método de la invención determina que la divergencia de la secuencia es estadísticamente superior a la divergencia endógena de la familia C*12. Esta divergencia indica que no pertenece a esa familia, por lo que la secuencia problema es una secuencia *de novo*.

Esta conclusión se mantiene en la subfamilia C*12:03, tal y como se aprecia en la Figura 14.

Ejemplo 10.1 Selección de un valor de corte significativo

Se seleccionó un valor de corte p de 1^{-6}, de forma que se identifica una secuencia problema como una secuencia que codifica para al menos un locus de HLA conocido cuando se obtiene un valor de p en el test de contraste mayor o igual a dicho valor de corte de 1^{-6} o, alternativamente, como una secuencia *de novo* que codifica para un nuevo locus de HLA cuando se obtiene un valor de p en el test de contraste menor a dicho valor de corte.

Ejemplo 11. Visualización de los resultados

Se creó expresamente un informe de resultados para la visualización de los resultados del método de identificación genotípica de HLA de la invención. A modo de ejemplo, se muestra el informe de resultados generado para la muestra VP373 se muestra en la Tabla 18.

Tabla 18. Informe de resultados de HLA

INFORME DE RESULTADOS DE HLA
Sexo: Hombre No. de muestra: KT-00000013_08_VP373 Ref: externa: VP373
1. Patología de origen SCID T- B-
2. Indicación clínica de estudio Paciente con inmunofenotipo indicativo de IDCS e infecciones víricas y bacterianas de repetición
3. Resultados

Gen		Alelo genético		Alelo serológico	
HLA_A	Major contig alleles	A*24:02:01:01 100.0%	A*02:01:01:05 99.46%	A*24	A*02
	Minor contig alleles	A*24:68 99.97%	A*02:01:01:01 99.46%	A*24	A*02
	Minor contig alleles	A*24:56 99.97%	A*02:90 99.43%	A*24	A*02
	Minor contig alleles	A*24:53 99.97%	A*02:704 99.43%	A*24	A*02
	Minor contig alleles	A*24:372 99.97%	A*02:661 99.43%	A*24	A*02
	Minor contig alleles	A*24:371 99.97%	A*02:660 99.43%	A*24	A*02
	Minor contig alleles	A*24:370N 99.97%	A*02:659 99.43%	A*24	A*02
	Minor contig alleles	A*24:369 99.97%	A*02:658 99.43%	A*24	A*02
	Minor contig alleles	A*24:364 99.97%	A*02:656 99.43%	A*24	A*02
	Minor contig alleles	A*24:362 99.97%	A*02:650 99.43%	A*24	A*02
HLA_B	Major contig alleles	B*57:01:01:01 99.9%	B*08:01:01:01 100.0%	B*57	B*08
	Minor contig alleles	B*57:03:01:02 99.85%	B*08:01:01:02 99.98%	B*57	B*08
	Minor contig alleles	B*58:01:01:01 98.54%	B*42:01:01 99.58%	B*58	B*42
	Minor contig alleles	B*58:01:01:03 98.53%	B*48:01:01:01 98.7%	B*58	B*48
	Minor contig alleles	B*15:21:01:01 98.19%	B*07:02:01:01 98.7%	B*15	B*07
	Minor contig alleles	B*15:01:01:01 97.97%	B*07:05:01:01 98.68%	B*15	B*07
	Minor contig alleles	B*15:08:01 97.95%	B*07:03 98.75%	B*15	B*07
	Minor contig alleles	B*15:15 97.92%	B*07:05:01:03 98.62%	B*15	B*07
	Minor contig	B*46:01:01	B*39:10:01	B*46	B*39

