(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 404 311**

(51) Int. CI.:

C12Q 1/68 (2006.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(96) Fecha de presentación y número de la solicitud europea: **12.04.2006 E 06749954 (1)**

(97) Fecha y número de publicación de la concesión europea: **23.01.2013 EP 1877576**

(54) Título: **Métodos para determinar variantes de secuencias usando secuenciación ultraprofunda**

(30) Prioridad:

12.04.2005 US 104781
06.06.2005 US 688042 P

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
27.05.2013

(73) Titular/es:

454 LIFE SCIENCES CORPORATION (100.0%)
20 COMMERCIAL STREET
BRANFORD CT 06405, US

(72) Inventor/es:

LEAMON, JOHN, HARRIS;
LEE, WILLIAM, LUN;
SIMONS, JAN, FREDRICK;
DESANY, BRIAN;
RONAN, MIKE, TODD;
DRAKE, JAMES;
LOHMAN, KENTON;
EGHOLM, MICHAEL y
ROTHBERG, JONATHAN

(74) Agente/Representante:

DE ELZABURU MÁRQUEZ, Alberto

DESCRIPCIÓN

Métodos para determinar variantes de secuencias usando secuenciación ultraprofunda.

Campo de la invención

La invención proporciona un método para detectar y analizar variantes de secuencias de baja frecuencia, incluyendo polimorfismos de nucleótido único (SNP), variantes de inserción/deleción (referidos como "indeles") y frecuencias alélicas, en una población de polinucleótidos diana en paralelo.

Antecedentes de la invención

El ADN genómico varía significativamente de un individuo a otro, excepto en hermanos idénticos. Muchas enfermedades humanas surgen de las variaciones genómicas. La diversidad genética entre los seres humanos y otras formas de vida explica las variaciones hereditarias observadas en la susceptibilidad a enfermedades. Las enfermedades derivadas de estas variaciones genéticas son la enfermedad de Huntington, la fibrosis quística, la distrofia muscular de Duchenne, y ciertas formas de cáncer de mama. Cada una de estas enfermedades está asociada con una mutación de un único gen. Enfermedades tales como la esclerosis múltiple, la diabetes, el Parkinson, la enfermedad de Alzheimer, y la hipertensión son mucho más complejas. Estas enfermedades pueden estar debidas a causas poligénicas (influencias de múltiples genes) o multifactoriales (influencias de múltiples genes y ambientales). Muchas de las variaciones en el genoma no dan como resultado un rasgo de una enfermedad. Sin embargo, como se ha descrito anteriormente, una sola mutación puede dar lugar a un rasgo de una enfermedad. La capacidad de explorar el genoma humano para identificar la ubicación de los genes que subyacen o están asociados con la patología de estas enfermedades es una herramienta muy poderosa en la medicina y la biología humana.

Varios tipos de variaciones de la secuencia, incluyendo inserciones y deleciones (indeles), diferencias en el número de secuencias repetidas, y diferencias de pares de bases individuales (SNP) dan como resultado la diversidad genómica. Las diferencias de pares de bases individuales, referidas como polimorfismos de nucleótido único (SNP) son el tipo más frecuente de variación en el genoma humano (presentándose en aproximadamente 1 de cada 10^3 bases). Según se utiliza en la presente memoria, un SNP puede ser cualquier posición genómica en la cual aparecen al menos dos o más alelos de nucleótidos alternativos. Según se utiliza la presente memoria, un SNP puede también referirse a cualquier variante inserción/deleción de base única (referida como "indel"), o un indel que incluye la inserción y/o deleción de entre 2 y 100 o más bases. Los SNP son muy adecuados para el estudio de variaciones de la secuencia, ya que son relativamente estables (es decir, exhiben tasas de mutación bajas) y debido a que pueden ser responsables de los rasgos heredados. Se entiende que en la discusión anterior, el término SNP también está destinado a ser aplicable a "indel" (definido a continuación).

Los polimorfismos identificados utilizando análisis basado en microsatélites, por ejemplo, se han utilizado para una variedad de propósitos. El uso de estrategias de ligamiento genético para identificar las ubicaciones de los factores mendelianos individuales ha tenido éxito en muchos casos (Benomar et al. (1995), Nat. Genet., 10:84-8; Blanton et al. (1991), Genomics, 11:857-69). La identificación de las localizaciones cromosómicas de los genes supresores de tumores en general se ha logrado mediante el estudio de la pérdida de heterozigosidad en tumores humanos (Cavenee et al. (1983), Nature, 305:779-784; Collins et al. (1996), Proc. Natl. Acad. Sci. USA, 93:14771-14775; Koufos et al. (1984), Nature, 309:170-172; y Legius et al. (1993), Nat. Genet., 3:122-126). Además, el uso de marcadores genéticos para inferir las posiciones cromosómicas de los genes que contribuyen a rasgos complejos, tales como diabetes de tipo I (Davis et al. (1994), Nature, 371:130-136; Todd et al. (1995), Proc. Natl. Acad. Sci. EE.UU., 92:8560-8565), se ha convertido en un foco de investigación en genética humana.

Si bien se ha avanzado considerablemente en la identificación de las bases genéticas de muchas enfermedades humanas, las metodologías actuales utilizadas para desarrollar esta información están limitadas por los costes prohibitivos y la gran cantidad de trabajo necesario para obtener información genotípica de grandes poblaciones de muestras. Estas limitaciones hacen que la identificación de mutaciones genéticas complejas que contribuyen a trastornos tales como la diabetes sea extremadamente difícil. Las técnicas para explorar el genoma humano para identificar las ubicaciones de los genes implicados en los procesos de enfermedad comenzaron a principios de la década de 1980 con el uso de polimorfismo de longitud de fragmentos de restricción (RFLP), (Botstein et al. (1980), Am. J. Hum. Genet., 32:314-31; Nakamura et al. (1987), Science, 235:1616-22). El análisis RFLP consiste en técnicas de transferencia Southern y otras. La transferencia de Southern es cara y lleva mucho tiempo cuando se realiza en un gran número de muestras, tales como las requeridas para identificar un genotipo complejo asociado con un fenotipo particular. Algunos de estos problemas se evitaron con el desarrollo de análisis marcadores de microsatélites basados en la reacción en cadena de la polimerasa (PCR). Los marcadores microsatélites son polimorfismos de longitud de secuencia simples (SSLP) que consisten en di-, tri-, y tetra-nucleótidos repetidos.

Otros tipos de análisis genómico se basan en el uso de marcadores que hibridan con las regiones hipervariables del ADN que tienen variación multialélica y alta heterozigosidad. Las regiones variables que son útiles para la huella dactilar de ADN genómico son repeticiones en tándem de una secuencia corta referida como minisatélite. El polimorfismo es debido a las diferencias alélicas en el número de repeticiones, que pueden surgir como resultado de intercambios mitóticos o meióticos desiguales o por el deslizamiento de ADN durante la replicación.

Actualmente, la identificación de las variaciones mediante secuenciación del ADN se ve obstaculizada por una serie de deficiencias. En los métodos actuales, la amplificación de una región de interés está seguida por secuenciación directa del producto de amplificación (es decir, una mezcla de secuencias variantes). Alternativamente, la etapa de secuenciación está precedida por una etapa de subclonación microbiana, es decir, por inserción recombinante de productos de amplificación en un vector adecuado para la propagación en el organismo anfitrión deseado.

La desventaja de la secuenciación directa del producto de amplificación consiste en una señal mixta que se produce en sitios variables en la secuencia. Las contribuciones relativas de los diferentes nucleótidos en dichas señales mixtas son difíciles o imposibles de cuantificar, incluso cuando la frecuencia del alelo de menor abundancia se aproxima a 50%. Además, si la variación es una inserción o deleción (en lugar de una sustitución de bases), el desplazamiento de fase resultante entre las diferentes moléculas dará lugar a una señal desordenada, ilegible.

La adición de una etapa de clonación microbiana supera los problemas asociados con la secuenciación directa, ya que no se encuentran señales mixtas. Sin embargo, esta estrategia requiere un mayor número de reacciones de secuenciación. Además, la etapa de clonación microbiana es costosa y consume tiempo, y puede también evitar ciertas variantes, y por tanto distorsionar la frecuencia relativa de las variantes. Si se desea la secuenciación de un gran número (es decir, cientos, miles, decenas de miles) de clones, el coste es extremadamente alto.

Cada uno de estos métodos actuales tiene inconvenientes importantes, ya que consumen mucho tiempo y tienen una resolución limitada. Si bien la secuenciación del ADN proporciona la mayor resolución, también es el método más costoso para la determinación de SNP. En este momento, la determinación de la frecuencia de SNP entre una población de 1.000 diferentes muestras es muy costosa y la determinación de la frecuencia de SNP entre una población de 100.000 muestras es prohibitiva. Por lo tanto, existe una necesidad continua en la técnica de métodos económicos de identificación y resecuenciación de variantes de secuencia presentes en las poblaciones de polinucleótidos, especialmente variantes presentes a frecuencias bajas.

Da Mota et al. (2002), Eur. J. Immunogen, 29:223-227 describen dos enfoques experimentales de genotipificación para identificar alelos del gen BoLA-DRB3 de ganado cebú de raza lechera tropical: (i) secuenciación directa de productos génicos de PCR, y (ii) secuenciación de fragmentos de PCR clonados.

Breve exposición de la invención

De acuerdo con la presente invención, se proporciona un método para la detección de variantes de secuencia que tienen una frecuencia de menos de 5% en una población de ácidos nucleicos, comprendiendo el método las etapas de:

(a) amplificación de un segmento polinucleotídico común a dicha población de ácidos nucleicos con un par de cebadores de ácido nucleico para la PCR que definen un locus para producir una primera población de amplicones que comprenden cada uno dicho segmento polinucleotídico;

(b) liberación de la primera población de amplicones en microrreactores acuosos en una emulsión de agua-en-aceite de manera que una pluralidad de los microrreactores acuosos comprenda (1) un único amplicón de la primera población de amplicones (2) una única perla, y (3) disolución de reacción amplificación que contiene los reactivos necesarios para realizar la amplificación del ácido nucleico;

(c) amplificación clonal de cada miembro de dicha primera población de amplicones mediante reacción en cadena de la polimerasa para producir una pluralidad de poblaciones de segundos amplicones en donde cada población de segundos amplicones deriva de un miembro de dicha primera población de amplicones;

(d) inmovilización de dichos segundos amplicones en una pluralidad de las perlas en los microrreactores de manera que cada cordón comprenda una población de dichos segundos amplicones;

(e) rotura de la emulsión para recuperar las perlas de los microrreactores;

(f) determinación en paralelo de una secuencia de ácido nucleico para los segundos amplicones sobre cada perla a una profundidad (es decir, número de lecturas de secuencia individuales) de más de 100 para producir una población de secuencias de ácido nucleico; y

(g) determinación de una incidencia de cada tipo de nucleótido en cada posición de dicho segmento polinucleotídico para detectar dichas variantes de secuencia en dicha población de ácido nucleico;

en donde el método no requiere un conocimiento previo de la composición de la secuencia de ácido nucleico de las variantes de secuencia,

y en donde dichos cebadores de ácido nucleico son cebadores bipartitos que comprenden una región 5' y una región 3', en donde dicha región 3' es complementaria a una región de dicho segmento polinucleotídico y en donde dicha región 5' es homóloga a un cebador de secuenciación o complemento del mismo.

De acuerdo con la invención, también se proporciona un método de identificación de una distribución de organismos en una población que comprende una pluralidad de diferentes organismos individuales, comprendiendo el método el uso de una muestra de ácido nucleico de dicha población y que comprende las etapas de:

(a) determinación de variantes de secuencia de un segmento de ácido nucleico que comprende un locus común a todos los organismos de dicha población utilizando el método anteriormente mencionado de la invención, en donde cada organismo comprende una secuencia de ácido nucleico diferentes en dicho locus; y

(b) identificación de la distribución de los organismos en dicha población basándose en dicha población de secuencias de ácido nucleico.

También de acuerdo con la invención, se proporciona un método para determinar una composición de una muestra de tejido, comprendiendo el método el uso de una muestra de ácido nucleico de dicha muestra de tejido y que comprende las etapas de:

(a) detección de variantes de secuencia de un segmento de ácido nucleico utilizando el método anteriormente mencionado de la invención, en donde dicho segmento comprende un locus común a todas las células en dicha muestra de tejido y en donde cada tipo de célula comprende una variante de secuencia diferente en dicho locus; y

(b) determinación de la composición de dicha muestra de tejido a partir de dicha frecuencia de nucleótidos. Los términos "región de polinucleótido" y "locus" se utilizan indistintamente en la presente memoria. Los métodos de amplificación y secuenciación de la presente invención facilitan la secuenciación de moléculas individuales. Esta resolución de moléculas individuales permite la detección de alta precisión y/o medición de la frecuencia de las variaciones que están presentes a frecuencias muy bajas en una mezcla de polinucleótido molde.

La presente invención permite la medición exacta de las variaciones de secuencia entre las mezclas de ácido nucleico, especialmente las variaciones que se producen a una frecuencia baja o muy baja. La inclusión de una etapa de amplificación dirigida a una región específica de interés en una muestra de ácido nucleico, acoplada a las llamadas tecnologías de secuenciación de moléculas individuales, permite un descubrimiento preciso, rápido y de bajo coste de variantes de secuencia, y la medición de frecuencias alélicas. Esta mejora sobre los métodos previamente conocidos se consigue, en parte, por el uso de una etapa de amplificación *in vitro* específica de la secuencia que precede a la secuenciación de las moléculas individuales.

Una característica destacada de la presente invención es la capacidad para determinar la secuencia de nucleótidos de una región de polinucleótido de interés a gran profundidad. Por profundidad se entiende el número de lecturas de secuencia individuales que abarca una región de interés dada. Por ejemplo, si se secuencian 1.000 moléculas por separado, la profundidad es igual a 1.000, y también puede ser referida como "1000 veces" o "X 1000". De acuerdo con la invención, la profundidad puede variar de aproximadamente 100 a aproximadamente varios millardos, por ejemplo de aproximadamente 100 a aproximadamente 1.000.000, de aproximadamente 100 a aproximadamente 10.000.000, de aproximadamente 100 a aproximadamente 100.000, o de aproximadamente 1.000 a aproximadamente 1.000.000. La profundidad puede ser mayor de aproximadamente 100, mayor de aproximadamente 1.000, mayor de aproximadamente 10.000, mayor de aproximadamente 100.000, mayor de aproximadamente 1.000.000, mayor de aproximadamente 10 millones, mayor de aproximadamente 100 millones, mayor de aproximadamente 1.000 millones. La profundidad de secuencia lograda mediante los métodos de la presente invención es mucho mayor que las profundidades alcanzables, prácticas o posibles mediante los métodos actuales. Específicamente, los métodos de la presente invención no requieren la clonación microbiana. Mediante la clonación microbiana se quiere significar la amplificación de polinucleótidos en organismos anfitriones microbianos, por ejemplo E. coli. Resultará evidente para el experto en la técnica que las profundidades logradas por la presente invención facilitan la detección de variantes de secuencia raras con relativa facilidad, velocidad y bajo coste.

La invención se refiere a la detección de un diversas variantes de secuencia (por ejemplo, variantes alélicas, variantes de polimorfismo de nucleótidos único, variantes de indel) mediante la identificación de secuencias de polinucleótidos específicas. La tecnología actual permite la detección de SNP, por ejemplo, mediante la reacción en cadena de polimerasa (PCR). Sin embargo, detección de SNP mediante PCR requiere el diseño de cebadores de PCR especiales que hibridan con un tipo de SNP y no con otro tipo de SNP. Además, aunque la PCR es una técnica poderosa, la PCR específica de alelos requiere conocimiento previo de la naturaleza (secuencia) de SNP, así como múltiples rondas de PCR y análisis de electroforesis en gel para determinar una frecuencia alélica. Por ejemplo, una frecuencia alélica de 5% (es decir, 1 de cada 20) requeriría un mínimo de 20 reacciones de PCR para su detección. La cantidad de PCR y electroforesis en gel necesaria para detectar una frecuencia alélica aumenta espectacularmente a medida que se reduce la frecuencia alélica, por ejemplo a 4%, 3%, 2%, 1%, 0,5%, 0,2%, o menos.

Ninguno de los métodos actuales ha proporcionado un método sencillo y rápido de detección de SNP, incluyendo SNP de baja abundancia, mediante la identificación de una secuencia de ADN específica.

Una técnica de PCR acoplada a una técnica de secuenciación de pirofosfato novedosa permite la detección de variantes de secuencia (SNP, indeles y otros polimorfismos de ADN) de una manera rápida, fiable y rentable. Además, el método de la invención puede detectar variantes de la secuencia que están presentes en una muestra de ADN en cantidades alélicas no estequiométricas, tales como, por ejemplo, variantes de ADN presentes en menos de aproximadamente 5% o menos de aproximadamente 1%. Las técnicas se pueden denominar convenientemente "secuenciación ultraprofunda".

El término alelo, según se utiliza en la presente memoria, incluye una variación de secuencia en un sitio variable, en donde la variación se puede producir dentro de un solo organismo, entre organismos individuales de la misma especie, o entre individuos de diferentes especies, entre tejidos normales y enfermos derivados de uno o más individuos, y entre los genomas virales.

Breve descripción de las figuras

Figura 1 describe un esquema de una realización de un proceso de amplificación en emulsión sobre perlas.

Figura 2 describe un esquema de una realización del método de secuenciación ultraprofunda.

Figura 3 describe la evaluación de calidad de los amplicones producidos con pares de cebadores SAD1F/R-DD14 (panel A), SAD1F/R-DE15 (panel B) y SAD1F/R-F5 (panel C). El análisis se realizó en un BioAnalyzer DNA 1000 BioChip representando los picos centrales los productos de PCR y los picos flanqueantes los marcadores de tamaño de referencia. Cada pico se midió para que estuviera dentro de 5 pb del tamaño teórico que oscilaba entre 156 y 181 pares de bases.

Figura 4 describe las frecuencias de nucleótidos (frecuencia de no emparejamientos) en amplicones que representan dos alelos diferentes en el locus MHC II mezclados en proporciones aproximadas (alelo C con respecto al alelo T) de 1:500 (A) y 1:1000 (B), o solamente el alelo T (A), amplificados clonalmente y secuenciados en la plataforma de secuenciación 454 de Life Sciences. Cada barra representa la frecuencia de desviación de la secuencia consenso y está codificada por colores de acuerdo con la sustitución de bases resultante (rojo = A; verde = C; azul = G, amarillo = T).

Figura 5 describe los mismos datos que se presentan en la Figura 4B y 4C, sin embargo después de la sustracción del fondo utilizando la muestra que solamente tiene el alelo T presentada en la Figura 4A.

Figura 6 describe diversas razones de alelos C con respecto a T del locus DD14 HLA mezclados y secuenciados en la plataforma 454 para determinar el rango dinámico. Las proporciones observadas experimentalmente se trazan frente a las proporciones deseadas (abscisa). El número real de lecturas de secuenciación para cada punto de datos se resumen en la Tabla 1.

Figura 7A: Una representación gráfica que muestra la ubicación del mapeo de lecturas con respecto al fragmento de 1,6 Kb del gen 16S que indica aproximadamente 12.000 mapeos de lectura para las 100 primeras bases del gen 16S. B: muestra resultados similares a 7A, excepto con los cebadores V3 que mapean una región de aproximadamente 1000 bases. C: muestra las ubicaciones de las lecturas en las que se utilizan tanto los cebadores V1 como V3.

Figura 8 describe un árbol filogenético que discrimina claramente entre las secuencias V1 (longitud más corta en la mitad izquierda de la figura) y V3 (longitud más larga en la mitad derecha de la figura) en todas menos una de los 200 secuencias.

Figura 9 describe un esquema de una realización del método de secuenciación ultraprofunda. Las flechas horizontales describen cebadores que flanquean la región de interés.

Figura 10 describe un esquema de otra realización del método de secuenciación ultraprofunda. Las flechas horizontales representan cebadores que flanquean la región de interés.

Descripción detallada de la invención

La invención se refiere a un método para detectar una o más variantes de secuencia que tienen una frecuencia de menos de 5% en una población de ácido nucleico. Las variantes de secuencia abarcan cualquier diferencia de secuencia entre dos moléculas de ácido nucleico. Como tales, se entiende que las variantes de secuencia también hacen referencia a al menos, polimorfismos de un solo nucleótido, inserciones/deleciones (indeles), frecuencias alélicas y frecuencias de nucleótidos - es decir, estos términos son intercambiables. Si bien se discuten diversas técnicas de detección a lo largo de esta memoria usando ejemplos específicos, se entiende que el proceso de la invención puede ser igualmente aplicable a la detección de cualquier variante de secuencia. Por ejemplo, una discusión de un procedimiento para la detección de SNP en esta descripción también puede ser aplicable a un procedimiento para la detección de indeles o frecuencias de nucleótidos.

Este procedimiento de la invención puede ser utilizado para amplificar y secuenciar moldes específicos elegidos como diana, tales como los que se encuentran dentro de, entre otros, genomas, muestras de tejidos, poblaciones celulares heterogéneas, poblaciones virales o muestras ambientales. Estos pueden incluir, por ejemplo, productos de PCR, genes candidatos, puntos calientes de mutación, regiones variables importantes desde el punto de vista evolutivo o medico. También se podría utilizar para aplicaciones tales como la amplificación del genoma completo con la posterior secuenciación del genoma completo mediante el uso de cebadores de amplificación variables o degenerados.

Hasta la fecha, el descubrimiento de variantes de secuencias novedosas, en los moldes elegidos como diana ha requerido o bien la preparación y la secuenciación de genomas completos, o bien la amplificación previa mediante PCR de una región de interés, seguido de secuenciación de una reserva de moléculas producto de PCR, o secuenciación de moléculas producto de PCR individuales después de su amplificación mediante subclonación microbiana. Los métodos de la invención permiten llevar a cabo el descubrimiento de variantes de secuencia novedosas, así como el análisis de variantes conocidas, a una profundidad sustancialmente mayor, con una sensibilidad, velocidad muy mejoradas y menor coste que las proporcionadas actualmente por la tecnología existente, evitando al mismo tiempo la subclonación microbiana.

En esta descripción, se puede definir un polimorfismo de un solo nucleótido (SNP) como una variación de secuencia que existe en al menos dos variantes, donde la variante menos común está presente en al menos 0,001% de la población. Se entiende que los métodos de la descripción se pueden aplicar a los "indeles." Por lo tanto, si bien la presente descripción se hace referencia a SNP, se entiende que esta descripción es igualmente aplicable si el término "SNP" se sustituye por el término "indel" en cualquier localización.

Según se utiliza en la presente memoria, se pretende que el término *"indel"* represente la presencia de una inserción o una deleción de uno o más nucleótidos dentro de una secuencia de ácido nucleico en comparación con una secuencia de ácido nucleico relacionada. Una inserción o una deleción incluye por lo tanto la presencia o ausencia de un nucleótido o nucleótidos únicos en una secuencia de ácido nucleico en comparación con una secuencia de ácido nucleico por lo demás idéntica en posiciones de nucleótidos adyacentes. Las inserciones y deleciones pueden incluir, por ejemplo, un único nucleótido, unos pocos nucleótidos o muchos nucleótidos, incluyendo 5, 10, 20, 50, 100 o más nucleótidos en cualquier posición concreta en comparación con una secuencia de referencia relacionada. Se entiende que el término también incluye más de una inserción o deleción dentro de una secuencia de ácido nucleico en comparación con una secuencia relacionada.

La estadística de Poisson indican que el límite inferior de detección (es decir, menos de un evento) para una placa de picotitulación de 60 mm X 60 mm completamente cargada (2 X 10^6 bases de alta calidad, que comprenden 200.000 x 100 lecturas de bases) es de tres eventos con un nivel de confianza de detección de 95% y de cinco eventos con un nivel de confianza de detección de 99% (véase la Tabla 1). Esto aumenta a escalas directamente con el número de lecturas, de manera que se esperan los mismos límites de detección para tres o cinco eventos en 10.000 lecturas, 1000 lecturas o 100 lecturas. Puesto que la cantidad real de lectura de ADN es mayor de 200.000, el límite de detección inferior real se espera en un punto aún más bajo debido al aumento de la sensibilidad del análisis. Como comparación, se ha informado de la detección de SNP a través de la secuenciación basada en pirofosfato para estados alélicos separados en un genoma tetraploide, siempre y cuando el alelo menos frecuente estuviera presente en 10% o más de la población (Rickert et al., 2002 BioTechniques. 32:592-603). Las secuenciación de ADN fluorescente convencional es aún menos sensible, experimentando problemas para resolver 50/50 (es decir, 50%) de alelos heterocigotos (Ahmadian et al., 2000 Anal. Biochem. 280:103-110).

Tabla 1: Probabilidad de detectar cero o uno o más eventos, basándose en el número de eventos en la población total. "*" Indica que la probabilidad de no detectar tres eventos es 5,0%, por lo tanto la probabilidad de detección de dicho evento es de 95%; de manera similar, "**" revela que la probabilidad de detectar uno o más eventos que se producen 5 veces es de 99,3%.

Copias de la secuencia	Porcentaje de probabilidad de detectar cero copias	Porcentaje de probabilidad de detectar una o más copias
1	36,8	63,2
2	13,5	86,5
3	5,0 *	95,0 *
4	1,8	98,2
5	0.7 **	99,3 **

Copias de la secuencia	Porcentaje de probabilidad de detectar cero copias	Porcentaje de probabilidad de detectar una o más copias
6	0,2	99,8
7	0,1	99,9
8	0,0	100,0
9	0,0	100,0
10	0,0	100,0

Como resultado, la utilización de una placa de picotitulación completa de 60 x 60 mm para detectar un único SNP permite la detección de un SNP presente en solo 0,002% de la población con un nivel de confianza del 95% o en 0,003% de la población con un nivel de confianza 99%. Naturalmente, el análisis múltiplex tiene una mayor aplicabilidad que esta profundidad de detección y la Tabla 2 muestra el número de SNP que se pueden escrutar simultáneamente en una única placa de picotitulación, con frecuencias alélicas mínimas detectables a un nivel de confianza de 95% y 99%.

Tabla 2

Clases de SNP	Número de lecturas	Frecuencia de SNP en la población con un nivel de confianza de 95%	Frecuencia de SNP en la población con un nivel de confianza de 99%
1	200000	0,002%	0,003%
2	10000	0,030%	0,050%
5	4000	0,075%	0,125%
10	2000	0,15%	0,25%
50	400	0,75%	1,25%
100	200	1,50%	2,5%
150	133	2,25%	3,75%
200	100	3,0%	5,0%
500	40	7,5%	12,5%
1000	20	15,0%	25,0%

Una ventaja de la invención es que se pueden eliminar o simplificar una serie de pasos, generalmente asociados con la preparación de muestras (por ejemplo, extracción y aislamiento del ADN de tejidos para su secuenciación). Por ejemplo, debido a la sensibilidad del método, ya no es necesario extraer el ADN del tejido utilizando la técnica tradicional de trituración del tejido y purificación química. En su lugar, se puede hervir una pequeña muestra de tejido de menos de un microlitro de volumen y utilizarla para la primera amplificación mediante PCR. El producto de esta amplificación en disolución se añade directamente a la reacción emPCR. Los métodos de la invención por lo tanto reducen el tiempo y el esfuerzo y la pérdida de producto (incluyendo la pérdida debida a error humano).

Otra ventaja de los métodos de la invención es que el método es altamente susceptible de multiplexación. Como se discute a continuación, los cebadores bipartitos de la invención permiten la combinación de conjuntos de cebadores

para múltiples genes con conjuntos de cebadores idénticos para la secuenciación basada en pirofosfato en una única amplificación en disolución. Alternativamente, el producto de múltiples preparaciones se puede colocar en una única reacción de PCR en emulsión. Como resultado, los métodos de la invención presentan un potencial considerable para aplicaciones de alto rendimiento.

Una realización de la invención está dirigida a un método para determinar una frecuencia alélica (incluyendo SNP y frecuencia indel). En la primera etapa, se produce una primera población de amplicones mediante PCR utilizando un primer conjunto de cebadores para amplificar una población diana de ácidos nucleicos que comprenden el locus que se va a analizar. El locus puede comprender una pluralidad de alelos, tales como, por ejemplo, 2, 4, 10, 15 o 20 o más alelos. Los primeros amplicones pueden tener cualquier tamaño, tal como, por ejemplo, entre aproximadamente 50 y aproximadamente 100 pares de bases, entre aproximadamente 100 pb y aproximadamente 200 pb, o entre aproximadamente 200 pb y aproximadamente 1 kb, o entre aproximadamente 500 pb y aproximadamente 5.000 pb, o entre aproximadamente 2000 y aproximadamente 20.000 pares de bases. Una ventaja del método es que no se requiere el conocimiento de la secuencia de ácido nucleico entre los dos cebadores.

En la siguiente etapa, se libera la población de los primeros amplicones en microrreactores acuosos en una emulsión de agua-en-aceite de manera que una pluralidad de microrreactores acuosos comprende (1) ADN suficiente para iniciar una reacción de amplificación dominada por un único molde o amplicón (2) una única perla, y (3) disolución de reacción de amplificación que contiene reactivos necesarios para realizar la amplificación de ácido nucleico (véase la discusión referente a EBCA (Amplificación Clónica Basada en Emulsión) a continuación). Los autores de la presente invención han encontrado que se puede lograr una reacción de amplificación dominada por un único molde o amplicón incluso si dos o más moldes están presentes en el microrreactor. Por lo tanto, también se describen en la presente memoria microrreactores acuosos que comprenden más de un molde. Preferiblemente, cada microrreactor acuoso tiene una única copia del molde de ADN para la amplificación.

Después de la etapa de liberación, se amplifica la primera población de amplicones en los microrreactores para formar los segundos amplicones. La amplificación se puede llevar a cabo, por ejemplo, usando EBCA (que implica PCR) (descrita en el documento WO 2004/069849) en un termociclador para producir los segundos amplicones. Después de la EBCA, los segundos amplicones se pueden unir a las perlas en los microrreactores. Las perlas, con los segundos amplicones unidos se liberan a una formación de cámaras de reacción (por ejemplo, una formación de al menos 10.000 cámaras de reacción) en una superficie plana. La liberación se puede ajustar de manera que una pluralidad de las cámaras de reacción comprenda no más de una única perla. Esto se puede lograr, por ejemplo, mediante el uso de una formación en la que las cámaras de reacción son lo suficientemente pequeñas como para acomodar solo una única perla.

Una reacción de secuenciación se puede realizar de forma simultánea en la pluralidad de cámaras de reacción para determinar una pluralidad de secuencias de ácido nucleico correspondiente a dicha pluralidad de alelos. Los métodos de secuenciación paralela en paralelo utilizando cámaras de reacción se describen en otra sección anterior y en los Ejemplos. Después de la secuenciación, la frecuencia alélica, para al menos dos alelos, se puede determinar mediante el análisis de las secuencias de la población diana de ácidos nucleicos. Como ejemplo, si se determinan 10.000 secuencias y 9.900 secuencias leen "AAA", mientras que 100 secuencias leen "AAG," se puede decir que el alelo "AAA" tiene una frecuencia de aproximadamente 99%, mientras que el alelo "aag" tendría una frecuencia de aproximadamente 1%. Esto se describe con más detalle en la siguiente descripción y en los ejemplos.

Una de las ventajas de los métodos de la invención es que permite un nivel de sensibilidad mayor que el alcanzado anteriormente. Si se utiliza una placa picotitulación, los métodos de la invención pueden secuenciar más de 100.000 o más de 300.000 copias diferentes de un alelo por placa picotitulación. La sensibilidad de detección debe permitir la detección de alelos de baja abundancia que puedan representar aproximadamente 1% o menos de las variantes alélicas. Otra ventaja de los métodos de la invención es que la reacción de secuenciación también proporciona la secuencia de la región analizada. Es decir, no es necesario tener un conocimiento previo de la secuencia del locus que se está analizando.

En una realización preferida, los métodos de la invención pueden detectar una frecuencia alélica que es menor de aproximadamente 5%, o menor de aproximadamente 2%. En una realización más preferida, el método puede detectar frecuencias alélicas de menos de aproximadamente 1%, tal como menos de aproximadamente 0,5%, menos de aproximadamente 0,2%, o menos de aproximadamente 0,02%. Los intervalos típicos de sensibilidad de detección pueden estar entre aproximadamente 0,01% y aproximadamente 100%, entre aproximadamente 0,01% y aproximadamente 50%, entre aproximadamente 0,01% y aproximadamente el 10%, tal como entre aproximadamente 0,1% y aproximadamente 5%.

La población diana de ácidos nucleicos puede ser de diversas fuentes. Por ejemplo, la fuente puede ser un tejido o fluido corporal de un organismo. El organismo puede ser cualquier organismo, incluyendo, pero no limitado a, mamíferos. Los mamíferos puede ser un ser humano o un ganado comercialmente valioso, tal como vacas, ovejas, cerdos, cabras, conejos, y similares. El método de la invención permitiría el análisis de muestras de tejidos y fluidos de plantas. Si bien se pueden analizar todas las plantas mediante los métodos de la invención, las plantas preferidas para los métodos de la invención incluyen especies de cultivos de valor comercial, incluyendo monocotiledóneas y dicotiledóneas. En una realización preferida, la población diana de ácidos nucleicos puede derivar de un cereal o

producto alimenticio para determinar el original y la distribución de los genotipos, alelos, o especies que forman el cereal o producto alimenticio. Tales cultivos incluyen, por ejemplo, maíz, maíz dulce, calabaza, melón, pepino, remolacha azucarera, girasol, arroz, algodón, colza, batata, judía, arveja, tabaco, soja, alfalfa, trigo, o similares.

Las muestras de ácido nucleico se pueden recoger a partir de múltiples organismos. Por ejemplo, la frecuencia alélica de una población de 1.000 individuos se puede llevar a cabo en un experimento que analiza una muestra de ADN mixta de 1.000 individuos. Naturalmente, para que una muestra de ADN mixta sea representativa de la frecuencia alélica de una población, cada miembro de la población (cada individuo) debe contribuir con la misma (o aproximadamente la misma) cantidad de ácido nucleico (mismo número de copias de un alelo) a la muestra reunida. Por ejemplo, en un análisis de la frecuencia alélica genómica, cada individuo puede contribuir al ADN con aproximadamente $1,0x10^6$ células a una muestra de ADN reunida.

En otra realización de la invención, se puede determinar el polimorfismo en un solo individuo. Esto es, el ácido nucleico diana puede ser aislado de un solo individuo. Por ejemplo, se puede examinar ácidos nucleicos reunidos a partir de muestras de múltiples tejidos de un individuo para determinar los polimorfismos y las frecuencias de nucleótidos. Esto puede ser útil, por ejemplo, para determinar el polimorfismo en un tumor, o un tejido que se sospecha que contiene un tumor, de un individuo. El método de la invención se pueden usar, por ejemplo, para determinar la frecuencia de un oncogén activado en una muestra de tejido (o ADN reunido de muestra de múltiples tejidos) de un individuo. En este ejemplo, una frecuencia alélica de 50% o más de oncogenes activados puede indicar que el tumor es monoclonal. La presencia de menos de 50% de un oncogén activado puede indicar que el tumor es policlonal, o que la muestra de tejido contiene una combinación de tejido tumoral y tejido normal (no tumoral). Además, en una biopsia de un tejido sospechoso, la presencia de, por ejemplo, 1% de un oncogén activado puede indicar la presencia de un tumor incipiente, o la presencia de una infiltración del tumor maligno. Además, la presencia de una fracción de células tumorales que tienen una mutación de resistencia a fármacos, en un tumor sensible por otra parte al fármaco, puede pronosticar una recaída del paciente con un tumor totalmente resistente fármaco. Dicha información pronóstica será de gran valor en la terapia y la investigación del cáncer.

La población diana de ácidos nucleicos puede ser cualquier ácido nucleico incluyendo, ADN, ARN y diversas formas de dichos ADN y ARN tales como, pero no limitadas a los plásmidos, cósmidos, genomas virales de ADN, genomas virales de ARN, genomas bacterianos, genomas fúngicos, genomas protozoicos, ADN mitocondrial, genomas de mamíferos, y genomas de plantas. El ácido nucleico se puede aislar de una muestra de tejido o de un cultivo *in vitro*. El ADN genómico se puede aislar de una muestra de tejido, un organismo completo, o una muestra de células. Si se desea, la población diana de ácido nucleico se puede normalizar de manera que contenga una cantidad igual de alelos de cada individuo que contribuye a la población.

Una ventaja de la invención es que el ADN genómico se puede utilizar directamente sin tratamiento adicional. Sin embargo, en una realización preferida, el ADN genómico puede estar sustancialmente libre de proteínas que interfieren con los procedimientos de PCR o hibridación, y también están sustancialmente libres de proteínas que dañan el ADN, tales como nucleasas. Preferiblemente, los genomas aislados también están libres de inhibidores no proteicos de la función de la polimerasa (por ejemplo, metales pesados) e inhibidores no proteicos de la hibridación que interferirían con una PCR. Las proteínas se pueden retirar de los genomas aislados mediante muchos métodos conocidos en la técnica. Por ejemplo, las proteínas se pueden retirar mediante una proteasa, tal como proteinasa K o pronasa, mediante el uso de un detergente fuerte tal como dodecil sulfato de sodio (SDS) o lauril sarcosinato de sodio (SLS) para lisar las células de las cuales se obtienen los genomas de aislados, o ambos. Las células lisadas se pueden extraer con fenol y cloroformo para producir una fase acuosa que contiene ácido nucleico, incluyendo los genomas aislados, que se pueden precipitar con etanol.

La población diana de ácido nucleico puede derivar de fuentes con orígenes desconocidos de ADN, tales como muestras de suelo, muestras de alimentos y similares. Por ejemplo, la secuenciación de un alelo encontrado en un patógeno en una muestra de ácido nucleico a partir de una muestra de alimento permitiría la determinación de la presencia de contaminación por patógenos en el alimento. Además, los métodos de la invención permitirían la determinación de la distribución de un alelo patogénico en el alimento. Por ejemplo, los métodos de la invención pueden determinar la cepa (especie) o la distribución de las cepas (especies) de un organismo concreto (por ejemplo, bacterias, virus, patógenos) en una muestra ambiental tal como una muestra de suelo (Véase el Ejemplo 5) o una muestra de agua de mar.

Una ventaja del método proporcionado en la presente memoria es que no se requiere para el método un conocimiento *a priori* de las mutaciones o variantes de secuencia en una población de ácidos nucleicos o polinucleótidos. Debido a que el método se basa en la secuenciación de ácidos nucleicos, se podrían detectar todas las mutaciones en una localización. Además, no se requiere la clonación microbiana para la secuenciación. Una muestra de ADN se puede amplificar y secuenciar in vitro en una serie de etapas sin necesidad de clonación, subclonación, y cultivo del ADN clonado.

Los métodos de la invención se pueden usar, por ejemplo, para la detección y cuantificación de variantes en muestras virales. Estas muestras virales pueden incluir, por ejemplo, un producto aislado viral de VIH. Otras aplicaciones del método incluyen estudios de población de variantes de secuencia. Las muestras de ADN se pueden recoger de una población de organismos y se pueden combinar y analizaron en un experimento para determinar las

frecuencias alélicas. Las poblaciones de organismos pueden incluir, por ejemplo, una población de seres humanos, una población de ganado, una población de cereal de una cosecha y similares. Otros usos incluyen la detección y cuantificación de mutaciones somáticas en biopsias de tumores (por ejemplo, cáncer de pulmón y colorrectal) o de biopsias que comprenden una población mixta de células tumorales y normales. Los métodos de la invención también se pueden usar para la re-secuenciación de alto nivel de confianza de genes de susceptibilidad clínicamente relevantes (por ejemplo, cáncer de mama, de ovario, colorrectal y de páncreas, melanoma).

Otro uso de la invención implica la identificación de polimorfismos asociados con una pluralidad de genomas distintos. Los genomas distintos se pueden aislar de poblaciones que están relacionadas por alguna característica fenotípica, origen familiar, proximidad física, raza, clase, etc. En otros casos, los genomas se seleccionan al azar de poblaciones tales que no tienen ninguna relación entre sí aparte de ser seleccionados de la misma población. En la presente memoria se describe que el método se puede realizar para determinar el genotipo (por ejemplo, contenido de SNP) de sujetos que tienen una característica fenotípica específica, tal como una enfermedad genética u otro rasgo.

Los métodos de la invención también se pueden utilizar para caracterizar la composición genética de un tumor sometiendo a ensayo la pérdida de heterozigosidad o para determinar la frecuencia alélica de un SNP particular. Adicionalmente, los métodos se pueden utilizar para generar un código de clasificación genómica para un genoma mediante la identificación de la presencia o ausencia de cada uno de un panel de SNP en el genoma y para determinar la frecuencia alélica de los SNP. Cada uno de estos usos se discute en la presente memoria con más detalle.

Un uso de los métodos descritos en la presente memoria comprende un método de genotipificación de alto rendimiento. "Genotipificación" es el procedimiento de identificación de la presencia o ausencia de secuencias genómicas específicas dentro del ADN genómico. Los distintos genomas se pueden aislar de los individuos de poblaciones que están relacionados por alguna característica fenotípica, por origen familiar, por la proximidad física, por raza, por clase, etc. con el fin de identificar polimorfismos (por ejemplo, los asociados con una pluralidad de genomas distintos) que se correlacionan con la familia, localización, raza, clase, etc. del fenotipo. Alternativamente, se pueden aislar al azar distintos genomas de poblaciones de manera que no tengan relación entre sí que no sea su origen en la población. La identificación de polimorfismos en tales genomas indica la presencia o ausencia de los polimorfismos en la población como un todo, pero no se corresponde necesariamente con un fenotipo concreto. Puesto que un genoma puede abarcar una región larga de ADN y puede implicar múltiples cromosomas, un método para detectar un genotipo necesitaría analizar una pluralidad de variantes de secuencia en múltiples localizaciones para detectar un genotipo con una fiabilidad del 99,99%.

Aunque la genotipificación se utiliza a menudo para identificar un polimorfismo asociado con un rasgo fenotípico concreto, esta correlación no es necesaria. La genotipificación solo requiere que un polimorfismo, que puede o no residir en una región codificante, está presente. Cuando la genotipificación se utiliza para identificar una característica fenotípica, se presume que el polimorfismo afecta al rasgo fenotípico que está siendo caracterizado. Un fenotipo puede ser deseable, perjudicial, o, en algunos casos, neutro. Los polimorfismos identificados de acuerdo con los métodos de la invención pueden contribuir a un fenotipo. Algunos polimorfismos se producen dentro de una secuencia codificante de la proteína y de este modo pueden afectar a la estructura de la proteína, lo que causa o contribuye a un fenotipo observado. Otros polimorfismos se producen fuera de la secuencia codificante de la proteína, pero afectan a la expresión del gen. Otros polimorfismos se producen simplemente cerca de los genes de interés y son útiles como marcadores de ese gen. Un único polimorfismo puede causar o contribuir a más de una característica fenotípica y, del mismo modo, una única característica fenotípica puede ser debida a más de un polimorfismo. En general, múltiples polimorfismos que se producen dentro del mismo haplotipo de un gen dado se correlacionan con el mismo fenotipo. Adicionalmente, que un individuo sea heterocigoto u homocigoto para un polimorfismo particular puede afectar a la presencia o ausencia de un rasgo fenotípico particular.

La correlación fenotípica se puede realizar mediante la identificación de una población experimental de sujetos que muestran una característica fenotípica y una población de control que no muestra esa característica fenotípica. Se dice que los polimorfismos que se producen dentro de una población experimental de sujetos que comparten una característica fenotípica y que no se producen en la población control son polimorfismos que se correlacionan con un rasgo fenotípico. Una vez que se ha identificado que un polimorfismo se correlaciona con un rasgo fenotípico, se pueden escrutar los genomas de los sujetos que tienen un potencial para desarrollar un rasgo o característica fenotípicos para determinar la aparición o no aparición del polimorfismo en los genomas de los sujetos con el fin de establecer si es probable que esos sujetos desarrollen con el tiempo la característica fenotípica. Estos tipos de análisis se pueden llevar a cabo en sujetos en riesgo de desarrollar un trastorno concreto, tal como la enfermedad de Huntington o cáncer de mama.