Gen		Alelo genético		Alelo serológico	
	alleles	97.9%	98.38%		
	Minor contig alleles	B*15:02:01	B*39:01:03:01	B*15	B*39
		98.06%	98.38%		
HLA_C	Major contig alleles	C*07:01:01:01	C*06:02:01:01	C*07	C*06
	Minor contig alleles	C*07:70	C*06:58	C*07	C*06
		99.98%	99.98%		
	Minor contig alleles	C*07:24	C*06:07	C*07	C*06
		99.98%	99.98%		
	Minor contig alleles	C*07:01:31	C*06:02:20	C*07	C*06
		99.98%	99.98%		
	Minor contig alleles	C*07:01:08	C*06:02:01:12	C*07	C*06
		99.98%	99.98%		
	Minor contig alleles	C*07:01:01:16	C*06:02:01:02	C*07	C*06
		99.98%	99.95%		
	Minor contig alleles	C*07:250	C*06:06	C*07	C*06
		99.98%	99.93%		
	Minor contig alleles	C*07:40	C*12:03:01:01	C*07	C*12
		99.95%	99.74%		
	Minor contig alleles	C*07:22	C*12:20	C*07	C*12
		99.95%	99.7%		
	Minor contig alleles	C*07:212	C*12:18:01	C*07	C*12
		99.95%	99.63%		
HLA_DRB1	Major contig alleles	DRB1*15:03:01:01	DRB1*15:03:01:01	DRB1*15	DRB1*15
		94.1%	94.38%		
	Minor contig alleles	DRB1*15:03:01:02	DRB1*15:03:01:02	DRB1*15	DRB1*15
		94.07%	94.35%		
	Minor contig alleles	DRB1*15:01:01:02	DRB1*15:01:01:02	DRB1*15	DRB1*15
		94.05%	94.35%		
	Minor contig alleles	DRB1*15:01:01:05	DRB1*15:01:01:05	DRB1*15	DRB1*15
		94.08%	94.38%		
	Minor contig alleles	DRB1*15:04	DRB1*15:04	DRB1*15	DRB1*15
		94.03%	94.33%		
	Minor contig alleles	DRB1*15:02:01:03	DRB1*15:02:01:03	DRB1*15	DRB1*15
		93.93%	94.21%		
	Minor contig alleles	DRB1*16:02:01:03	DRB1*16:02:01:03	DRB1*16	DRB1*6
		93.98%	94.29%		
	Minor contig alleles	DRB1*15:02:01:02	DRB1*15:02:01:02	DRB1*15	DRB1*15
		93.92%	94.19%		
		DRB1*16:02:01:02	DRB1*16:02:01:02	DRB1*1	DRB1*1

Gen		Alelo genético		Alelo serológico	
	Minor contig alleles	93.97%	94.27%	6	6
	Minor contig alleles	DRB1*16:01:01 93.97%	DRB1*16:01:01 94.27%	DRB1*16	DRB1*16
HLA_DPA1	Minor contig alleles	DPA1*02:07:01:01 99.39%	DPA1*02:07:01:01 99.39%	DPA1*02	DPA1*02
	Major contig alleles	DPA1*02:01:01:01 99.97%	DPA1*02:01:01:01 99.97%	DPA1*02	DPA1*02
	Minor contig alleles	DPA1*02:01:01:03 99.95%	DPA1*02:01:01:03 99.95%	DPA1*02	DPA1*02
	Minor contig alleles	DPA1*02:01:01:02 99.95%	DPA1*02:01:01:02 99.95%	DPA1*02	DPA1*02
	Minor contig alleles	DPA1*02:08 99.94%	DPA1*02:08 99.94%	DPA1*02	DPA1*02
	Minor contig alleles	DPA1*02:02:02:01 99.28%	DPA1*02:02:02:01 99.28%	DPA1*02	DPA1*02
	Minor contig alleles	DPA1*02:02:02:03 99.27%	DPA1*02:02:02:03 99.27%	DPA1*02	DPA1*02
	Minor contig alleles	DPA1*02:02:02:02 99.27%	DPA1*02:02:02:02 99.27%	DPA1*02	DPA1*02
	Minor contig alleles	DPA1*01:03:01:02 96.34%	DPA1*01:03:01:02 96.34%	DPA1*01	DPA1*01
	Minor contig alleles	DPA1*01:03:01:03 96.33%	DPA1*01:03:01:03 96.33%	DPA1*01	DPA1*01
HLA_DPB1	Major contig alleles	DPB1*01:01:01:01 99.98%	DPB1*01:01:01:01 99.98%	DPB1*01	DPB1*01
	Minor contig alleles	DPB1*667:01 99.97%	DPB1*667:01 99.97%	DPB1*667	DPB1*667
	Minor contig alleles	DPB1*127:01 99.97%	DPB1*127:01 99.97%	DPB1*127	DPB1*127
	Minor contig alleles	DPB1*01:01:01:03 99.97%	DPB1*01:01:01:03 99.97%	DPB1*01	DPB1*01
	Minor contig alleles	DPB1*01:01:01:02 99.94%	DPB1*01:01:01:02 99.94%	DPB1*01	DPB1*01
	Minor contig alleles	DPB1*90:01:01 99.93%	DPB1*90:01:01 99.93%	DPB1*90	DPB1*90
	Minor contig alleles	DPB1*01:01:01:04 99.9%	DPB1*01:01:01:04 99.9%	DPB1*01	DPB1*01
		DPB1*26:01:02	DPB1*26:01:02	DPB1*2	DPB1*2