Los métodos descritos en la presente memoria se pueden utilizar para la asociación de un rasgo fenotípico con un SNP. Un rasgo fenotípico abarca cualquier tipo de enfermedad, afección o característica genética, cuya presencia o ausencia pueden ser determinadas positivamente en un sujeto. Los rasgos fenotípicos que son enfermedades o afecciones genéticas incluyen enfermedades multifactoriales de las cuales un componente puede ser genético (por ejemplo, debido a la presencia en el sujeto de un SNP), y la predisposición a tales enfermedades. Estas

enfermedades incluyen, por ejemplo, pero no se limitan a, asma, cáncer, enfermedades autoinmunitarias, inflamación, ceguera, úlceras, enfermedades del corazón o cardiovasculares, trastornos del sistema nervioso, y susceptibilidad a la infección por microorganismos patógenos o virus. Las enfermedades autoinmunitarias incluyen, pero no se limitan a, artritis reumatoide, esclerosis múltiple, diabetes, lupus eritematoso generalizado y enfermedad de Graves. Los cánceres incluyen, pero no se limitan a, cánceres de la vejiga, cerebro, mama, colon, esófago, riñón, sistema hematopoyético, por ejemplo, leucemia, hígado, pulmón, cavidad oral, ovario, páncreas, próstata, piel, estómago y útero. Un rasgo fenotípico puede incluir también la susceptibilidad a fármacos u otros tratamientos terapéuticos, apariencia, estatura, color (por ejemplo, de las plantas con flores), la fuerza, la velocidad (por ejemplo, de los caballos de carreras), el color de pelo, etc. Se han descrito muchos ejemplos de los rasgos fenotípicos asociados con la variación genética, véase, por ejemplo, la Patente de los Estados Unidos Núm. 5.908.978 (que identifica la asociación de la resistencia a enfermedades en algunas especies de plantas asociadas con variaciones genéticas) y la Patente de los Estados Unidos Núm. 5.942.392 (que describe marcadores genéticos asociados con el desarrollo de la enfermedad de Alzheimer).

La identificación de asociaciones entre variaciones genéticas (por ejemplo, aparición de SNP) y rasgos fenotípicos es útil para muchos fines. Por ejemplo, la identificación de una correlación entre la presencia de un alelo de SNP en un sujeto y el desarrollo ulterior por parte del sujeto de una enfermedad es particularmente útil para la administración de tratamientos tempranos, o la institución de cambios de estilo de vida (por ejemplo, la reducción del colesterol o de los alimentos grasos, con el fin de evitar enfermedades cardiovasculares en sujetos que tienen una predisposición mayor que la normal a dicha enfermedad), o el control estricto de un paciente para el desarrollo de cáncer u otra enfermedad. También puede ser útil en el escrutinio prenatal para determinar si un feto padece o está predispuesto a desarrollar una enfermedad grave. Además, este tipo de información es útil para el escrutinio de animales o plantas criados para de mejorar o exhibir las características deseadas.

Un método para determinar un SNP o una pluralidad de SNP asociados con una pluralidad de genomas es el escrutinio para determinar la presencia o ausencia de un SNP en una pluralidad de muestras genómicas derivadas de organismos con el rasgo. Con el fin de determinar qué SNP están relacionados con un rasgo fenotípico particular, las muestras genómicas se aíslan de un grupo de individuos que exhiben el rasgo fenotípico particular, y las muestras se analizan para determinar la presencia de SNP comunes. La muestra genómica obtenida de cada individuo se puede combinar para formar una muestra genómica reunida. A continuación, los métodos de la invención se utilizan para determinar una frecuencia alélica para cada SNP. La muestra genómica reunida se escruta usando paneles de SNP en un método de alto rendimiento para determinar si la presencia o ausencia de un SNP concreto (alelo) se asocian con el fenotipo. En algunos casos, puede ser posible predecir la probabilidad de que un sujeto concreto exhiba el fenotipo relacionado. Si un alelo polimórfico concreto está presente en 30% de los individuos que desarrollan la enfermedad de Alzheimer, pero solo en 1% de la población, en ese caso, un individuo que tenga ese alelo tiene una mayor probabilidad de desarrollar la enfermedad de Alzheimer. La probabilidad también puede depender de varios factores, tales como si los individuos no afectados con la enfermedad de Alzheimer tienen este alelo y si otros factores están asociados con el desarrollo de la enfermedad de Alzheimer. Este tipo de análisis puede ser útil para determinar la probabilidad de que se manifieste un fenotipo concreto. Con el fin de aumentar la capacidad de predicción de este tipo de análisis, se pueden analizar múltiples SNP asociado con un fenotipo particular e identificar los valores de correlación.

También es posible identificar los SNP que segregan con una enfermedad concreta. Se pueden detectar múltiples sitios polimórficos y examinarlos para identificar una conexión física entre ellos o entre un marcador (SNP) y un fenotipo. Esto se puede utilizar para mapear un locus genético ligado o asociado a un rasgo fenotípico para una posición cromosómica y revelar de ese modo uno o más genes asociados con el rasgo fenotípico. Si dos sitios polimórficos segregan al azar, en ese caso o bien están en cromosomas separados o bien están lo bastante distantes entre sí en el mismo cromosoma como para no co-segregar. Si dos sitios co-segregan con una frecuencia significativa, en ese caso, están ligados entre sí en el mismo cromosoma. Estos tipos de análisis de ligamiento pueden ser útiles para el desarrollo de mapas genéticos que pueden definir regiones del genoma importantes para un fenotipo - incluyendo un genotipo de enfermedad.

El análisis de ligamiento se puede realizar en miembros de familias que muestran altas tasas de un fenotipo particular o una enfermedad particular. Las muestras biológicas se pueden aislar de los miembros de la familia que presentan un rasgo fenotípico, así como de sujetos que no presentan el rasgo fenotípico. Cada una de estas muestras se puede utilizar para generar frecuencias alélicas de SNP individuales. Los datos se pueden analizar para determinar si los diferentes SNP están asociados con el rasgo fenotípico y si cualquier SNP segrega o no con el rasgo fenotípico.

Los métodos para analizar los datos de ligamiento han sido descritos en muchas referencias, incluyendo Thompson & Thompson, Genetics in Medicine (5ª edición), WB Saunders Co., Philadelphia, 1991; y Strachan, "Maping the Human Genome" in the Human Genome (Bios Scientific Publishers Ltd., Oxford) capítulo 4, y se resumen en la solicitud de patente publicada PCT Núm. WO 98/18967 por Affymetrix, Inc. El análisis de ligamiento que implica el cálculo del log de los valores de posibilidades (valores LOD) revela la probabilidad de ligamiento entre un marcador y un locus genético en una fracción de recombinación, en comparación con el valor cuando el marcador y el locus genético no están ligados. La fracción de recombinación indica la probabilidad de que los marcadores estén ligados.

Se han desarrollado programas de ordenador y tablas matemáticas para el cálculo de las puntuaciones LOD de diferentes valores de la fracción de recombinación y la determinación de la fracción de recombinación basada en una puntuación LOD concreta, respectivamente. Véanse, por ejemplo, Lathrop, PNAS, USA 81, 3443-3446 (1984); Smith et al., Mathematical Tables for Research Workers in Human Genetics (Churchill, Londres, 1961); Smith, Ann. Hum. Genet. 32, 127-1500 (1968). El uso de valores LOD para el mapeo genético de rasgos fenotípicos se describe en la solicitud de patente publicada PCT Núm. WO 98/18967 de Affymetrix, Inc. En general, un valor de puntuación LOD positivo indica que dos loci genéticos están ligados y una puntuación LOD de +3 o mayor es una fuerte evidencia de que dos loci están ligados. Un valor negativo indica que el ligamiento es menos probable.

Los métodos descritos en la presente memoria son también útiles para evaluar la pérdida de heterozigosidad en un tumor. La pérdida de heterocigosidad en un tumor es útil para determinar el estado del tumor, por ejemplo, si el tumor es un tumor agresivo, metastásico. El método se puede realizar mediante el aislamiento de ADN genómico de la muestra de tumor obtenida a partir de una pluralidad de sujetos que tienen tumores del mismo tipo, así como de tejido normal (es decir, no canceroso) obtenido a partir de los mismos sujetos. Estas muestras de ADN genómico se pueden utilizar en el método de detección de SNP de la invención. La ausencia de un alelo de SNP del tumor en comparación con los alelos de SNP generados a partir de tejido normal indica si se ha producido la pérdida de heterocigosidad. Si un alelo de SNP está asociado con un estado metastásico de un cáncer, la ausencia del alelo de SNP se puede comparar con su presencia o ausencia en una muestra de tumor no metastásico o una muestra de tejido normal. Se ha generado una base de datos de SNP que se producen en tejidos normales y tumorales y se puede comparar la aparición de SNP en una muestra de un paciente con la base de datos para fines de diagnóstico o pronóstico.

Es útil poder diferenciar tumores primarios no metastásicos de tumores metastásicos, puesto que la metástasis es una causa importante de fracaso de tratamiento en pacientes con cáncer. Si la metástasis se puede detectar temprano, ésta se puede tratar agresivamente con el fin de retardar la progresión de la enfermedad. La metástasis es un proceso complejo que implica la separación de células de un tumor primario, el movimiento de las células a través de la circulación, y la eventual colonización de las células tumorales en sitios tisulares locales o distantes. Adicionalmente, es deseable poder detectar una predisposición a desarrollar un cáncer concreto de manera que se pueda iniciar la supervisión y el tratamiento temprano. Muchos cánceres y tumores están asociados con alteraciones genéticas.

Los tumores sólidos progresan desde la tumorigénesis a una fase metastásica y a una fase en la que se pueden producir diversas aberraciones genéticas, por ejemplo, Smith et al., Breast Cancer Res. Terat., 18 Suppl. 1, S5-14, 1991. Se cree que las aberraciones genéticas alteran el tumor de manera que éste puede avanzar a la siguiente fase, es decir, confiriendo ventajas proliferativas, la capacidad de desarrollar resistencia a fármacos o el aumento de la angiogénesis, proteolisis, o capacidad metastásica. Estas aberraciones genéticas son referidas como "pérdida de heterocigosidad". La pérdida de heterocigosidad puede estar causada por una deleción o recombinación que da como resultado una mutación genética que desempeña un papel en el progreso del tumor. Se cree que la pérdida de heterocigosidad para los genes supresores de tumores juega un papel en el progreso del tumor. Por ejemplo, se cree que las mutaciones en el gen supresor del tumor retinoblastoma localizado en el cromosoma 13q14 causa el progreso de retinoblastomas, osteosarcomas, cáncer de pulmón de células pequeñas, y cáncer de mama. Del mismo modo, se ha demostrado que el brazo corto del cromosoma 3 está asociado con el cáncer, tal como el cáncer de pulmón de células pequeñas, el cáncer renal y el cáncer de ovario. Por ejemplo, colitis ulcerosa es una enfermedad que se asocia con un mayor riesgo de cáncer que presumiblemente implica un progreso de múltiples etapas que implica cambios genéticos acumulados (Patente de los estados Unidos Núm. 5.814.444). Se ha demostrado que los pacientes aquejados de colitis ulcerosa de larga duración presentan un mayor riesgo de cáncer, y que un marcador temprano es la pérdida de heterozigosidad de una región del brazo corto distal del cromosoma 8. Esta región es el sitio de un supuesto gen supresor tumoral que también puede estar implicado en el cáncer de próstata y de mama. La pérdida de heterocigosidad se puede detectar fácilmente mediante la realización de los métodos descritos en la presente memoria de forma rutinaria en pacientes aquejados de colitis ulcerosa. Se pueden realizar análisis similares usando muestras obtenidas de otros tumores que se sabe o se cree que están asociados con la pérdida de heterozigosidad. Esto es particularmente ventajoso para el estudio de la pérdida de heterozigosidad debido a que se pueden escrutar de una vez miles de muestras tumorales.

La descripción incluye, en parte, métodos para el procesamiento de ácidos nucleicos para determinar una frecuencia alélica. Uno de estos métodos se puede definir ampliamente en las siguientes tres etapas: (1) Preparación de la muestra - preparación de los primeros amplicones; (2) PCR en emulsión con perlas - preparación de los segundos amplicones; (3) secuenciación por síntesis - determinación de secuencias múltiples a partir de los segundos amplicones para determinar una frecuencia alélica. Cada uno de estas etapas se describe con más detalle a continuación y en la sección de Ejemplos.

1. Preparación del molde de ácido nucleico

Moldes de ácido nucleico

El ácido nucleico molde se puede construir a partir de cualquier fuente de ácido nucleico, por ejemplo, cualquier célula, tejido u organismo, y se puede generar mediante cualquier método reconocido en la técnica.

Alternativamente, las genotecas de moldes se pueden elaborar mediante la generación de una genoteca de ADN complementario (ADNc) a partir de ARN, por ejemplo, ARN mensajero (ARNm). Los métodos de preparación de muestras se pueden encontrar en la solicitud de Patente de los Estados Unidos en tramitación con la presente con el Núm. de Serie 10/767.779 y en la solicitud PCT PCT/US04/02570 y también está publicada en el documento WO/04070007.

Los métodos de la presente invención comprenden la amplificación selectiva de una región de polinucleótido de interés a partir de una población de primeras moléculas polinucleotídicas. La amplificación dará lugar a una población de segundas moléculas de polinucleótido que se derivan de una pluralidad de primeras moléculas que comprenden región de interés. Incluso aunque cada una de las primeras moléculas amplificadas comprenda la región de interés, se apreciará que pueden existir una o más variaciones de secuencia entre las primeras moléculas dentro de la región de interés. El número de primeras moléculas individuales en la población amplificada de este modo puede variar de 2 a varios miles de millones, de forma ventajosa, más de aproximadamente 100, más de aproximadamente 1.000, más de aproximadamente 10.000, más de aproximadamente 100.000, más de aproximadamente 1.000.000, o más aproximadamente mil millones de moléculas.

La amplificación selectiva significa que la amplificación se dirige a una región de interés y por lo tanto amplifica preferiblemente o específicamente esa región de interés. Idealmente, solo la región de interés experimentará amplificación. Sin embargo, el experto en la técnica apreciará que también se puede producir la amplificación no específica sustancial de otras regiones, como se observa con frecuencia en las reacciones de amplificación de ácidos nucleicos. Tales productos de reacción no específicos se pueden evitar mediante la optimización de las condiciones de reacción, por ejemplo mediante modificaciones de la temperatura, el diseño y la concentración del cebador, la concentración de los componentes del tampón y los nucleótidos, y similares. El experto en la técnica estará familiarizado con las estrategias para la optimización de las reacciones de amplificación, incluyendo el uso de cebadores anidados para la mejora de la especificidad de la amplificación. Alternativamente, se puede separar cualquiera de los productos de amplificación no específicos de los productos deseados, por ejemplo, mediante selección por tamaño por medio de electroforesis en gel o técnicas cromatográficas. Dependiendo del grado de la amplificación no específica y del diseño experimental específico, puede no ser necesaria en absoluto la eliminación de productos no específicos.

La reacción de amplificación selectiva se puede llevar a cabo mediante diversos métodos conocidos en la técnica, incluyendo métodos isotérmicos y métodos que requieren termociclación. Por ejemplo, un método de termociclación fácilmente conocido por los expertos en la técnica es la reacción en cadena de la polimerasa (PCR). Un ejemplo de un método isotérmico para la amplificación selectiva es la amplificación isotérmica mediada por bucles (LAMP) descrita por Notomi et al., Nucleic Acids Res. 2000; 28 (12): E63. LAMP emplea la síntesis de ADN con desplazamiento de la hebra autorrecurrente cebada por un conjunto de cebadores específicos para la diana diseñado específicamente. El tamaño de la región de polinucleótidos de interés, es decir, su longitud, oscilará entre aproximadamente 20 y aproximadamente 40.000 nucleótidos, por ejemplo entre aproximadamente 50 y aproximadamente 10.000 nucleótidos, entre aproximadamente 80 y aproximadamente 1.000 nucleótidos, o entre aproximadamente 100 y aproximadamente 500 nucleótidos. Se prefiere una longitud de entre aproximadamente 50 y aproximadamente 2.000 nucleótidos. El producto de amplificación puede estar en la forma de un polinucleótido de hebra sencilla o de doble hebra, o ambos. Estos y otros métodos de amplificación de ADN se describen en: *DNA Amplification*: *Current Techonology and Applications*, V. Demidov y N. Broude, eds, Horizon Bioscience, 2004. También se contemplan las combinaciones de cualquiera de tales métodos de amplificación diferentes.

Con independencia del método utilizado, la amplificación selectiva dará como resultado la síntesis de una o varias poblaciones de segundas moléculas de polinucleótidos. El número de segundas moléculas polinucleótidos individuales en la población amplificada de este modo puede variar de 2 a varios miles de millones, de forma ventajosa, más de aproximadamente 100, más de aproximadamente 1.000, más de aproximadamente 10.000, más de aproximadamente 100.000, más de aproximadamente 1.000.000, o más de aproximadamente mil millones de moléculas. La región de polinucleótidos amplificada puede variar de 2 a varios miles de millones de nucleótidos, comprendiendo ventajosamente al menos aproximadamente 25, al menos aproximadamente 50, al menos aproximadamente 150, al menos aproximadamente 300, al menos aproximadamente 500, al menos aproximadamente 1.000, al menos aproximadamente 5.000, o al menos aproximadamente 10.000 nucleótidos.

La amplificación selectiva también pueden estar dirigida a una pluralidad de regiones de interés, ya sea en reacciones separadas o en una sola reacción (es decir, multiplexación). Si tal pluralidad de regiones se amplificó por separado, los productos de amplificación se pueden combinar (reunir) en cualquier punto antes de la etapa de determinación de la secuencia.

Un método preferido para la preparación del molde de ácido nucleico consiste en realizar la PCR en una muestra para amplificar una región que contiene (se sabe o se sospecha) el alelo o los alelos de interés. La técnica de PCR puede aplicarse a cualquier muestra de ácido nucleico (ADN, ARN, ADNc) utilizando cebadores oligonucleotídicos separados entre sí. Los cebadores son complementarios a las hebras opuestas de una molécula de ADN de doble hebra y están separados típicamente por aproximadamente 50 a 2.000 nucleótidos, o más. Sin embargo, la amplificación por PCR de regiones tan grandes como de 35000 bases es posible mediante el uso de ADN

polimerasas de corrección de pruebas (Barnes, WM (1994) Proc. Natl. Acad. Sci. USA 91:2216). El método de PCR se describe en varias publicaciones, incluyendo Saiki et al., Science (1985) 230:1350-1354; Saiki et al., Nature (1986) 324:163-166; y Scharf et al, Science (1986) 233:1076 - 1078. Véanse también las Patentes de los Estados Unidos Núms. 4.683.194; 4.683.195; y 4.683.202. Los métodos adicionales para la amplificación por PCR se describen en: PCR Technology: Principles and Applications for DNA Amplification ed. HA Erlich, Freeman Press, New York, NY (1992); PCR Protocols: A guide to Methods and Applications, eds. Innis, Gelfland, Snisky y White, Academic Press, San Diego, California (1990); Mattila et al. (1991) Nucleic Acids Res. 19: 4967; Eckert, KA y Kunkel, TA (1991) PCR Methods and Applications 1: 17, y; PCR, eds. McPherson, Quirkes y Taylor, IRL Press, Oxford.

2. Amplificación de moldes ácido nucleico

La población o las poblaciones de segundas moléculas de polinucleótidos se pueden someter a continuación a análisis de la secuencia, por medio del cual se secuencian por separado las segundas moléculas de polinucleótidos individuales.

Opcionalmente sin embargo, antes del análisis de secuencia, las segundas moléculas de polinucleótidos individuales se someten a una segunda ronda de amplificación *in vitro*, dando como resultado la síntesis de una o varias poblaciones de terceras moléculas de polinucleótidos. Esta segunda ronda de amplificación se puede producir mediante cualquiera de los numerosos métodos conocidos en la técnica que permiten que una tercera población de moléculas derivada de cada segunda molécula permanezca separada de las terceras poblaciones molécula que resultan de las otras segundas moléculas. Este tipo de amplificación es referido comúnmente como amplificación clonal. "Clonal", según se utiliza en la presente memoria, significa que comprende una pluralidad de moléculas idénticas, o copias, tal como, por ejemplo, que comprende una pluralidad de moléculas de ácidos nucleicos idénticas amplificadas a partir de una única molécula de ácido nucleico ancestral. Específicamente, cada población es clonal ya que representa una sola segunda molécula de polinucleótido en la siguiente determinación de secuencia.

En una realización, la segunda ronda de amplificación se puede llevar a cabo sobre un soporte sólido o semi-sólido, tal como, por ejemplo, mediante un método de amplificación conocido como amplificación puente, como se describe en la Publicación de la Solicitud de Patente de los Estados Unidos Núm. 2005/0100900, y en 2003/0022207, y en Publicación de la Solicitud de Patente de los Estados Unidos Núm. 2004/0096853. Por consiguiente, las segundas moléculas de polinucleótidos se pueden recocer a moléculas cebadoras oligonucleotídicas apropiadas, que están inmovilizadas sobre un soporte sólido. El cebador se puede prolongar después y la molécula y el cebador se pueden separar el uno del otro. El cebador prolongado se puede recocer después a otro cebador inmovilizado (formando un "puente") y el otro cebador se puede prolongar. Ambos cebadores prolongados se pueden separar después el uno del otro y se pueden utilizar para proporcionar cebadores prolongados adicionales. El procedimiento puede repetirse para proporcionar una o varias poblaciones inmovilizadas, amplificadas de terceras moléculas de polinucleótidos. Si el recocido inicial de las segundas moléculas de polinucleótidos se llevó a cabo de manera que las moléculas recocidas estuvieran a una distancia suficiente entre sí, la población o las poblaciones de terceros polinucleótidos tenderán a permanecer separadas entre sí en forma de colonias y por lo tanto serán clonales. De este modo, aunque las colonias puedan estar en estrecha proximidad entre sí en un único soporte sólido o semi-sólido, en condiciones de partida apropiadas la mayoría de las colonias no obstante será distinta y representará productos de amplificación clonales. Estas colonias que comprenden productos de amplificación puente se pueden someter a análisis de la secuencia de nucleótidos.

En otra realización, la segunda ronda de amplificación se puede llevar a cabo mediante amplificación en una emulsión (documentos WO 2004/069849 y WO 2005/073410). La emulsión puede contener millones de reacciones individuales. La emulsión puede contener micropartículas con las que se asocian los productos de amplificación de forma clonal.

En otra realización, la segunda ronda de amplificación se puede realizar sobre un soporte semi-sólido, por ejemplo mediante la tecnología polony descrita en las Patentes de los Estados Unidos Núms. 6.432.360; 6.485.944; y 6.511.803. Por ejemplo, los cebadores oligonucleotídicos se inmovilizan sobre un soporte semi-sólido, los ácidos nucleicos molde se siembran sobre el soporte semi-sólido y se hibridan con los cebadores, que son prolongados usando una ADN polimerasa y desoxinucleótidos trifosfato, a continuación se desnaturalizan. Varias rondas de recocido, prolongación y desnaturalización conducen a la amplificación clonal *in situ* sobre el soporte semi-sólido. Los productos de amplificación se restringen espacialmente a las proximidades inmediatas de la molécula molde de la cual derivan. Esto da como resultado la creación de colonias de PCR, conocidas en la técnica como polonias. La secuencia de polinucleótidos de las moléculas de ácido nucleico de cada polonia se pueden determinar después mediante diversos métodos conocidos en la técnica, incluyendo métodos de secuenciación por síntesis, por ejemplo como describen por Mitra et al. (2003) Analyt. Biochem. 320:55-65.

En una realización preferida, la segunda ronda de amplificación se puede llevar a cabo mediante un sistema de amplificación novedoso, denominado en la presente memoria EBCA (Amplificación Clonal Basada en Emulsión o amplificación en emulsión con perlas) utilizado para realizar esta segunda amplificación. La EBCA (documento WO 2004/069849 y documento WO 2005/073410) se lleva a cabo anclando un ácido nucleico molde (p. ej., ADN) que se va a amplificar a un soporte sólido, preferiblemente en forma de una perla generalmente esférica. Una genoteca de ADN molde de hebra sencilla preparada de acuerdo con los métodos de preparación de muestras de esta invención

es un ejemplo de una fuente adecuada de la genoteca de moldes de ácido nucleico de partida que se va a anclar a una perla para su uso en este método de amplificación.

La perla esta unida a un gran número de una especie de cebador individual (es decir, cebador B en la Figura 1) que es complementaria a una región del ADN molde. El ADN molde se recuece al cebador unido a la perla. Las perlas se suspenden en una mezcla de reacción acuosa y después se encapsulan en una emulsión de agua-en-aceite. La emulsión se compone de microgotas en fase acuosa discretas, aproximadamente de 60 a 200 µm de diámetro, delimitadas por una fase oleosa termoestable. Cada microgota contiene, preferiblemente, disolución de reacción de amplificación (es decir, los reactivos necesarios para la amplificación del ácido nucleico). Un ejemplo de una amplificación sería una mezcla de reacción de PCR (polimerasa, sales, dNTP) y un par de cebadores de PCR (cebador A y cebador B). Véase, la Figura 1A. Un subconjunto de la población de microgotas también contiene las perla de ADN que comprenden el molde de ADN. Este subconjunto de microgotas es la base para la amplificación. Las microcápsulas que no están dentro de este subconjunto no tienen ADN molde y no participarán en la amplificación. La técnica de amplificación es la PCR y los cebadores de PCR pueden estar presentes en una relación de 8:1 o 16:1 (es decir, 8 o 16 de un cebador a 1 del segundo cebador) para realizar la PCR asimétrica.

En esta visión general, el ADN se recuece a un oligonucleótido (cebador B) que está inmovilizado en una perla. Durante la termociclación (Figura 1B), se rompe el enlace entre el molde de ADN de hebra sencilla y el cebador B inmovilizado sobre la perla, liberando el molde en la disolución microencapsulada circundante. La disolución de amplificación, en este caso la disolución de PCR, contiene adicionalmente cebador A y cebador B en fase de disolución. Los cebadores B en fase de disolución se unen fácilmente a la región b' complementaria del molde ya que las cinéticas de unión son más rápidas para los cebadores en fase de disolución que para los cebadores inmovilizados. En la PCR de fase temprana, ambas hebras A y B se amplifican igualmente bien (Figura 1C).

En la PCR de fase intermedia (es decir, entre los ciclos 10 y 30) los cebadores B se agotan, deteniendo la amplificación exponencial. La reacción se entra a continuación en una amplificación asimétrica y la población de amplicones para a estar dominada por hebras A (Figura 1D). En la PCR de fase tardía (Figura 1E), después de 30 a 40 ciclos, la amplificación asimétrica aumenta la concentración de hebras A en disolución. Las hebras A en exceso comienzan a recocerse a los cebadores B inmovilizados en las perlas. Las polimerasas termoestables utilizan a continuación la hebra A como molde para sintetizar una hebra B del amplicón unida a la perla, inmovilizada.

En la PCR de fase final (Figura 1F), la ciclación térmica continuada fuerza el recocido adicional a los cebadores unidos a las perlas. La amplificación en fase de disolución puede ser mínima en esta etapa, pero la concentración de hebras B inmovilizadas aumenta. A continuación, la emulsión se rompe y el producto inmovilizado se vuelve de hebra sencilla mediante desnaturalización (por calor, pH, etc.), que elimina la hebra A complementaria. Los cebadores A se recuecen a la región A' de la hebra inmovilizada, y la hebra inmovilizado se carga con las enzimas de secuenciación, y cualquier proteínas accesoria necesaria. Las perlas se secuencian después usando técnicas de pirofosfato reconocidas (descritas, p. ej., en las Patentes de los Estados Unidos Núms. 6.274.320, 6.258.568 y 6.210.891.

Los cebadores utilizados para la amplificación son bipartitos - que comprenden una sección 5' y una sección 3'. La sección 3' del cebador contiene la secuencia específica de la diana (véase la Figura 2) y realiza la función de cebador de la PCR. La sección 5' del cebador comprende secuencias que son útiles para el método de secuenciación o el método de inmovilización. Por ejemplo, en la Figura 2, la sección 5' de los dos cebadores utilizados para la amplificación contiene secuencias (marcadas como 454 directa y 454 inversa) que son complementarias a los cebadores de una perla o un cebador de secuenciación. Es decir, la sección 5', que contiene la secuencia directa o inversa, permite que los amplicones se unan a las perlas que contienen los oligos inmovilizados que son complementarios a la secuencia directa o inversa. Además, la reacción de secuenciación puede iniciarse usando cebadores de secuenciación que sean complementarios a las secuencias cebadoras directa e inversa. De este modo, se puede utilizar en todas las reacciones un conjunto de perlas que comprenden secuencias complementarias a la sección 5' del cebador bipartito. De una manera similar, se puede utilizar un conjunto de cebadores de secuenciación que comprende secuencias complementarias a las sección 5' del cebador bipartito para secuenciar cualquiera de los amplicones elaborados usando el cebador bipartito. En la realización más preferida, todos los conjuntos de cebadores bipartitos utilizados para la amplificación tendrían el mismo conjunto de secciones 5' tales como el cebador 454 directo y el cebador 454 inverso mostrados en la Figura 2. En este caso, todos los amplicones se pueden analizar utilizando perlas convencionales recubiertas con oligos complementarios a la sección 5'. Los mismos oligos (inmovilizados sobre perlas o no inmovilizados) se pueden utilizar como oligos de secuenciación.

Rotura de la emulsión y recuperación de perlas

Después de la amplificación de la molde, la emulsión se "rompe" (también referido como "desemulsificación" en la técnica). Existen muchos métodos para romper una emulsión (véase, p. ej., la Patente de los Estados Unidos Núm. 5.989.892 y referencias allí citadas) y un experto en la técnica sería capaz de seleccionar el método apropiado. Un método preferido para romper la emulsión se describe en detalle en la sección Ejemplos.

Una vez que se ha roto, las perlas que contienen el molde amplificado se pueden resuspender a continuación en disolución acuosa para su uso, por ejemplo, en una reacción de secuenciación de acuerdo con las tecnologías conocidas. (Véanse, Sanger, F. et al., Proc. Natl. Acad. Sci. USA 75, 5463-5467 (1977); Maxam, A. M. y Gilbert, W. Proc Natl Acad Sci USA 74, 560-564 (1977); Ronaghi, M. et al., Science 281, 363, 365 (1998); Lysov, I. et al., Dokl Akad Nauk SSSR 303, 1508-1511 (1988); Bains W. y Smith G. C. J. TheorBiol 135, 303-307 (1988); Drnanac, R. et al., Genomics 4, 114-128 (1989); Khrapko, K. R. et al., FEBS Lett 256. 118-122 (1989); Pevzner P. A. J Biomol Struct Dyn 7, 63-73 (1989); Southern, E. M. et al., Genomics 13, 1008-1017 (1992)). Si las perlas se van a utilizar en una reacción de secuenciación basada en pirofosfato (descrita, por ejemplo, en las Patentes de los Estados Unidos Núms. 6.274.320, 6.258.568 y 6.210.891, es necesario eliminar la segunda hebra del producto de la PCR y recocer un cebador de secuenciación al molde de hebra sencilla que está unido a la perla.

En este punto, el ADN amplificado sobre la perla se puede secuenciar o bien directamente sobre la perla o bien en un recipiente de reacción diferente. En una realización de la presente invención, el ADN se secuencia directamente sobre la perla transfiriendo la perla a un recipiente de reacción y sometiendo el ADN a una reacción de secuenciación (por ejemplo, secuenciación con pirofosfato o de Sanger). Alternativamente, las perlas se pueden aislar y el ADN se puede separar de cada perla y secuenciar. En cualquier caso, las etapas de secuenciación se pueden llevar a cabo sobre cada perla individual.

3. Métodos de secuenciación de ácidos nucleicos

De acuerdo con la invención, cada una de una pluralidad o población de segundas moléculas de polinucleótidos, u opcionalmente cada uno de una pluralidad o población de terceras moléculas de polinucleótidos, se somete a análisis de secuencia de nucleótidos. La secuencia de las segundas (opcionalmente y tercera) moléculas de polinucleótidos se determina mediante los métodos de la invención, que oscilan de 2 a varios miles de millones, ventajosamente de más de aproximadamente 100, más de aproximadamente 1000, más de aproximadamente 10.000, más de aproximadamente 100.000, más de aproximadamente 1 millón, o más de aproximadamente mil millones. La secuencia puede comprender al menos dos nucleótidos consecutivos, preferiblemente al menos aproximadamente 5, por lo menos aproximadamente 25, al menos aproximadamente 50, al menos aproximadamente 100, al menos aproximadamente 150, al menos aproximadamente 200, al menos aproximadamente 300, al menos aproximadamente 500, al menos aproximadamente 1000, al menos aproximadamente 5000, al menos aproximadamente 10.000, o al menos aproximadamente 100.000 nucleótidos consecutivos y se determinan a partir de cada una de las segundas (u opcionalmente terceras) moléculas de polinucleótidos.

El experto en la técnica estará familiarizados con los diferentes métodos para la secuenciación de polinucleótidos. Estos incluyen, pero no se limitan a, la secuenciación de Sanger (también referida como secuenciación didesoxi) y varios métodos de secuenciación por síntesis (SBS) como revisó Metzger (Metzger ML 2005, *Genome Research* 1767), secuenciación por hibridación, por ligación (por ejemplo, documento WO 2005/021786), por degradación (por ejemplo, Patentes de los Estados Unidos Núms. 5.622.824 y 6.140.053) y secuenciación por nanoporos.

Cualquiera de los métodos de amplificación y secuenciación de polinucleótidos conocidos en la técnica se puede utilizar de acuerdo con la presente invención, siempre y cuando el enfoque seleccionado de como resultado la determinación de la secuencia de moléculas de polinucleótidos individuales, u opcionalmente la determinación de la secuencia de poblaciones de polinucleótidos clonales obtenidos mediante amplificación de dichas moléculas de polinucleótidos individuales. Cualquier amplificación se produce *in vitro,* en oposición a la clonación microbiana.

En ciertas realizaciones, la secuenciación de polinucleótidos se consigue mediante cualquiera de un grupo de métodos referido como síntesis por secuenciación (SBS). La SBS hace referencia a métodos para determinar la identidad de uno o más nucleótidos en un polinucleótido o en una población de polinucleótidos, en donde los métodos comprenden la síntesis por etapas de una hebra sencilla de polinucleótido complementaria al polinucleótido molde cuya secuencia de nucleótidos se va a determinar. Un cebador oligonucleotídico se diseña para que se recueza a una posición complementaria, predeterminada de la molécula molde de muestra. El complejo de cebador/molde se presenta con un nucleótido en presencia de una enzima polimerasa de ácido nucleico. Si el nucleótido es complementario a la posición en la molécula molde de la muestra que está directamente adyacente al extremo 3' del cebador oligonucleotídico, en ese caso la polimerasa prolongará el cebador con el nucleótido. Alternativamente, el complejo de cebador/molde se presenta con todos los nucleótidos de interés (típicamente A, G, C y T) de una vez, y se incorpora el nucleótidos que es complementario a la posición en la molécula molde de la muestra directamente adyacente al extremo 3' del cebador oligonucleotídico. En cualquiera de los casos, los nucleótidos pueden ser bloqueados químicamente (por ejemplo en la posición 3'-O) para evitar la prolongación adicional, y no es necesario desbloquearlos antes de la siguiente ronda de síntesis. Cualquier incorporación del nucleótido se puede detectar mediante una variedad de métodos conocidos en la técnica, p. ej., detectando la liberación de pirofosfato (PPi) (Patentes de los Estados Unidos Núms. 6.210.891; 6.258.568; y 6.828.100) a través de quimioluminiscencia, o mediante el uso de marcas detectables unidas a los nucleótidos. Las marcas detectables incluyen etiquetas de masas (por ejemplo, Patentes de los Estados Unidos Núms. 5.622.824 y 6.140.053) y marcas fluorescentes o quimioluminiscentes. La marca detectable se une directamente o indirectamente a los nucleótidos. En el caso de las marcas fluorescentes, la marca puede ser excitada directamente por un estímulo luminoso externo,

o indirectamente por emisión de un donador fluorescente (FRET) o luminiscente (LRET) (Patentes de los Estados Unidos Núms. 6.982.146). Después de la detección de la marca detectable, la marca tiene que ser inactivada, o separada de la reacción, de modo que no interfiera o se mezcle con la señal de una marca posterior. La separación de la marca se puede lograr, por ejemplo, mediante escisión química (por ejemplo, Publicación de la Solicitud de Patente de los Estados Unidos Núm. 2003/0124594) o fotoescisión. La inactivación de la marca se puede lograr, por ejemplo, mediante fotoblanqueo. De acuerdo con la invención, se puede utilizar cualquier método de SBS conocido en la técnica en la secuenciación de los segundos polinucleótidos, o de las población o poblaciones de terceros polinucleótidos.

De acuerdo con la invención, la secuenciación de polinucleótidos también se puede lograr mediante un método basado en nanoporos. El principio subyacente a la secuenciación por nanoporos es que una molécula de ADN o de ARN de hebra sencilla puede ser impulsada mediante electroforesis a través de un poro de escala nanométrica de tal manera que la molécula atraviesa el poro de una forma lineal estricta. Debido a que una molécula que se transloca obstruir o bloquea parcialmente el nanoporo, ésta altera las propiedades eléctricas del poro. Este cambio en las propiedades eléctricas depende de la secuencia de nucleótidos, y se puede medir. El nanoporo puede comprender una molécula de proteína, o puede estar en estado sólido. Una ventaja de los métodos basados en nanoporos es que se puede lograr longitudes de lectura muy largas, p. ej., se puede leer miles, decenas de miles o cientos de miles de nucleótidos consecutivos a partir de una sola molécula. Los métodos de caracterización de polinucleótido a través de nanoporos se discuten por ejemplo, en las Publicaciones de Solicitudes de Patente de los Estados Unidos Núms. 2006/0063171, U.S. 2006/0068401, y U.S. 2005/0202444.

Un método de secuenciación es un método de SBS referido como secuenciación basada en pirofosfato. En la secuenciación basada en pirofosfato, la secuencia de ADN de la muestra y el cebador de prolongación se someten a una reacción de la polimerasa en presencia de un nucleótido trifosfato mediante la cual el nucleótido trifosfato solo se incorporará y liberará pirofosfato (PPi) si es complementario a la base en la posición diana, añadiéndose el nucleótido trifosfato o bien alícuotas separadas de la mezcla de cebadores de la muestra o bien sucesivamente a la misma mezcla de cebadores de la muestra. A continuación se detecta la liberación de PPi para indicar qué nucleótido se incorpora.

En una realización, una región del producto de la secuencia se determina mediante el recocido de un cebador de secuenciación a una región del ácido nucleico molde, y después poniendo en contacto el cebador de secuenciación con una ADN polimerasa y un nucleótido de trifosfato conocido, es decir, dATP, dCTP, dGTP, dTTP, o un análogo de uno de estos nucleótidos. La secuencia se puede determinar mediante la detección de un subproducto de reacción de secuencia, como se describe a continuación.

El cebador de secuencia puede tener cualquier longitud o composición de bases, siempre que sea capaz de recocerse específicamente a una región del molde de ácido nucleico amplificado. No se requiere ninguna estructura particular para el cebador de secuenciación siempre que sea capaz de cebar específicamente una región en el ácido nucleico molde amplificado. Preferiblemente, el cebador de secuenciación es complementario a una región del molde que se encuentra entre la secuencia que se va a caracterizar y la secuencia hibridable al cebador de anclaje. El cebador de secuenciación se prolonga con la ADN polimerasa para formar un producto de secuencia. La prolongación se lleva a cabo en presencia de uno o más tipos de nucleótidos trifosfato, y si se desea, proteínas de unión coadyuvantes.

La incorporación del dNTP se determina preferiblemente analizando la presencia de un subproducto de secuenciación. En una realización preferida, la secuencia de nucleótidos del producto de la secuenciación se determina mediante la medición del pirofosfato inorgánico (PPi) liberado de un nucleótido trifosfato (dNTP) a medida que se incorpora el dNMP a un cebador de secuencia prolongada. Este método de secuenciación, denominado tecnología Pyrosequencing™ (Pyrosequencing AB, Estocolmo, Suecia) se puede llevar a cabo en disolución (fase líquida) o en forma de una técnica en fase sólida. Los métodos de secuenciación basados en PPi se describen generalmente, por ejemplo, en el documento WO 9813523A1, Ronaghi, et al., 1996. *Anal. Biochem*. 242: 84-89, Ronaghi, et al., 1998. *Science* 281: 363-365 (1998) y en la Publicación de la Solicitud de Patente de los Estados Unidos Núm. 2001/0024790. Véanse también, p. ej., las Patentes de los Estados Unidos Núms. 6.210.891 y 6.258.568.

En una realización preferida, la secuenciación del ADN se lleva a cabo utilizando el aparato de secuenciación de la empresa 454 (454 Life Sciences) y los métodos que se describen en las solicitudes de patente en tramitación con la presente USSN: 10/768.729, USSN: 10/767.779, USSN: 10/767.899, y USSN: 10/767.894- Todas las cuales se presentaron el 28 de Enero de 2004.

A menos que se defina lo contrario, todos los términos técnicos y científicos utilizados en la presente memoria tienen el mismo significado comúnmente comprendido por un experto normal en la técnica a la cual pertenece esta invención. La definición comúnmente comprendida incluiría aquellas definidas en los documentos USSN: 60/476.602, presentada el 6 de Junio de 2003; USSN: 60/476.504, presentada el 6 de Junio de 2003; USSN: 60/443.471, presentada el 29 de Enero de 2003; USSN: 60/476.313, presentada el 6 de Junio de 2003; USSN: 60/476.592, presentada el 6 de Junio de 2003; USSN: 60/465.071, presentada el 23 de Abril de 2003; USSN:

60/497.985, presentada el 25 de Agosto de 2003; USSN: 10/767.779 presentada el 28 de Enero de 2004; 10/767.899 presentada el 28 de Enero de 2004; USSN: 10/767.894 presentada el 28 de Enero de 2004.