Gen		Alelo genético		Alelo serológico	
	Minor contig alleles	99.88%	99.88%	6	6
	Minor contig alleles	DPB1*01:01:02:02 99.86%	DPB1*01:01:02:02 99.86%	DPB1*01	DPB1*01
	Minor contig alleles	DPB1*01:01:02:01 99.84%	DPB1*01:01:02:01 99.84%	DPB1*01	DPB1*01
HLA_DQA1	Major contig alleles	DQA1*05:01:01:02 98.3%	DQA1*05:01:01:02 95.21%	DQA1*05	DQA1*05
	Minor contig alleles	DQA1*05:03:01:02 98.29%	DQA1*05:03:01:02 95.2%	DQA1*05	DQA1*05
	Minor contig alleles	DQA1*05:07 98.27%	DQA1*05:07 95.18%	DQA1*05	DQA1*05
	Minor contig alleles	DQA1*05:06:01:01 98.27%	DQA1*05:06:01:01 95.18%	DQA1*05	DQA1*05
	Minor contig alleles	DQA1*05:01:01:01 98.27%	DQA1*05:01:01:01 95.18%	DQA1*05	DQA1*05
	Minor contig alleles	DQA1*05:06:01:02 98.26%	DQA1*05:06:01:02 98.17%	DQA1*05	DQA1*05
	Minor contig alleles	DQA1*05:05:01:03 97.06%	DQA1*05:05:01:03 94.2%	DQA1*05	DQA1*05
	Minor contig alleles	DQA1*05:05:01:09 97.0%	DQA1*05:05:01:09 94.14%	DQA1*05	DQA1*05
	Minor contig alleles	DQA1*05:11 96.98%	DQA1*05:11 94.12%	DQA1*05	DQA1*05
	Minor contig alleles	DQA1*05:05:01:01 96.94%	DQA1*05:05:01:01 94.08%	DQA1*05	DQA1*05
4. Transcritos de referencia usados en el análisis de resultados HLA-A (NM_002116.7); HLA-B (NM_005514.6); HLA-C (NM_002117.5); HLA-DRB1 (NM_002124.3); HLA-DQB1 (NM_002123.4); HLA-DPB1 (NM_002121.5); HLA-DQA1 (NM_002122.3); HLA-DPA1 (NM_033554); HLA-DRB3 (NM_022555); HLA-DRB4 (NM_021983); HLA-DRB5 (NM_002125)					
5. Notas Release IMGT-HLA					

LISTA DE REFERENCIAS

Bauer, D. F. (1972). Constructing confidence sets using rank statistics. Journal of the American Statistical Association, 67, 687–690

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. Journal of Molecular Evolution, 17(6):368–376

Gascuel, O. (1997). Concerning the NJ algorithm and its unweighted version, UNJ. in Birkin et. al. Mathematical Hierarchies and Biology, 149-170

Gu, X. et al. (1999). Locus Specificity of Polymorphic Alleles and Evolution by a Birth-and-Death Process in Mammalian MHC Genes. Mol Biol Evol, 16(2):147-56

Hosomichi, K. et al. (2013). Phase-defined complete sequencing of the HLA genes by next-generation sequencing. BMC Genomics, May 28;14:355

Katoh, K. et al. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol Biol Evol, 30(4):772-80

Li, H. et al. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics, 25(14),1754–1760

McKenna, A. et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res., 20(9):1297-303

McKenzie, L. M. et al. (1999). Taxonomic hierarchy of HLA class I allele sequences. Genes Immun. Nov;1(2):120-9.

Patterson, M. et al. (2015). WhatsHap: Weighted Haplotype Assembly for Future-Generation Sequencing Reads. J. Comput. Biol., 22(6):498-509

Pearson, W. R. (2016). Finding Protein and Nucleotide Similarities with FASTA. Curr Protoc Bioinformatics, 53:3.9.1-25

Rousseeuw, P. J. (1987). Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis. Journal of Computational and Applied Mathematics, 20, 53-65

Saitou, N. et al. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution, 4, 406-425

Schliep K. P. (2011). phangorn: phylogenetic analysis in R. Bioinformatics, 27(4), 592-593

Schliep, K. P. et al. (2017). Intertwining phylogenetic trees and networks. Methods in Ecology and Evolution, 8:1212-1220

Shiina, T. et al. (2012). Super high resolution for single molecule-sequence-based typing of classical HLA loci at the 8-digit level using next generation sequencers. Tissue Antigens, 80(4):305-16

Sneath, P.H.A. et al. (1973). Numerical Taxonomy: The Principles and Practice of Numerical Classification, p. 278 ff; Freeman, San Francisco.

Studier, J. A. et al. (1988). A Note on the Neighbor-Joining Algorithm of Saitou and Nei. Molecular Biology and Evolution, 6, 729-731