Ejemplos

Ejemplo 1Secuenciación del locus HLA

Se diseñaron cinco pares de cebadores de PCR para abarcar los SNP descritos públicamente, conocidos en el locus de la clase II del MHC. Los cebadores se diseñaron utilizando el programa Primer3 (Whitehead Institute for Biomedical Research) utilizando secuencias genómicas de aprox. 200 pares de bases de longitud que abarcan las regiones diana como de entrada. Cada cebador consistió en una porción 3' específica del locus que tenía una longitud que oscilaba de 20 a 24 bases y una porción 5' de 19 bases constante (que se muestra en minúsculas) que incluye una clave de 4 bases (resaltada en negrita). Los cebadores se obtuvieron de Integrated DNA Technologies (Coralville, IA):

SAD1F-DC1 gcctccctcgcgcca **TCAG**ACCTCCCTCTGTGTCCTTACAA (SEQ ID NO: 1)

SAD1R-DC1 gccttgccagcccgc **TCAG**GGAGGGAATCATACTAGCACCA (SEQ ID NO: 2)

SAD1F-DD14 gcctccctcgcgcca **TCAG**TCTGACGATCTCTGTCTTCTAACC (SEQ ID NO: 3)

SAD1R-DD14 gccttgccagcccgc **TCAG**GCCTTGAACTACACGTGGCT (SEQ ID NO: 4)

SAD1F-DE15 gcctccctcgcgcca **TCAG**ATTTCTCTACCACCCCTGGC (SEQ ID NO: 5)

SAD1R-DE15 gccttgccagcccgc **TCAG**AGCTCATGTCTCCCGAAGAA (SEQ ID NO: 6)

SAD1F-GA9 gcctccctcgcgcca **TCAG**AAAGCCAGAAGAGGAAAGGC (SEQ ID NO: 7)

SAD1R-GA9 gccttgccagcccgc **TCAG**CTTGCAGATTGGTCATAAGG (SEQ ID NO: 8)

SAD1F-F5 gcctccctcgcgcca **TCAG**ACAGTGCAAACACCACCAAA (SEQ ID NO: 9)

SAD1R-F5 gccttgccagcccgc **TCAG**CCAGTATTCATGGCAGGGTT (SEQ ID NO: 10)

El ADN genómico humano (Cornell Medical Institute for Research, Camden, NJ) de 4 individuos se cuantificó basándose en la densidad óptica a 260 nm y se utilizaron 100 ng (aproximadamente 15.000 equivalentes de genoma haploides) como molde para cada reacción de amplificación de PCR. Las reacciones de PCR se llevaron a cabo en condiciones de reacción convencionales (Tris-SO_4 60 mM, pH 8,9, $(NH_4)_2SO_4$ 18 mM), $MgSO_4$ 2,5 mM, dNTP 1 mM, 0,625 uM de cada cebador, 4,5 unidades de polimerasa Platinum Taq High Fidelity (Invitrogen, Carlsbad, CA)) con el siguiente perfil de temperatura: 3 min 94°C; 30 ciclos de 30 s 94°C, 45 s 57°C, 1 min 72°C; 3 min 72°C. Los productos de amplificación se purificaron utilizando un kit de purificación QIAquick PCR (Qiagen, Valencia, CA), y se verificaron sus tamaños esperados (156 a 181 pares de bases) en un aparato 2100 Bioanalyzer de microfluidos utilizando un 500 DNA LabChip® (Agilent Technologies, Inc, Palo Alto, CA). Los amplicones purificados se cuantificaron con un kit de cuantificación de ADN bicatenario PicoGreen® (Molecular Probes, Eugene, OR) y se diluyeron a 10^7 copias por microlitro.

La EBCA (Amplificación Clonal Basada en Emulsión) se realizó como se ha descrito anteriormente con 0,5 amplicones por perla, usando los cebadores de amplificación SAD1F (GCC TCC CTC CCA GCG (SEQ ID NO: 11)) y SAD1R y perlas de captura de Sepharose con cebador de captura SADR1 (GCC TTG CCA GCC CGC (SEQ ID NO: 12)) (Amersham Biosciences, Piscataway, NJ). Todas las manipulaciones adicionales, incluyendo la rotura de las emulsiones y la secuenciación en la placa PicoTiter se realizaron como se ha descrito anteriormente.

Ejemplo 2 Detección de Mutaciones Sensibles

Para demostrar la capacidad del sistema actual (es decir, la plataforma 454) para detectar las variantes de secuencia de baja abundancia, específicamente sustituciones de una única base, se diseñaron experimentos para secuenciar alelos conocidos mezclados a diferentes proporciones.

Se sometieron a ensayo los 6 pares de cebadores enumerados anteriormente para determinar la eficacia de amplificación y se realizó un análisis adicional utilizando los pares SAD1F/R-DD14, SAD1F/R-DE15 y SAD1F/R-F5 que producían todos distintos productos de amplificación (Figura 3). Se amplificaron un total de 8 muestras de ADN genómico humano y se secuenciaron en la plataforma 454 para determinar los genotipos de cada locus. Para simplificar la configuración experimental se realizaron todos los análisis adicionales utilizando el par de cebadores SAD1F/R-DD14 (Figura 3A) y dos muestras que se había demostrado que eran homocigotas para el alelo C o T en el locus concreto.

Los amplicones primarios de cada muestra se cuantificaron y se mezclaron en proporciones específicas que oscilaban de 10:90 a 1:1000, típicamente con el alelo T en exceso. Después de mezclar, las muestras se diluyeron a

una concentración de trabajo de 2 x 10^6 copias por microlitro y se sometieron a EBCA y se secuenciaron en la plataforma 454. La Figura 2 presenta los datos de secuenciación obtenidos a partir de la mezcla del alelo C en proporciones aproximadas 1:500 y 1:1000 en el alelo T. En ambos casos, aproximadamente se generaron más o menos 10.000 lecturas de secuenciación de alta calidad y se sometieron a análisis Blast para identificar sustituciones de nucleótidos contra una secuencia de referencia (en este caso la secuencia que portaba el alelo T). Para la visualización de los resultados se traza la frecuencia de sustitución en una forma codificada por colores con respecto a la secuencia de referencia. Los datos demuestran que en ambas muestras las sustituciones de una sola base de baja frecuencia eran fácilmente identificables (Figura 4A-C). Además, se encontró que el fondo era relativamente consistente entre las muestras permitiendo la sustracción de fondo. Esto produjo típicamente una razón de señal-a-ruido incluso para el alelo 1:1000 que superó 10 (Figura 5A y B). La experimentación adicional utilizando muestras de genotipos conocidos ha confirmado la capacidad para detectar sustituciones de un único nucleótido hasta al menos un nivel de abundancia de 0,1%. El nivel de confianza adicional en los cambios de baja abundancia se puede obtener a partir de la secuenciación de un molde en ambas direcciones. Típicamente, la diferencia entre las frecuencias de los dos conjuntos independientes de datos bidireccionales está a un nivel de abundancia de 20% a 1%.

Para demostrar una respuesta lineal a lo largo de un intervalo más amplio de razones alélicas, se mezclaron amplicones que representan los alelos T y C del locus DD14 HLA a relaciones 1:10, 1:20, 1:50 y 1:200 (10%, 5%, 2% y 0,5%), se amplificaron mediante EBCA y se secuenciaron. La Figura 6 muestra que se observó un aumento lineal en el número relativo del alelo de baja frecuencia a lo largo de todo el intervalo (R^2= 0,9927). Las frecuencias absolutas registradas se desviaron algo de las razones esperadas (Véase la Tabla de más abajo) y se atribuyeron a las dificultades observadas comúnmente al intentar cuantificar con precisión, pequeñas cantidades alícuotas y mixtas de ADN.

Porcentaje de C Esperado	Total de Lecturas	C Esperado	C Observado	T Observado	Porcentaje de C observado
0,00%	101450	0	1	101449	0,00%
0,50%	72406	361	193	72213	0,27%
2,00%	103292	2045	1049	102243	1,02%
2,00%	57115	1131	578	56537	1,01%
5,00%	112378	5452	3340	109038	2,97%
10,00%	104906	9760	7311	97595	6,97%

Resumen de secuenciación utilizado para generar el gráfico de la Figura 6. Los números en las columnas 2-5 indican el número total de moldes secuenciados, y los números esperados y observados para cada alelo, respectivamente.

Ejemplo 3 Proyecto 16S bacteriano - un método para examinar poblaciones de bacterias

Los estudios de poblacionales bacterianas son aplicaciones esenciales para muchos campos, incluyendo el control de procesos industriales, además de la investigación médica, medioambiental y agrícola. Un método común utiliza la secuencia de gen de ARN ribosómico 16S para distinguir especies bacterianas (Jonasson, Olofsson et al. 2002; Grahn, Olofsson et al. 2003). Otro método examina de un modo similar la secuencia intermedia entre los genes de ARN ribosómico 16S y 23S (Garcia-Martinez, Bescos et al. 2001). Sin embargo, para la mayoría de los investigadores es imposible encontrar un censo completo de poblaciones bacterianas complejas usando las tecnologías de preparación y secuenciación de muestras actuales, los requisitos de mano de obra para dicho proyecto o bien son prohibitivamente costosos o bien obligan a un submuestreo drástico de las poblaciones.

En la actualidad, no se utilizan habitualmente métodos de alto rendimiento para examinar poblaciones bacterianas. La práctica común utiliza uno o varios cebadores universales para amplificar el gen de ARN ribosómico 16S (o regiones dentro del gen), que se subclonan con posterioridad en vectores y se secuencian. A menudo se realizan digestiones de restricción en los vectores en un esfuerzo para reducir la carga de secuenciación mediante la eliminación de vectores que muestran patrones de restricción idénticos. Las secuencias resultantes se comparan con una base de datos de genes conocidos procedentes de diferentes organismos; las estimaciones de la composición de la población se extraen de la presencia de secuencias génicas específicas de la especie o el

género. Los métodos de esta descripción tienen el potencial de revolucionar el estudio de las poblaciones de bacterias mediante una drástica reducción de los costes de mano de obra a través de la eliminación de las etapas de clonación y digestión de restricción, incrementando la producción informativa al proporcionar secuencias completas de las regiones de ARN 16S (y posiblemente intergénicas y 23S) permitiendo posiblemente la diferenciación de subcepas no obtenible previamente y proporcionando potencialmente estimaciones de la densidad de especies mediante la conversión del sobremuestreo de la secuencia en una abundancia relativa.

Un método preferido de secuenciación de ácidos nucleicos son los métodos de secuenciación basados en pirofosfato desarrollados por 454 Life Sciences. La utilización de los métodos de la invención, junto con todos los aspectos de la tecnología 454 enormemente paralela (alguno de los cuales se describe en esta memoria) puede aumentar en gran medida el rendimiento y reducir el coste de la identificación de la comunidad. La tecnología de 454 elimina la necesidad de clonar un gran número de productos de PCR individuales, mientras que el pequeño tamaño del gen 16S (1,4 kb) permite procesar simultáneamente decenas de miles de muestras. El procedimiento se ha demostrado satisfactoriamente en la forma indicada a continuación.

Inicialmente, se obtuvo ADN 16S de *Escherichia coli* a partir de células *E. coli* TOP10 competentes (Invitrogen, Carlsbad, CA.) transformadas con el vector pCR2.1, cultivadas en placa sobre placas de LB/ampicilina (50 mg/ml) e incubadas durante la noche a 37°C. Se recogió una sola colonia y se inoculó en 3 ml de caldo LB/ampicilina y se sacudió a 250 RPM durante 6 horas a 37°C. Un microlitro de esta disolución se utilizó como molde para la amplificación de las regiones V1 y V3 de la secuencia 16S.

Se diseñaron cebadores de PCR bipartitos para dos regiones variables en el gen 16S, indicadas como V1 y V3 como describen Monstein et al (Monstein, Nikpour-Badr et al. 2001). Se fusionaron cinco etiquetas de cebado que comprendían los cebadores directo o inverso de 19 bases (cebadores de amplificación de 15 bases, seguidos por una clave de 4 bases (TCGA) 3') específicos de 454 a los cebadores directo e inverso específicos de la región que flanquean las regiones V1 y V3 variables. Esto se puede representar como: 5' - (cebador de Amplificación directo o inverso de 15 bases) - (clave de 4 bases) - (cebador V1 o V3 directo o inverso) - 3'. Los cebadores usados para producir amplicones de 16S contienen las siguientes secuencias, representando las secuencias en mayúscula los cebadores específicos V1 y V3, las cuatro bases en negrita identifican la clave, y las cuatro bases en minúscula indican los cebadores de amplificación 454:

SAD-V1 fusión (directo): gcctccctcgcgcca **TCAG**GAAGAGTTTGATCATGGCTCAG (SEQ ID NO: 13)

SAD-V1 fusión (inverso): gccttgccagcccgc **TCAG**TTACTCACCCGTCCGCCACT (SEQ ID NO: 14)

SAD-V3 fusión (directo): gcctccctcgcgcca **TCAG**GCAACGCGAAGAACCTTACC (SEQ ID NO: 15)

SAD-V3 fusión (inverso): gccttgccagcccgc **TCAG**ACGACAGCCATGCAGCACCT (SEQ ID NO: 16)

Los amplicones V1 y V3 se generaron por separado en reacciones de PCR que contenían los siguientes reactivos: 1X tampón HiFi, $MgSO_4$ 2,5 mM (Invitrogen), dNTPs 1 mM (Pierce, Milwaukee WI.), cebador bipartito directo e inverso 1 µM cada uno para las regiones V1 o V3 (IDT, Coralville, IA), 0,15 U/µl Platinum HiFi Taq (Invitrogen). Se añadió 1 µl de caldo *E. coli*/LB/ampicilina a la mezcla de reacción y se llevaron a cabo 35 ciclos de PCR (94°C durante 30 segundos, 55°C durante 30 segundos, y 68°C durante 150 segundos, seguido el ciclo final de 10°C indefinidamente). Posteriormente, se hizo correr 1 µl de la mezcla de reacción amplificada en el Agilent 2100 Bioanalyzer (Agilent, Palo Alto, CA) para estimar la concentración del producto final, y garantizar que se había generado el producto del tamaño adecuado, 155 pb para el V1, 145 pb para el V3).

A continuación se combinaron productos V1 y V3, se emulsionaron a concentraciones de molde que oscilaban de 0,5 a 10 moléculas de molde por perla de captura de ADN y se amplificaron a través del procedimiento EBCA (Amplificación Clonal Basada en Emulsión) como se esboza en la sección Protocolo de EBCA a continuación. Las perlas amplificadas clonalmente resultantes se secuenciaron con posterioridad en el 454 Genome Sequencer (454 Life Sciences, Branford CT).

Las secuencias obtenidas a partir de las perlas amplificadas se alinearon contra la secuencia del gen 16S de *Escherichia coli* (Entrez gi174375). Los alineamientos aceptables (o "mapeados") se distinguieron de los alineamientos rechazados (o "no mapeados") calculando la puntuación de alineamiento para cada secuencia. La puntuación es el logaritmo medio de la probabilidad de que una señal observada se corresponda con el homopolímero esperado, o:

$$S = \sum \ln[P(s|h)]/N$$

donde S es la puntuación de alineamiento computada, P es la probabilidad de un flujo específico, s es la señal medida a ese flujo, h es la longitud del homopolímero de referencia esperado a ese flujo, y N es el número total de flujos alineados. La puntuación de alineamiento para cada secuencia se comparó después con una puntuación máxima de alineamiento, o MAS; las puntuaciones de alineamiento menores que la MAS fueron

consideradas "reales" y se imprimieron en el archivo de salida. Para este proyecto, se utilizó una MAS de 1,0 (aproximadamente equivalente a 95% de identidad).

Para las secuencias generadas con los cebadores específicos V1, de las 13702 secuencias generadas, 87,75% u 11973 lecturas se mapearon en el genoma con una puntuación de alineamiento menor de 1,0, y una longitud de lectura mayor de 21 bases. En la Figura 7A se muestra una representación gráfica que muestra la localización de las lecturas que se mapean en el fragmento del gen 16S de 1,6 Kb, indicando aproximadamente 12.000 lecturas que se mapean en las primeras 100 bases del gen 16S.

Utilización de BLAST para la secuencia consenso no modificada (AAGAGTTT**t**GATCATGGCTCAGATTGAACGCT-GGCGGCAGGCCTAACACATGCA AGTCGA ACGGTAACAGGA (SEQ ID NO: 17)) contra la base de datos 16S (http://greengenes.llnl.gov) emparejada con *Escherichia coli* como primer organismo conocido

```
>lcl|009704 X80724 Escherichia coli cepa Seattle 1946 ATCC 25922.
        Longitud = 1452
Puntuación = 125 bits (63),Esperado = 1e-28
Identidades = 70/71 (98%),Espacio = 1/71 (1%)
  Hebra = Plus / Plus
Problema: 7  tttgatcatggctcagattgaacgctggcggcaggcctaacacatgcaagtcgaacggta 66
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
 Sujeto: 3  tttgatcatggctcagattgaacgctggcggcaggcctaacacatgcaagtcgaacggta 62

Problema: 67 acgaggaacga 77 (SEQ ID NO:18)
             || ||||||||
 Sujeto: 63 ac-aggaacga 72 (SEQ ID NO:19)

>lcl|090202 AY319393 Escherichia coli cepa  5.2 gen de ARN ribosómico
          16S, secuencia parcial
        Longitud = 1399
Puntuación = 123 bits (62), Esperado = 5e-28
 Identidades  = 62/62 (100%)
  Hebra  = Plus / Plus
```

La secuencia consenso de V1 se editó para AAGAGTTT**T**GATCATGGCTCAGATTGAACGCT-GGCGGCAGGCCTAACACATGCA AGTCGAACGGTAACAGGA (SEQ ID NO: 20), puesto que la cuarta "T" en la posición 9 (marcado en negrita y subrayado) de un tramo del homopolímero se revisó y eliminó, en base a una puntuación del nivel de confianza excesivamente baja. Los resultados de BLAST de la secuencia de V1 editada demostraron la mejora de los éxitos contra los genes 16S de *Escherichia coli*.

```
>lcl|076948 AE016770 Escherichia coli CFT073 sección 16 de 18 del genoma
completo
        Longitud = 1542
Puntuación = 141 bits (71), Esperado = 1e-33
 Identidades = 71/71 (100%)
  Hebra = Plus / Plus
Problema: 1  aagagtttgatcatggctcagattgaacgctggcggcaggcctaacacatgcaagtcgaa 60
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
 Sujeto: 6  aagagtttgatcatggctcagattgaacgctggcggcaggcctaacacatgcaagtcgaa 65

Problema: 61 cggtaacagga 71 (SEQ ID NO:21)
             ||||||||||
 Sujeto: 66 cggtaacagga 76 (SEQ ID NO:22)
```

Se obtuvieron resultados similares con los cebadores específicos V3. De las 17329 lecturas, 71,00% se mapearon en el genoma de referencia de 16S en condiciones de análisis idénticas, a las utilizadas con los moldes de V1 anteriores. Éste es un número menor que 87,75% de las lecturas de V1 que se habían mapeado, y esto puede revelar una mayor divergencia entre la muestra V3 y las secuencias de referencia que entre la muestra V1 y las secuencias de referencia. La secuencia consenso: CAACGCGAAGAACCTTACCTGGTCTTGACATCCACGAA-GTTTAC**T**AGAGATGAG AATGTGCCGTTCGGGAACCG**G**TGAGACAGGTGCTGCATGGCTGTCGTCTg (SEQ ID NO: 23), se mapeó en las regiones 966-1067 del genoma de referencia, como se muestra en la Figura 7B.

A diferencia de de la secuencia V1 los resultados BLAST de la secuencia de consenso no modificada no se emparejaban con *Escherichia coli* como primer organismo conocido, sino más bien como el segundo organismo.

```
>lcl|088104 AJ567617 Escherichia coli gen de ARN 16S parcial, clon
            MBAE104
         Longitud = 1497
Puntuación = 147 bits (74), Esperado = 3e-35
  Identidades = 98/102 (96%), Espacios= 3/102 (2%)
    Hebra = Plus / Plus
Problema: 1    caacgcgaagaaccttacctggtcttgacatccacgaagtttactagagatgagaatgtg 60
               |||||||||||||||||||||||||||||||||||||||| | ||||||||||||||||
 Sujeto:956    caacgcgaagaaccttacctggtcttgacatccacgaagttttc-agagatgagaatgtg 1014

Problema:61    ccgttcgggaaccggtgagacaggtgctgcatggctgtcgtc 102 (SEQ ID NO:24)
               || |||||||||| |||||||||||||||||||||||||||
 Sujeto:1015   cc-ttcgggaacc-gtgagacaggtgctgcatggctgtcgtc 1054 (SEQ ID NO:25)
```

La secuencia de consenso se revisó y se editó para CAACGCGAAGAACCTTACCTGGTCTTGACATCC-ACGAAGTTTACAGAGATGAGA ATGTGCCGTTCGGGAACCGTGAGACAGGTGCTGCATGGCTGTCGTCTg (SEQ ID NO: 26) (con la eliminación de dos bases) basándose en las puntuaciones del nivel confianza, y se volvió a utilizar BLAST. BLAST dio como resultado el éxito más altamente puntuado existente contra *E. coli.*

```
>lcl|088104 AJ567617 Escherichia coli gen de ARN 16S parcial, clon
            MBAE104
         Longitud = 1497
Puntuación = 174 bits (88), Esperado = 1e-43
  Identidades = 98/100 (98%), Espacios= 1/100 (1%)
    Hebra = Plus / Plus
Problema:1     caacgcgaagaaccttacctggtcttgacatccacgaagtttacagagatgagaatgtgc 60
               |||||||||||||||||||||||||||||||||||||||| ||||||||||||||||||
 Sujeto: 956   caacgcgaagaaccttacctggtcttgacatccacgaagttttcagagatgagaatgtgc 1015

Problema:61    cgttcgggaaccgtgagacaggtgctgcatggctgtcgtc 100 (SEQ ID NO:27)
               | ||||||||||||||||||||||||||||||||||||||
 Sujeto:1016   c-ttcgggaaccgtgagacaggtgctgcatggctgtcgtc 1054 (SEQ ID NO:28)
```

Se llevó a cabo un segundo experimento para demostrar la capacidad para utilizar cebadores de PCR mixtos en células bacterianas no procesadas, donde las células de *E. coli* se hicieron crecer hasta saturación y se añadió 1 µl de una dilución 1:1000 del caldo bacteriano a la mezcla de reacción de EBCA en lugar del molde. Los cebadores utilizados en la reacción de EBCA consistieron en cebadores bipartitos específicos de V1 y V3 a 0,04 mM cada uno, así como los cebadores de amplificación directo e inverso de 454 a 0,625 y 0,04 µM respectivamente. Por lo demás, se siguió el protocolo de EBCA esbozado más abajo.

Los datos mostraron que las regiones V1 y V3 podrían ser amplificadas satisfactoriamente, secuenciadas y distinguidas de forma simultánea de una reserva no tratada de células bacterianas. De las 15484 lecturas, 87,66% se mapearon en el genoma de referencia 16S, con las secuencias localizadas en las posiciones de V1 y V3 distintivas mostradas en la Figura 7C.

La capacidad para distinguir entre las secuencias V1 y V3 se evaluó reuniendo 100 lecturas de las secuencias tanto V1 como V3, y convirtiendo de los datos de la señal sin procesar en una cadena binaria, indicando "1" que una base estaba presente a un flujo dado, e indicando "0" que ésta estaba ausente. Los tramos de homopolímero se desplomaron a un valor positivo único, por lo que "A", "AA" y "AAAAA" (SEQ ID NO: 29), recibieron todos una puntuación idéntica de "1". Las cadenas binarias desplomadas se agruparon a continuación a través de la metodología Hierarchical Ordered Partitioning and Collapsing Hybrid (HOPACH) (Pollard y van der Laan 2005) en el paquete estadístico R (Team 2004). El árbol filogenético resultante, que se muestra en la Figura 8, discrimina claramente entre las secuencias V1 (etiquetas rojas de longitud más corta) y V3 (etiquetas azules de mayor longitud) en todas menos 1 de las 200 secuencias.

La capacidad para discriminar esto claramente entre dos regiones similares del mismo gen dentro del mismo organismo sugiere que esta tecnología debe demostrar capacidad para discriminar entre regiones variables de distintos organismos, proporcionando una valiosa herramienta de diagnóstico.

Ejemplo 4 Protocolo EBCA

4.1 Preparación de perlas de captura de ADN

Las perlas empaquetadas de una columna de afinidad Sepharose HP de 1 mL activada con éster de N-hidroxisuccinimida (NHS) (Amersham Biosciences, Piscataway, NJ) se retiraron de la columna y se activaron como se describe en la documentación del producto (Amersham Pharmacia Protocol Núm. 71700600AP). Se unieron a las perlas 25 µl de un cebador de captura HEG marcado con amina 1 mM (5'-Amina-3 espaciadores de hexa-etilenglicol

de 18 átomos secuenciales CCATCTGTTGCGTGCGTGTC-3' (SEQ ID NO: 30)) (IDT Technologies, Coralville, IA, USA) en tampón fosfato 20 mM, pH 8,0, después de lo cual se seleccionaron perlas de 25-36 µm mediante paso en serie a través de secciones de malla para filtro con poro de 36 y 25 µm (Sefar America, Depew, NY, USA). Las perlas de captura de ADN que pasaron a través del primer filtro, pero fueron retenidas por el segundo se recogieron en un tampón de almacenamiento de perlas (Tris 50 mM, Tween al 0,02%, de azida de sodio al 0,02%, pH 8), se cuantificaron con un Contador Multisizer 3 Coulter (Beckman Coulter, Fullerton, CA, USA) y se almacenaron a 4°C hasta que fueron necesarias.

4.2 Especies de moldes de unión para las perlas de captura de ADN

Las moléculas molde se recocieron con cebadores complementarios sobre las perlas de Captura de ADN en una campana de flujo laminar tratada con UV. Se transfirieron seiscientas mil perlas de captura de ADN suspendidas en tampón de almacenamiento de perlas a un tubo de PCR de 200 µL, se centrifugaron en una minicentrífuga de sobremesa durante 10 segundos, el tubo se hizo girar 180° y se centrifugó durante 10 segundos más para asegurar la formación de un sedimento uniforme. El sobrenadante separó a continuación, y las perlas se lavaron con 200 µL de Tampón de Recocido (Tris 20 mM, pH 7,5 y de acetato de magnesio 5 mM), se sometieron a vórtice durante 5 segundos para resuspender las perlas, y se sedimentaron como antes. Se retiró todo menos aproximadamente 10 µL del sobrenadante por encima de las perlas, y se añadieron 200 µL adicionales de tampón de recocido. Las perlas se sometieron a vórtice nuevamente durante 5 segundos, se dejaron reposar durante 1 minuto, después se sedimentaron como antes. Se descartó todo menos aproximadamente 10 µL de sobrenadante, y se añadieron a las perlas 0,48 µL de 2 x 10^7 moléculas por µL de la genoteca de moldes. El tubo sometió a vórtice durante 5 segundos para mezclar el contenido, después de lo cual los moldes se recocieron a las perlas en un programa de desnaturalización/recocido controlado realizado en un termociclador MJ (5 minutos a 80°C, seguido de una disminución de 0,1°C/seg hasta 70°C, 1 minuto a 70C, por disminución de 0,1°C/s hasta 60°C, mantenimiento a 60°C durante 1 minuto, disminución de 0,1°C/s hasta 50°C, mantenimiento a 50°C durante 1 minuto, disminución de 0,1° C/s hasta 20°C, mantenimiento a 20°C). Una vez completado el procedimiento de recocido, las perlas se almacenaron en hielo hasta que fueron necesarias.

4.3 Preparación y formulación de la mezcla de reacción para PCR

Para reducir la posibilidad de contaminación, la mezcla de reacción de PCR se preparó en una campana de flujo laminar tratada con UV localizada en una sala limpia para PCR. Para cada reacción para PCR en emulsión de 600.000 perlas, se prepararon en un tubo de 1,5 mL 225 µl de mezcla de reacción (1X Tampón Platinum HiFi (Invitrogen), dNTP 1 mM (Pierce), $MgSO_4$ 2,5 mM (Invitrogen), BSA de calidad para biología molecular acetilada al 0,1% (Sigma), Tween-80 al 0,01% (Acros Organics), 0,003 U/µL de pirofosfatasa termoestable (NEB), cebadores directo 0,625 mM (5 '- CGTTTCCCCTGTGTGCCTTG-3' (SEQ ID NO: 31)) e inverso 0,039 mM (5'-CCATCTGTTGCG TGCGTGTC-3 '(SEQ ID NO: 32)) (IDT Technologies, Coralville, IA, USA) y 0,15 U/µL de polimerasa Taq Platinum Hi-Fi (Invitrogen)). Se retiraron 25 µL de la mezcla de reacción y se almacenaron en un individuo para PCR de 200 µL individual para su uso como control negativo. Tanto la mezcla de reacción como los controles negativos se almacenaron en hielo hasta que fueron necesarios. Adicionalmente, se prepararon 240 µL de mezcla de amplificación simulada (1X tampón Platinum HiFi (Invitrogen), $MgSO_4$ 2,5 mM (Invitrogen), BSA al 0,1%, Tween al 0,01%) para cada emulsión en un tubo de 1,5 ml, y se almacenaron de manera similar a temperatura ambiente hasta que fueron necesarios.

4.4 Emulsificación y amplificación

El procedimiento de emulsificación crea una emulsión de agua-en-aceite estable al calor con aproximadamente 10.000 microrreactores de PCR discretos por microlitro que sirven como matriz para la amplificación clonal de una sola molécula de las moléculas individuales de la genoteca diana. La mezcla de reacción y las perlas de captura de ADN para una sola reacción se emulsionaron de la siguiente manera: en una campana de flujo laminar tratada con UV, se añadieron 200 µL de una disolución para PCR al tubo que contenía las 600.000 perlas de captura de ADN. Las perlas se resuspendieron a través de la acción repetida de la pipeta, después de lo cual se permitió que la mezcla de PCR-perlas reposara a temperatura ambiente durante al menos 2 minutos, permitiendo que las perlas se equilibraron con la disolución de PCR. Mientras tanto, se tomaron alícuotas de 400 µL de aceite de emulsión (60% (p/p) DC 5225C Formulation Aid (Dow Chemical Co., Midland, MI), 30% (p/p) DC 749 Fluid (Dow Chemical Co., Midland, MI), y 30% (p/p) AR20 Silicone Oil (Sigma)) en un tubo de centrífuga de 2 mL con la parte superior plana (Dot Scientific). Los 240 µL de la mezcla de amplificación simulada se añadieron después a 400 µL de aceite de emulsión, el tubo se tapó de manera segura y se colocó un adaptador TissueLyser de 24 pocillos (Qiagen) de un TissueLyser MM300 (Retsch GmbH & Co. KG, Haan, Alemania). La emulsión se homogeneizó durante 5 minutos a 25 oscilaciones/seg para generar las emulsiones extremadamente pequeñas, o "microfinas", que confieren estabilidad adicional a la reacción.

Durante la formación microfina, se añadieron 160 µL de la mezcla de amplificación de PCR a la mezcla de moldes recocidos y perlas de captura de ADN. Las perlas y la mezcla de reacción para PCR combinadas se sometieron a vórtice brevemente y se dejó que se equilibraran durante 2 minutos. Una vez que hubieron formado las microfinas, se añadieron la mezcla de amplificación, los moldes y las perlas de captura de ADN al material emulsionado. La velocidad del TissueLyser se redujo a 15 oscilaciones por segundo y la mezcla de reacción se homogeneizó durante

5 minutos. La velocidad de homogeneización inferior creó gotitas de agua en la mezcla de aceite con un diámetro medio de 100 a 150 µm, suficientemente grandes para contener perlas de captura de ADN y la mezcla de amplificación.

Se tomaron alícuotas de la emulsión en 7 a 8 tubos para PCR separados que contenían cada uno aproximadamente 80 µL. Los tubos se sellaron y se colocaron en un termociclador MJ junto con los 25 µl de control negativo elaborados previamente. Se utilizaron los siguientes tiempos de ciclos: 1X (4 minutos a 94°C) - Iniciación de Arranque en Caliente "Hot Start Initiation", 40x (30 segundos a 94°C, 60 segundos a 58°C, 90 segundos a 68C) - Amplificación, 13X (30 segundos a 94C, 360 segundos a 58°C) – Prolongación mediante Hibridación. Una vez completado el programa de PCR, las reacciones se retiraron y las emulsiones se rompieron inmediatamente (como se describe a continuación) o las reacciones se almacenaron a 10°C durante hasta 16 horas antes de iniciar el procedimiento de rotura.

4.5 Rotura de la emulsión y recuperación de perlas

Se añadieron 50 µL de alcohol isopropílico (Fisher) a cada tubo de PCR que contenía la emulsión de material amplificado, y se sometieron a vórtice durante 10 segundos para reducir la viscosidad de la emulsión. Los tubos se centrifugaron durante varios segundos en una microcentrífuga para eliminar cualquier material emulsionado atrapado en la tapa del tubo. La mezcla de emulsión-alcohol isopropílico se retiró de cada tubo en una Jeringa BD Desechable de 10 ml (Fisher Scientific) equipada con una aguja roma de calibre 16 (Bricomedical Supplies). Se añadieron 50 µl adicionales de alcohol isopropílico a cada tubo de PCR, se sometieron a vórtice, se centrifugaron como antes, y se añadieron al contenido de la jeringa. El volumen del interior de la jeringa se aumentó a 9 ml con alcohol isopropílico, después de lo cual la jeringa se invirtió y se introdujo 1 ml de aire de la jeringa para facilitar la mezcla del isopropanol y la emulsión. La aguja de punta roma se retiró, se ancló un soporte para filtro Swinlock de 25 mm (Whatman) que contenía Nitex Sieving Fabric (Sefar America, Depew, NY, USA) con un poro de 15 µm a la jeringa tipo Luer, y la aguja de punta roma se fijó en el lado opuesto de la unidad Swinlock.

El contenido de la jeringa se expelió suave pero completamente a través de la unidad de filtro Swinlock y la aguja a un recipiente de desechos con lejía. Se volvieron a transferir seis mililitros de alcohol isopropílico de nueva aportación a la jeringa a través de la aguja de punta roma y la unidad de filtro Swinlock, y la jeringa se invirtió 10 veces para mezclar el alcohol isopropílico, las perlas y componentes de la emulsión restantes. El contenido de la jeringa se expelió de nuevo a un recipiente de desechos, y se repitió dos veces el procedimiento de lavado con 6 ml de alcohol isopropílico adicional en cada lavado. La etapa de lavado se repitió con 6 ml de etanol del 80%/1X tampón de recocido (etanol del 80%, Tris-HCl 20 mM, pH 7,6, acetato de magnesio 5 mM). Las perlas se lavaron a continuación con 6 ml de 1X tampón de recocido con Tween al 0,1% (Tween-20 al 0,1%, Tris-HCl 20 mM, pH 7,6, acetato de magnesio 5 mM), seguido de un lavado con 6 mL con agua PicoPure.

Después de expeler el lavado final al contenedor de desechos, se introdujeron 1,5 ml de EDTA 1 mM en la jeringa, y se retiró la unidad de filtro Swinlock y se dejó a un lado. El contenido de la jeringa se transfirió seriadamente a un tubo de centrífuga de 1,5 ml. El tubo se centrifugó periódicamente durante 20 segundos en una minicentrífuga para sedimentar las perlas y se eliminó el sobrenadante, después de lo cual el contenido restante de la jeringa se añadió al tubo de centrífuga. La unidad Swinlock se volvió a conectar al filtro y se introdujeron 1,5 ml de EDTA en la jeringa. El filtro Swinlock se retiró por última vez, y se añadieron al tubo de centrífuga las perlas y EDTA, sedimentando las perlas y eliminando el sobrenadante según fuera necesario.

4.6 Eliminación de la segunda hebra

El ADN amplificado, inmovilizado sobre las perlas de captura, se volvió monocatenario mediante la eliminación de la segunda hebra a través de incubación en una disolución de fusión alcalina. Se añadió 1 ml de Disolución de Fusión recién preparada (NaOH 0,125 M, NaCl 0,2 M) a las perlas, el sedimento se resuspendió sometiendo a vórtice a un ajuste medio durante 2 segundos, y el tubo se colocó en un rodillo de tubo Thermolyne LabQuake durante 3 minutos. Las perlas se sedimentaron después como anteriormente, y el sobrenadante se retiró cuidadosamente y se descartó. La disolución de fusión residual se diluyó después mediante la adición de 1 ml de Tampón de Recocido (Tris-acetato 20 mM, pH 7,6, acetato de magnesio 5 mM), después de lo cual las perlas se sometieron a vórtice a velocidad media durante 2 segundos, y las perlas se sedimentaron, y sobrenadante se eliminó como antes. El lavado con Tampón de Recocido se repitió, excepto que solo se retiraron 800 µl del Tampón de Recocido después de la centrifugación. Las perlas y el Tampón de Recocido restante se transfirieron a un tubo de PCR de 0,2 ml, y se utilizaron inmediatamente o se almacenaron a 4°C durante hasta 48 horas antes de continuar con el procedimiento de enriquecimiento posterior.

4.7 Enriquecimiento de perlas

Hasta este punto la masa de perlas estaba compuesta tanto por las perlas con hebras de ADN amplificado, inmovilizado, como por las perlas nulas sin producto amplificado. El procedimiento de enriquecimiento se utilizó para capturar selectivamente las perlas con cantidades secuenciables de ADN molde mientras que se rechazaban las perlas nulas.

Las perlas de hebra sencilla de la etapa anterior se sedimentaron por centrifugación durante 10 segundos en una minicentrífuga de sobremesa, después de lo cual el tubo se hizo girar 180º y se centrifugó durante 10 segundos más para asegurar la formación del sedimento uniforme. A continuación se eliminó tanto sobrenadante como fuera posible sin perturbar las perlas. Se añadieron 15 µl de Tampón de Recocido a las perlas, seguido de 2 µl de tampón de enriquecimiento HEG de 40 bases, biotinilado 100 mM, (5' Biotina – espaciador de hexa-etilenglicol de 18 átomos - CGTTTCCCCTGTGTGCCTTGCCATCTGTTCCCTCCCTGTC-3' (SEQ ID NO: 33), IDT Technologies, complementario a los sitios de amplificación y secuenciación combinados (cada uno de 20 bases de longitud) en el extremo 3' del molde inmovilizado en la perla. La disolución se mezcló sometiendo a vórtice a un ajuste medio durante 2 segundos, y los cebadores de enriquecimiento se recocieron a las hebras de ADN inmovilizadas utilizando un programa de desnaturalización/recocido controlado en un termociclador MJ (30 segundos a 65°C, disminución de 0,1°C/seg hasta 58°C, 90 segundos a 58°C, y mantenimiento a 10°C).

Mientras los cebadores se estaban recociendo, se resuspendió una disolución de partida de perlas de estreptavidina magnéticas SeraMag-30 (Seradyn, Indianápolis, IN, USA) mediante agitación suave, y se añadieron 20 µl de perlas SeraMag a un tubo de microcentrífuga de 1,5 ml que contenía 1 ml de Fluido Potenciador (NaCl 2 M, Tris-HCl 10 mM, EDTA 1 mM, pH 7,5). La mezcla de perlas SeraMag se sometió a vórtice durante 5 segundos, y el tubo que se colocó en un imán Dynal MPC-S, sedimentando las perlas paramagnéticas contra el lado del tubo de microcentrífuga. El sobrenadante se retiró cuidadosamente y se descartó sin perturbar las perlas SeraMag, el tubo se retiró del imán, y se añadieron 100 µL de fluido potenciador. El tubo se sometió a vórtice durante 3 segundos para resuspender las perlas y el tubo se almacenó sobre hielo hasta que fue necesario.

Una vez completado el programa de recocido, se añadieron 100 µl de tampón de recocido al tubo de PCR que contenía las perlas de Captura de ADN y el cebador de enriquecimiento, el tubo se sometió a vórtice durante 5 segundos, y el contenido se transfirió a un tubo de microcentrífuga nuevo de 1,5 ml. El tubo de PCR en el que se había recocido el cebador de enriquecimiento a las perlas de captura se lavó una vez con 200 µl de tampón de recocido, y la disolución de lavado se añadió al tubo de 1,5 ml. Las perlas se lavaron tres veces con 1 ml de tampón de recocido, se sometieron a vórtice durante 2 segundos, se sedimentaron como antes, y el sobrenadante se eliminó cuidadosamente. Después del tercer lavado, las perlas se lavaron dos veces con 1 ml de fluido potenciador enfriado con hielo, se sometieron a vórtice, se sedimentaron, y se eliminó el sobrenadante como antes. Las perlas se resuspendieron a continuación en 150 µl de de fluido potenciador enfriado con hielo y la disolución de perlas se añadió a las perlas lavadas SeraMag.

La mezcla de perlas se sometió a vórtice durante 3 segundos y se incubó a temperatura ambiente durante 3 minutos en un rodillo para tubos LabQuake, mientras que las perlas SeraMag recubiertas de estreptavidina unidas a los cebadores de enriquecimiento biotinilados se recocieron a los moldes inmovilizados sobre las perlas de captura de ADN. Las perlas se centrifugaron a continuación a 2.000 rpm durante 3 minutos, después de lo cual las perlas se "golpearon" suavemente hasta que las perlas se resuspendieron. Las perlas resuspendidas se colocaron después sobre hielo durante 5 minutos. Después de la incubación sobre hielo, se añadió Fluido Potenciador frío a las perlas hasta un volumen final de 1,5 ml. El tubo se insertó en un imán Dynal MPC-S, y las perlas se dejaron en reposo durante 120 segundos para permitir que las perlas sedimentaran contra el imán, después de lo cual el sobrenadante (que contenía exceso SeraMag y perlas de captura de ADN nulas) se retiró cuidadosamente y se descartó.

El tubo se retiró del imán MPC-S, se añadió 1 ml de fluido potenciador frío a las perlas, y las perlas se resuspendieron con golpecitos suaves. Fue esencial no someter a vórtice las perlas, ya que el vórtice puede romper la unión entre SeraMag y perlas de captura de ADN. Las perlas se devolvieron al imán, y se eliminó el sobrenadante. Este lavado se repitió tres veces más para asegurar la eliminación de todas las perlas de captura nulas. Para eliminar los cebadores de enriquecimiento recocidos y las perlas SeraMag de las perlas de captura de ADN, las perlas se resuspendieron en 1 ml de disolución de fusión, se sometieron a vórtice durante 5 segundos, y se sedimentaron con el imán. El sobrenadante, que contenía las perlas enriquecidas, se transfirió a un tubo de microcentrífuga 1,5 ml separado, las perlas se sedimentaron y se descartó el sobrenadante. Las perlas enriquecidas se resuspendieron a continuación en 1X Tampón de Recocido con Tween-20 al 0,1%. Las perlas se sedimentaron de nuevo sobre el MPC, y el sobrenadante se transfirió a un tubo de 1,5 ml nuevo, asegurando la eliminación máxima de las perlas SeraMag restantes. Las perlas se centrifugaron, después de lo cual se eliminó el sobrenadante, y las perlas se lavaron 3 veces con 1 ml de 1X Tampón de Recocido. Después del tercer lavado, se retiraron 800 µl del sobrenadante, y las perlas restantes y la disolución se transfirieron a un tubo de PCR de 0,2 ml.

El rendimiento medio para el procedimiento de enriquecimiento fue de 33% de las perlas originales añadidas a la emulsión, o 198.000 perlas enriquecidas por reacción emulsionada. Como el formato PTP 60 x 60 mm requería 900.000 perlas enriquecidas, se procesaron cinco emulsiones de 600.000 perlas por PTP 60 x 60 mm secuenciado.

4.8 Recocido del cebador de secuenciación

Las perlas enriquecidas se centrifugaron a 2.000 rpm durante 3 minutos y se decantó el sobrenadante, después de lo cual se añadieron 15 µl de tampón de recocido y 3 µl de cebador de secuenciación (SAD1F 100 mM (5'-GCC TCC CTC CCA GCG-3' (SEQ ID NO: 34), IDT Technologies). El tubo se sometió después a vórtice durante 5 segundos, y se colocó en un termociclador MJ para el siguiente programa de recocido de 4 etapas: 5 minutos a 65°C,

disminución de 0,1°C/s hasta 50°C, 1 minuto a 50°C, disminución por 0,1°C/seg hasta 40°C, mantenimiento a 40°C durante 1 minuto, disminución de 0,1°C/s hasta 15°C, mantenimiento a 15°C.

Una vez completado el programa de recocido, las perlas se retiraron de termociclador y se sedimentaron mediante centrifugación durante 10 segundos, girando el tubo 180°, y se centrifugaron durante 10 segundos más. El sobrenadante se descartó, y se añadieron 200 µl de tampón de recocido. Las perlas se resuspendieron con un vórtice de 5 segundos, y las perlas se sedimentaron como antes. El sobrenadante se eliminó, y las perlas se resuspendieron en 100 µl de tampón de recocido, momento en el cual las perlas se cuantificaron con un contador Multisizer 3 Coulter. Las perlas se almacenaron a 4°C y fueron estables durante al menos una semana.

4.9 Incubación de perlas de ADN con ADN polimerasa Bst, Fragmento Grande y proteína SSB

Se preparó tampón para el lavado de las perlas (100 ml) mediante la adición de apirasa (Biotage) (actividad final de 8,5 unidades/litro) a 1x tampón de análisis que contenía BSA al 0,1%. La lámina de fibra óptica se retiró del agua PicoPure y se incubó en el tampón para el lavado de perlas. Se centrifugaron novecientas mil de las perlas de ADN preparadas previamente y el sobrenadante se retiró cuidadosamente. Después las perlas se incubaron en 1.290 µl de tampón para el lavado de perlas que contenía 0,4 mg/ml de polivinilpirrolidona (PM 360.000), DTT 1 mM, 175 g de proteína de unión a las cadenas sencillas de *E. coli* (SSB) (United States Biochemicals) y 7000 unidades de ADN polimerasa *Bst*, Fragmento Grande (New England Biolabs). Las perlas se incubaron a temperatura ambiente en un agitador rotativo durante 30 minutos.

4.10 Preparación de perlas de enzima y cargas de micropartículas

Se prepararon UltraGlow Luciferase (Promega) y ATP sulfurilasa *Bst* en la propia empresa como fusiones de proteína transportadora de biotina carboxilo (BCCP). La región de BCCP 87 aminoácidos contiene un residuo de lisina al que está unida covalentemente una biotina durante la expresión *in vivo* de las proteínas de fusión en *E. coli.* La luciferasa biotinilada (1,2 mg) y la sulfurilasa (0,4 mg) se mezclaron previamente y unieron a 4°C a 2,0 ml de perlas paramagnéticas Dynal M280 (10 mg/ml, Dynal SA, Noruega) de acuerdo con las instrucciones del fabricante. Las perlas unidas a enzima se lavaron 3 veces en 2.000 µl de tampón para el lavado de perlas y se resuspendieron en 2.000 µl de tampón para el lavado de perlas.

Se prepararon micropartículas Seradyn (Powerbind SA, 0,8 µm, 10 mg/ml, Seradyn Inc) como sigue: se lavaron 1.050 µl de la provisión de partida con 1.000 µl de 1X tampón de análisis que contenía BSA al 0,1%. Las micropartículas se centrifugaron a 9.300 g durante 10 minutos y se eliminó el sobrenadante. El lavado se repitió 2 veces más y las micropartículas se volvieron a suspender en 1.050 µl de 1X tampón de análisis que contenía BSA al 0,1%. Las perlas y micropartículas se almacenaron sobre hielo hasta su uso.

4.11 Depósito de perlas

Se sometieron a vórtice las perlas con enzima Dynal y las micropartículas Seradyn durante un minuto y se mezclaron 1.000 µl de cada una en un tubo de microcentrífuga nuevo, se sometieron a vórtice brevemente y se almacenaron sobre hielo. Las perlas con enzima/Seradyn (1.920 µl) se mezclaron con las perlas de ADN (1.300 µl) y el volumen final se ajustó a 3.460 µl con tampón para el lavado de perlas. Las perlas se depositaron en capas ordenadas. La lámina de fibra óptica se retiró del tampón para el lavado de las perlas y se depositó la Capa 1, una mezcla de ADN y la perlas con enzima/Seradyn. Después de la centrifugación, se aspiró el sobrenadante de la Capa 1 fuera de la lámina de fibra óptica y se depositó la Capa 2, perlas con enzima Dynal. Esta sección describe con detalle cómo se centrifugaron las distintas capas.

Capa 1. Se ajustó cuidadosamente una junta que crea dos áreas activas de 30x60 mm sobre la superficie de una lámina de fibra óptica de 60x60 mm a los pasadores de acero inoxidable asignados en la parte superior del portapiezas. La lámina de fibra óptica se colocó en el portapiezas con el lado liso sin grabar de la lámina hacia abajo y se ajustó la parte superior del portapiezas/junta se ajustó sobre el lado grabado de la lámina. La parte superior del portapiezas se aseguró después correctamente con los tornillos proporcionados, apretando los extremos opuestos de manera que se atornillaran manualmente. La mezcla de ADN-perlas con enzima se cargó sobre la lámina fibra óptica a través de dos puertos de entrada proporcionados en la parte superior del portapiezas. Se tuvo un cuidado extremo para minimizar las burbujas durante la carga de la mezcla de perlas. Cada depósito se completó con un suave empuje continuo del émbolo de la pipeta. Todo el conjunto se centrifugó a 2.800 rpm en una centrífuga Beckman Coulter Allegra 6 con un rotor GH 3.8-A durante 10 minutos. Después de la centrifugación, el sobrenadante se eliminó con una pipeta.

Capa 2. Se mezclaron las perlas con enzima Dynal (920 µl) con 2.760 µl de tampón para el lavado de perlas y se cargaron 3.400 µl de suspensión de enzima-perlas sobre la lámina de fibra óptica como se ha descrito previamente. El conjunto de la lámina se centrifugó a 2.800 rpm durante 10 min y el sobrenadante se decantó. La lámina de fibra óptica se retiró del portapiezas y se almacenó en tampón para el lavado de perlas hasta que estuvo lista para ser cargada en el aparato.

4,12 Secuenciación en el aparato 454

Todos los reactivos de flujo se prepararon en 1x tampón de análisis con 0,4 mg/ml de polivinilpirrolidona (PM 360.000), DTT 1 mM y Tween 20 al 0,1%. El sustrato (D-luciferina 300 µM (Regis) y fosfosulfato de adenosina 2,5 µM (Sigma)) se preparó en 1X tampón de análisis con 0,4 mg/ml de polivinilpirrolidona (PM 360.000), DTT 1 mM y Tween 20 al 0,1%. El lavado con apirasa se prepara mediante la adición de apirasa para una actividad final de 8,5 unidades por litro en 1X tampón de análisis con 0,4 mg/ml de polivinilpirrolidona (PM 360.000), DTT 1 mM y Tween 20 al 0,1%. Se prepararon desoxinucleótidos dCTP, dGTP y dTTP (GE Biosciences) a una concentración final de 6,5 mM, el α-tiotrifosfato de desoxiadenosina (dATPαS, Biolog) y el pirofosfato de sodio (Sigma) se prepararon a una concentración final de 50 µM y 0,1 µM, respectivamente, en el tampón sustrato.

El aparato de secuenciación 454 consiste de tres conjuntos principales: un subsistema de fluidos, un cartucho con una lámina de fibra óptica/cámara de flujo, y un subsistema de formación de imágenes. Las líneas de entrada de reactivos, un colector de múltiples válvulas, y una bomba peristáltica forman del subsistema de fluidos. Los reactivos individuales se conectan a las líneas de entrada de reactivos apropiados, lo que permite la dispensación de reactivos en la cámara de flujo, un reactivo cada la vez, a una velocidad de flujo y una duración previamente programadas. El cartucho con la lámina de fibra óptica/cámara de flujo tiene un espacio de 250 µm entre el lado grabado de la lámina y el techo de la cámara de flujo. La cámara de flujo también incluye medios para el control de la temperatura de los reactivos y la lámina de fibra óptica, así como una cubierta hermética a la luz. El lado pulido (no grabado) de la lámina se colocó directamente en contacto con el sistema de formación de imágenes.

La liberación cíclica de los reactivos de secuenciación en los pocillos de la lámina de fibra óptica y el lavado de los subproductos de la reacción de secuenciación de los pocillos se consiguió mediante una operación de pre-programada del sistema de fluidos. El programa fue escrito en forma de un paquete de instrucciones de lenguaje de control de interfaz (ICL), especificando el nombre del reactivo (Lavado, dATPαS, dCTP, dGTP, dTTP, y patrón de PPi), la velocidad y la duración del flujo de cada etapa del paquete de instrucciones. La velocidad de flujo se ajustó a 4 ml/min para todos los reactivos y la velocidad lineal dentro de la cámara de flujo fue de aproximadamente ~ 1 cm/s. El orden de flujo de los reactivos de secuenciación se organizó en núcleos donde el primer núcleo consistía en un flujo de PPi (21 segundos), seguido de 14 segundos de flujo de sustrato, 28 segundos de lavado con apirasa y 21 segundos de flujo de sustrato. El primer flujo de PPi estuvo seguido de 21 ciclos de flujos de dNTP (DC-sustrato-lavado con apirasa-sustrato dA-sustrato-lavado con apirasa-sustrato-dG-sustrato-lavado con apirasa-sustrato-dT-sustrato-lavado con apirasa-sustrato), cuando cada flujo de dNTP estaba compuesto de 4 núcleos individuales. Cada núcleo dura 84 segundos (dNTP-21 segundos, flujo de sustrato-14 segundos, lavado con apirasa-28 segundos, flujo de sustrato-21 segundos); se captura una imagen después de 21 segundos y después de 63 segundos. Después de 21 ciclos de flujo de dNTP, se introduce un núcleo de PPi, y seguido a continuación de otros 21 ciclos de flujo de dNTP. El final de la ronda de secuenciación está seguido de un tercer núcleo de PPi. El tiempo total de la ronda fue de 244 minutos. Los volúmenes de reactivo requeridos para completar esta ronda son los siguientes: 500 ml de cada disolución de lavado, 100 ml de cada disolución de nucleótidos. Durante la ronda, todos los reactivos se mantuvieron a temperatura ambiente. La temperatura de la cámara de flujo y el tubo de entrada de la cámara de flujo se controla a 30°C y todos los reactivos que entran a la cámara de flujo son pre-calentado a 30°C.

Ejemplo 5 Análisis de muestras de suelo

El ácido nucleico se extrajo de organismos del suelo para su análisis utilizando los métodos de la invención. La extracción se realizó utilizando un kit de extracción de ADN de Epicentre (Madison, WI, USA) siguiendo las directrices del fabricante.

En resumen, se añadieron 550 ul de Resina de Eliminación del Inhibidor a cada Columna de Centrifugación vacía de Epicentre. Las columnas se centrifugaron durante un minuto a 2.000 xg para cargar la columna. El flujo continuo se retiró y se añadieron otros 550 ul de Resina de Eliminación del Inhibidor a cada columna, seguido de centrifugación durante 2 minutos a 2.000 x g.

Se recogieron 100 mg de suelo en un tubo de 1,5 ml y se añadieron 250 ul de tampón de extracción de ADN del Suelo con 2 ul de proteinasa K. La disolución se sometió a vórtice y se añadieron 50 ul de tampón de Lisis del Suelo y se sometió a vórtice de nuevo. El tubo se incubó a 65C durante 10 minutos y después se centrifugó durante 2 minutos a 1.000 x g. Se transfirieron 180 ul del sobrenadante a un nuevo tubo y se añadieron 60 ul de Reactivo de Precipitación de Proteína mezclando cuidadosamente por inversión del tubo. El tubo se incubó sobre hielo durante 8 minutos y se centrifugó durante 8 minutos a velocidad máxima. Se transfirieron 100-150 ul del sobrenadante directamente a la Columna de Centrifugación preparada y la columna se centrifugó durante 2 minutos a 2.000 xg en el tubo de 1,5 ml. La columna se descartó y se recogió el producto eluido. Se añadieron 6 ul de Disolución de Precipitación de ADN al producto eluido y el tubo se mezcló mediante un breve vórtice. Después de una incubación a temperatura ambiente durante 5 minutos, el tubo se centrifugó durante 5 minutos a velocidad máxima. Se eliminó el sobrenadante y el sedimento se lavó con 500 ul de Disolución de Lavado del Sedimento. El tubo se invirtió para mezclar la disolución y después se centrifugó durante 3 minutos a velocidad máxima. Se eliminó el sobrenadante y se repitió la etapa de lavado. El sobrenadante se eliminó de nuevo y el sedimento final se resuspendió en 300 ul de tampón TE.

La muestra de ADN producida se puede utilizar para los métodos de la invención incluyendo, al menos, los métodos para detectar la frecuencia de nucleótidos en un locus.

Referencias

BioAnalyzer User Manual (Agilent): hypertext transfer protocol://world wide web.chem.agilent.com/temp/rad31B29/00033620.pdf

BioAnalyzer DNA and RNA LabChip Usage (Agilent): hypertext transfer protocol://world wide web.agilent.com/chem/labonachip

BioAnalyzer RNA 6000 Ladder (Ambion): hypertext transfer protocol://world wide web.ambion.com/techlib/spec/sp_7152.pdf

Biomagnetic Techniques in Molecular Biology, Technical Handbook, 3ª edición (Dynal, 1998): hypertext transfer protocol://world wide web.dynal.no/kunder/dynal/DynalPub36.nsf/cb927fbab127a0ad4125683b004b011c/4908f5b1a665858a41256adf005779f2/$FILE/Dynabeads M-280 Streptavidin.pdf.

Dinauer et al., 2000 Sequence-based typing of HLA class II DQB1. Tissue Antigens 55:364.

Garcia-Martinez, J., I. Bescos, et al. (2001). "RISSC: a novel database for ribosomal 16S-23S RNA genes spacer regions." Nucleic Acids Res 29(1): 178-80.

Grahn, N., M. Olofsson, et al. (2003). "Identification of mixed bacterial DNA contamination in broad-range PCR amplification of 16S rDNA V1 and V3 variable regions by pyrosequencing of cloned amplicons." FEMS Microbiol Lett 219(1): 87-91.

Hamilton, S.C., J.W. Farchaus and M.C. Davis. 2001. DNA polymerases as engines for biotechnology. BioTechniques 31:370.

Jonasson, J., M. Olofsson, et al. (2002). "Classification, identification and subtyping of bacteria based on pyrosequencing and signature matching of 16S rDNA fragments." Apmis 110(3): 263-72.

MinElute kit (QIAGEN): hypertext transfer protocol://world wide web.qiagen.com/literature/handbooks/minelute/1016839_HBMinElute_Prot_Gel.pdf.

Monstein, H., S. Nikpour-Badr, et al. (2001). "Rapid molecular identification and subtyping of Helicobacter pylori by pyrosequencing of the 16S rDNA variable V1 and V3 regions." FEMS Microbiol Lett 199(1): 103-7.

Norgaard et al., 1997 Sequencing-based typing of HLA-A locus using mRNA and a single locus-specific PCR followed by cycle-sequencing with AmpliTaq DNA polymerse. Tissue Antigens. 49:455-65.

Pollard, K. S. and M. J. van der Laan (2005). "Clsuter Analysis of Genomic Data with Applications in R." U.C. Berkeley Division of Biostatistics Working Paper Series # 167.

QiaQuick Spin Handbook (QIAGEN, 2001): hypertext transfer protocol://world wide web.qiagen.com/literature/handbooks/qqspin/1016893HBQQSpin_PCR_mc_prot.pdf.

Quick Ligation Kit (NEB): hypertext transfer protocol://world wide web.neb.com/neb/products/mod_enzymes/M2200.html.

Shimizu et al., 2002 Universal fluorescent labeling (UFL) method for automated microsatellite analysis. DNA Res. 9:173-78.

Steffens et al., 1997 Infrared fluorescent detection of PCR amplified gender identifying alleles. J. Forensic Sci. 42:452-60.

Team, R. D. C. (2004). R: A language and environment for statistical computing. Vienna, Austria, R Foundation for Statistical Computing.

Tsang et al., 2004 Development of multiplex DNA electronic microarray using a universal adaptor system for detection of single nucleotide polymorphisms. Biotechniques 36:682-88.

LISTA DE SECUENCIAS

<110> Leamon, John H
Lee, William L
Simons, Jan F
Desany, Brian
Ronan, Mike T
Drake, James
Lohman, Kenton
Egholm, Michael
Rothberg, Jonathan

<120> Métodos Para Determinar Variantes de Secuencia Utilizando Secuenciación Ultraprofunda

<130> 21465-515 UTIL

<140> 11/104,781
<141> 2005-04-12

<160> 34

<170> PatentIn versión 3.3

<210> 1
<211> 41
<212> ADN
<213> Artificial

<220>
<223> oligonucleótidos sintetizados

<400> 1
gcctccctcg cgccatcaga cctccctctg tgtccttaca a 41

<210> 2
<211> 41
<212> ADN
<213> Artificial

<220>
<223> oligonucleótidos sintetizados

<400> 2
gccttgccag cccgctcagg gagggaatca tactagcacc a 41

<210> 3
<211> 43
<212> ADN
<213> Artificial

<220>
<223> oligonucleótidos sintetizados

<400> 3
gcctccctcg cgccatcagt ctgacgatct ctgtcttcta acc 43

<210> 4
<211> 39
<212> ADN
<213> Artificial

<220>
<223> Oligonucleótidos sintetizados

<400> 4
gccttgccag cccgctcagg ccttgaacta cacgtggct 39

<210> 5
<211> 39
<212> ADN
<213> Artificial

<220>
<223> Oligonucleótidos sintetizados

<400> 5
gcctccctcg cgccatcaga tttctctacc acccctggc 39

<210> 6
<211> 39
<212> ADN
<213> Artificial

<220>
<223> oligonucleótidos sintetizados

<400> 6
gccttgccag cccgctcaga gctcatgtct cccgaagaa 39

<210> 7
<211> 39
<212> ADN
<213> Artificial

<220>
<223> Oligonucleótidos sintetizados

<400> 7
gcctccctcg cgccatcaga aagccagaag aggaaaggc 39

<210> 8
<211> 39
<212> ADN
<213> Artificial

<220>
<223> Oligonucleótidos sintetizados

<400> 8
gccttgccag cccgctcagc ttgcagattg gtcataagg 39

<210> 9
<211> 39
<212> ADN
<213> Artificial

<220>
<223> Oligonucleótidos sintetizados

<400> 9
gcctccctcg cgccatcaga cagtgcaaac accaccaaa 39

<210> 10
<211> 39
<212> ADN
<213> Artificial

<220>

<223> Oligonucleótidos sintetizados

<400> 10
gccttgccag cccgctcagc cagtattcat ggcagggtt 39

<210> 11
<211> 15
<212> ADN
<213> Artificial

<220>
<223> Oligonucleótidos sintetizados

<400> 11
gcctccctcg cgcca 15

<210> 12
<211> 15
<212> ADN
<213> Artificial

<220>
<223> Oligonucleótidos sintetizados

<400> 12
gccttgccag cccgc 15

<210> 13
<211> 41
<212> ADN
<213> Artificial

<220>
<223> Oligonucleótidos sintetizados

<400> 13
gcctccctcg cgccatcagg aagagtttga tcatggctca g 41

<210> 14
<211> 39
<212> ADN
<213> Artificial

<220>
<223> Oligonucleótidos sintetizados

<400> 14
gccttgccag cccgctcagt tactcacccg tccgccact 39

<210> 15
<211> 39
<212> ADN
<213> Artificial

<220>
<223> Oligonucleótidos Sintetizados

<400> 15
gcctccctcg cgccatcagg caacgcgaag aaccttacc 39

<210> 16
<211> 39
<212> ADN
<213> Artificial

<220>
<223> oligonucleótidos sintetizados

<400> 16
gccttgccag cccgctcaga cgacagccat gcagcacct 39

<210> 17
<211> 72
<212> ADN
<213> Artificial

<220>
<223> Oligonucleótidos sintetizados

<400> 17
aagagttttg atcatggctc agattgaacg ctggcggcag gcctaacaca tgcaagtcga 60

acggtaacag ga 72

<210> 18
<211> 71
<212> ADN
<213> Escherichia coli

<400> 18
tttgatcatg gctcagattg aacgctggcg gcaggcctaa cacatgcaag tcgaacggta 60

acgaggaacg a 71

<210> 19
<211> 70
<212> ADN
<213> Escherichia coli

<400> 19
tttgatcatg gctcagattg aacgctggcg gcaggcctaa cacatgcaag tcgaacggta 60

acaggaacga 70

<210> 20
<211> 72
<212> ADN
<213> Artificial

<220>
<223> Oligonucleótidos Sintetizados

<400> 20
aagagttttg atcatggctc agattgaacg ctggcggcag gcctaacaca tgcaagtcga 60

acggtaacag ga 72

<210> 21
<211> 71
<212> ADN
<213> Escherichia coli

<400> 21
aagagtttga tcatggctca gattgaacgc tggcggcagg cctaacacat gcaagtcgaa 60

cggtaacagg a 71

<210> 22
<211> 71
<212> ADN
<213> Escherichia coli

<400> 22
aagagtttga tcatggctca gattgaacgc tggcggcagg cctaacacat gcaagtcgaa 60

cggtaacagg a 71

<210> 23
<211> 104
<212> ADN
<213> Artificial

<220>
<223> Oligonucleótidos sintetizados

<400> 23
caacgcgaag aaccttacct ggtcttgaca tccacgaagt ttactagaga tgagaatgtg 60

ccgttcggga accggtgaga caggtgctgc atggctgtcg tctg 104

<210> 24
<211> 102
<212> ADN
<213> Escherichia coli

<400> 24
caacgcgaag aaccttacct ggtcttgaca tccacgaagt ttactagaga tgagaatgtg 60

ccgttcggga accggtgaga caggtgctgc atggctgtcg tc 102

<210> 25
<211> 99
<212> ADN
<213> Escherichia coli

<400> 25
caacgcgaag aaccttacct ggtcttgaca tccacgaagt tttcagagat gagaatgtgc 60

cttcgggaac cgtgagacag gtgctgcatg gctgtcgtc 99

<210> 26
<211> 102
<212> ADN
<213> Artificial

<220>
<223> Oligonucleótidos sintetizados

<400> 26
caacgcgaag aaccttacct ggtcttgaca tccacgaagt ttacagagat gagaatgtgc 60

cgttcgggaa ccgtgagaca ggtgctgcat ggctgtcgtc tg 102

<210> 27
<211> 100
<212> ADN
<213> Escherichia coli

<400> 27
caacgcgaag aaccttacct ggtcttgaca tccacgaagt ttacagagat gagaatgtgc 60

cgttcgggaa ccgtgagaca ggtgctgcat ggctgtcgtc 100

<210> 28
<211> 99
<212> ADN
<213> Escherichia coli

<400> 28
caacgcgaag aaccttacct ggtcttgaca tccacgaagt tttcagagat gagaatgtgc 60

cttcgggaac cgtgagacag gtgctgcatg gctgtcgtc 99

<210> 29
<211> 5
<212> ADN
<213> Artificial

<220>
<223> Oligonucleótidos sintetizados

<400> 29
aaaaa 5

<210> 30
<211> 20
<212> ADN
<213> Artificial

<220>
<223> Oligonucleótidos sintetizados

<400> 30
ccatctgttg cgtgcgtgtc 20

<210> 31
<211> 20
<212> ADN
<213> Artificial

<220>
<223> oligonucleótidos sintetizados

<400> 31
cgtttcccct gtgtgccttg 20

<210> 32
<211> 20
<212> ADN
<213> Artificial

<220>
<223> Oligonucleótidos Sintetizados

<400> 32
ccatctgttg cgtgcgtgtc 20

<210> 33
<211> 40
<212> ADN
<213> Artificial

<220>
<223> Oligonucleótidos Sintetizados

<400> 33
cgtttcccct gtgtgccttg ccatctgttc cctccctgtc 40

<210> 34
<211> 15
<212> ADN
<213> Artificial

<220>
<223> Oligonucleótido sintetizado

<400> 34
gcctccctcg cgcca 15

REIVINDICACIONES

1. Un método para la detección de variantes de secuencia que tienen una frecuencia de menos de 5% en una población de ácidos nucleicos, comprendiendo el método las etapas de:

(a) amplificar un segmento de polinucleótido común a dicha población de ácido nucleico con un par de cebadores de ácidos nucleicos para PCR que definen un locus para producir una primera población de amplicones comprendiendo cada amplicón dicho segmento de polinucleótido;

(b) liberar la primera población de amplicones en microrreactores acuosos en una emulsión de agua-en-aceite de manera que una pluralidad de los microrreactores acuosos comprende (1) un único amplicón de la primera población de amplicones, (2) una única perla, y (3) disolución de reacción de amplificación que contiene los reactivos necesarios para realizar la amplificación de ácido nucleico;

(c) amplificar clonalmente cada miembro de dicha primera población de amplicones mediante reacción en cadena de la polimerasa para producir una pluralidad de poblaciones de segundos amplicones en donde cada población de segundos amplicones deriva de un miembro de dicha primera población de amplicones;

(d) inmovilizar dichos segundos amplicones en una pluralidad de las perlas en los microrreactores de manera que cada perla comprenda una población de dichos segundos amplicones;

(e) romper la emulsión para recuperar las perlas de los microrreactores;

(f) determinar en paralelo una secuencia de ácido nucleico para los segundos amplicones en cada perla, a una profundidad (es decir, número de lecturas de secuencia individuales) de más de 100 para producir una población de secuencias de ácido nucleico; y

(g) determinar una incidencia de cada tipo de nucleótido en cada posición de dicho segmento de polinucleótido para detectar dichas variantes de secuencia en dicha población de ácido nucleico;

en donde el método no requiere conocimiento previo de la composición de secuencia de ácido nucleico de las variantes de secuencia,

y en donde dichos cebadores de ácido nucleico son cebadores bipartitos que comprenden una región 5' y una región 3', en donde dicha región 3' es complementaria a una región en dicho segmento de polinucleótido y en donde dicha región 5' es homóloga a un cebador de secuenciación o un complemento del mismo.

2. Un método de acuerdo con la reivindicación 1, en donde:

la etapa (e) comprende la liberación de las perlas con los segundos amplicones unidos a una formación de cámaras de reacción sobre una superficie plana y la realización de una reacción de secuenciación de forma simultánea sobre la pluralidad de cámaras de reacción para determinar una pluralidad de secuencias de ácido nucleico correspondientes a una pluralidad de alelos.

3. Un método de acuerdo con la reivindicación 1 o 2, en donde el método comprende determinar la secuencia de nucleótidos de una región de polinucleótidos de interés a una profundidad (es decir, número de lecturas de secuencias individuales) de más de 1000.

4. El método de la reivindicación 1, en donde dicha región 5' es homóloga a un oligonucleótido de captura o un complemento del mismo sobre dicho soporte sólido móvil.

5. El método de la reivindicación 1, en donde dicha perla tiene un diámetro seleccionado del grupo que consiste en aproximadamente 1 a aproximadamente 500 micras, entre aproximadamente 5 y aproximadamente 100 micras, entre aproximadamente 10 y aproximadamente 30 micras y entre aproximadamente 15 y aproximadamente 25 micras.

6. El método de la reivindicación 1, en donde dicha perla comprende un oligonucleótido que hibrida e inmoviliza dicha primera población de amplicones, segundos amplicones, o ambos.

7. El método de la reivindicación 1, en donde dicha etapa de determinación de una secuencia de ácido nucleico se lleva a cabo liberando la pluralidad de soportes sólidos móviles en una formación de al menos 10.000 cámaras de reacción sobre una superficie plana, en donde una pluralidad de las cámaras de reacción comprenden no más de un único soporte sólido móvil; y determinando una secuencia de ácido nucleico de los amplicones sobre cada uno de dichos soportes sólidos móviles.

8. El método de la reivindicación 1, en donde dicha etapa de determinación de una secuencia de ácido nucleico se lleva a cabo realiza mediante secuenciación basada en pirofosfato.

9. El método de la reivindicación 1, en donde dicha población de ácido nucleico comprende ADN, ARN, ADNc o una combinación de los mismos.

10. El método de la reivindicación 1, en donde la población de ácido nucleico deriva de una pluralidad de organismos.

11. El método de la reivindicación 1, en donde la población de ácido nucleico deriva de un organismo.

12. El método de la reivindicación 11, en donde dicha población de ácido nucleico deriva de múltiples muestras de tejido de dicho organismo.

13. El método de la reivindicación 11, en donde dicha población de ácido nucleico deriva de un solo tejido de dicho organismo.

14. El método de la reivindicación 1, en donde la población de ácido nucleico es de un tejido enfermo.

15. El método de la reivindicación 14, en donde dicho tejido enfermo comprende tejido tumoral.

16. El método de la reivindicación 1, en donde dicha población de ácido nucleico deriva de un cultivo bacteriano, un cultivo viral, o de una muestra ambiental.

17. El método de la reivindicación 1, en donde la primera población de amplicones tiene de 30 a 500 bases de longitud.

18. El método de la reivindicación 1, en donde dicha primera población de amplicones comprende más de 1000 amplicones, más de 5000 amplicones, o más de 10000 amplicones.

19. El método de la reivindicación 1, en donde cada uno de dichos soportes sólidos se une a al menos 10.000 miembros de dicha pluralidad de segundos amplicones.

20. El método de la reivindicación 1, en donde la secuencia de ácido nucleico de dicho segmento de polinucleótido es indeterminado o parcialmente indeterminado antes de dicho método.

21. Un método para identificar una distribución de organismos en una población que comprende una pluralidad de organismos individuales diferentes, comprendiendo el método el uso de una muestra de ácido nucleico de dicha población y que comprende las etapas de:

(a) determinar las variantes de secuencia de un segmento de ácido nucleico que comprende un locus común a todos los organismos de dicha población utilizando el método de la reivindicación 1, en donde cada organismo comprende una secuencia de ácido nucleico diferente en dicho locus; y

(b) identificar la distribución de los organismos en dicha población basándose en dicha población de secuencias de ácido nucleico.

22. El método de la reivindicación 21, en donde dicha población es una población de organismos seleccionados del grupo que consiste de bacterias, virus, organismos unicelulares, plantas y levaduras.

23. Un método para determinar una composición de una muestra de tejido, comprendiendo el método la utilización de una muestra de ácido nucleico de dicha muestra de tejido y que comprende las etapas de:

(a) detectar variantes de secuencia de un segmento de ácido nucleico utilizando el método de la reivindicación 1, en donde dicho segmento comprende un locus común a todas las células en dicha muestra de tejido y en donde cada tipo de célula comprende una variante de secuencia diferente en dicho locus; y

(b) determinar la composición de dicha muestra de tejido a partir de dicha frecuencia de nucleótidos.

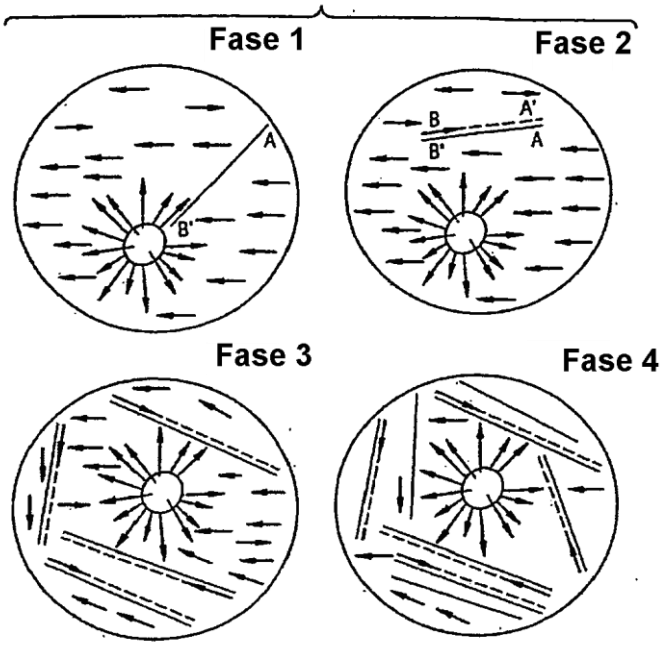


Figura 1A

Figura 1B

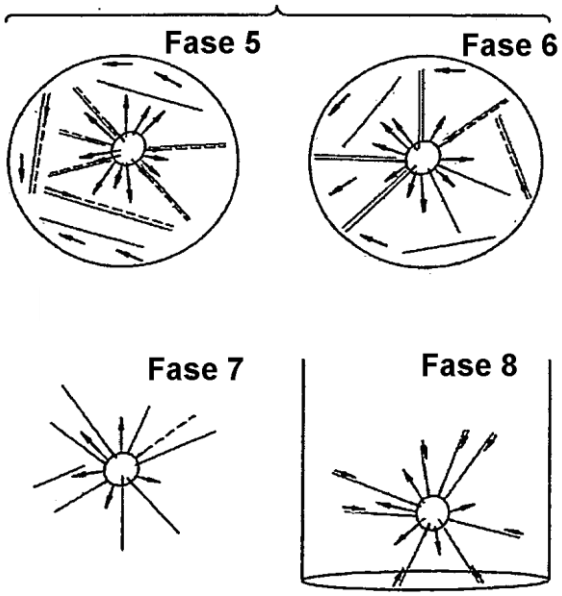


Figura 2

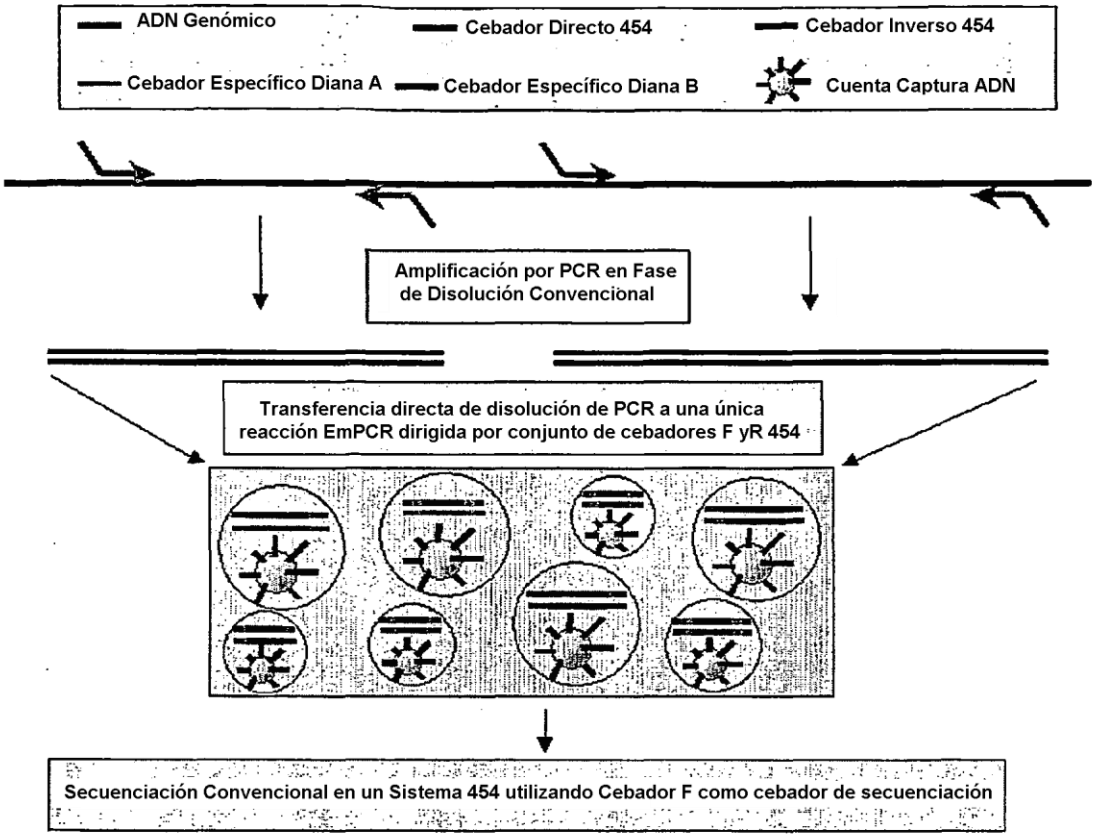
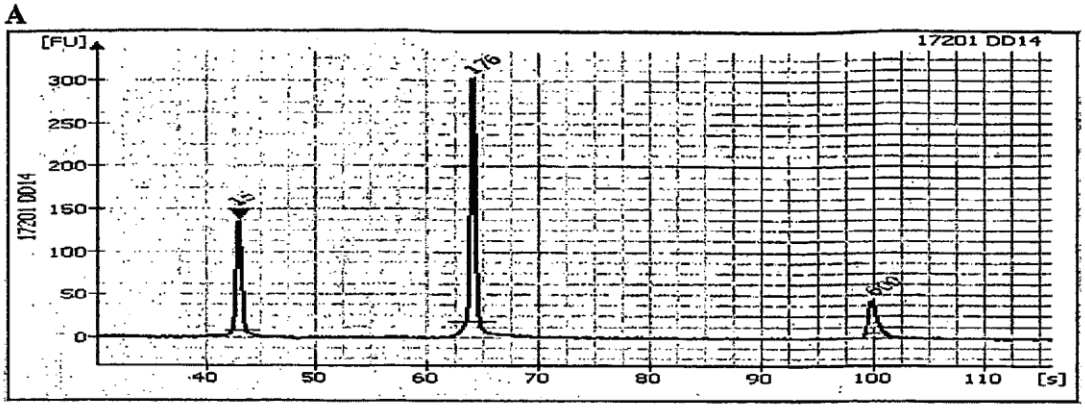
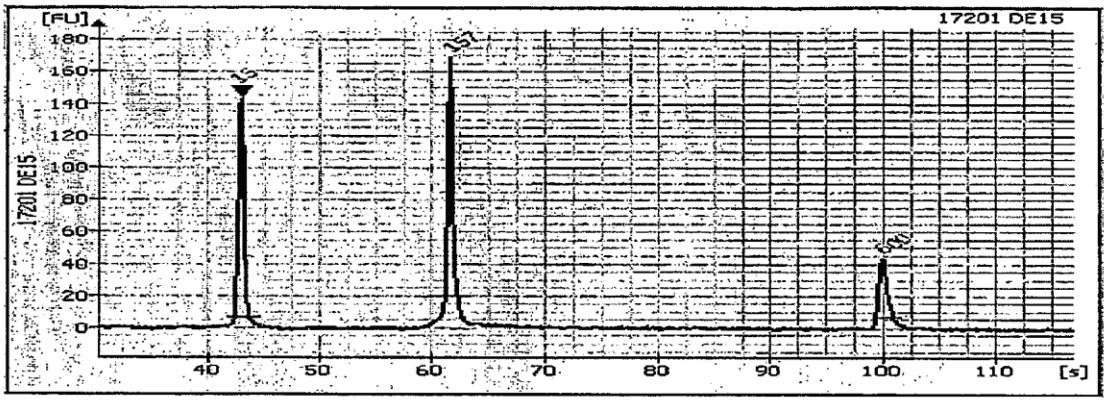

ADN Genómico
Cebador Directo 454
Cebador Inverso 454
Cebador Específico Diana A
Cebador Específico Diana B
Cuenta Captura ADN
Amplificación por PCR en Fase de Disolución Convencional
Transferencia directa de disolución de PCR a una única reacción EmPCR dirigida por conjunto de cebadores F yR 454
Secuenciación Convencional en un Sistema 454 utilizando Cebador F como cebador de secuenciación

Figura 3

A


17201 DD14
[FU]
300
250
200
150
100
50
0
40
50
60
70
80
90
100
110
[s]

B


17201 DE15
[FU]
180
160
140
120
100
80
60
40
20
0
40
50
60
70
80
90
100
110
[s]

C

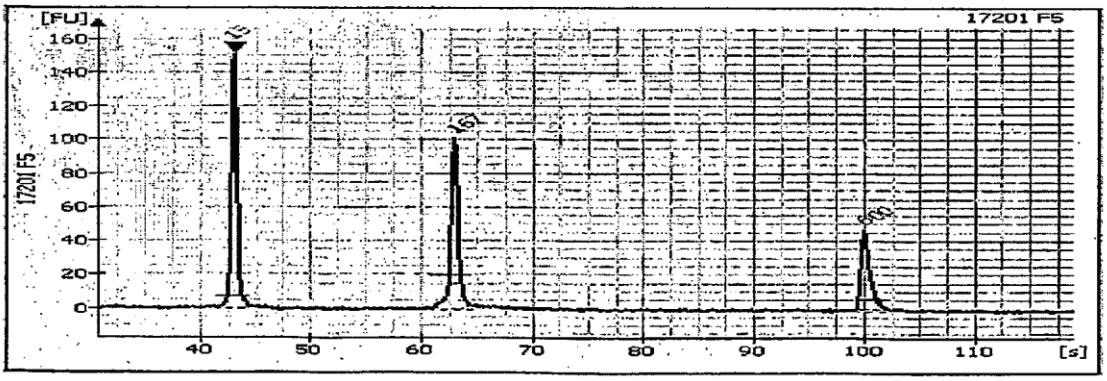

17201 F5
[FU]
160
140
120
100
80
60
40
20
0
40
50
60
70
80
90
100
110
[s]

Figura 4

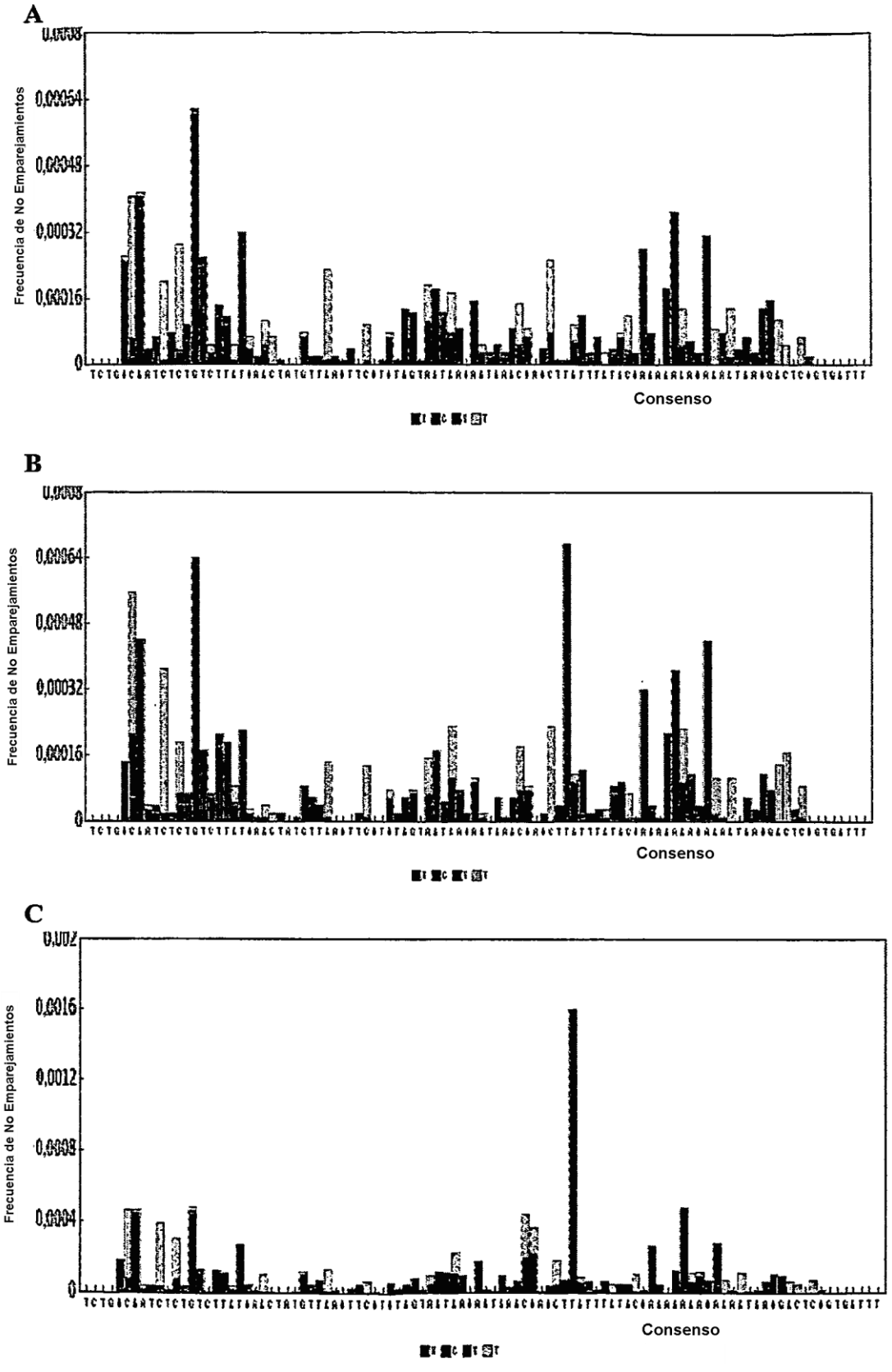

A
Frecuencia de No Emparejamientos
Consenso
B
Frecuencia de No Emparejamientos
Consenso
C
Frecuencia de No Emparejamientos
Consenso

Figura 5

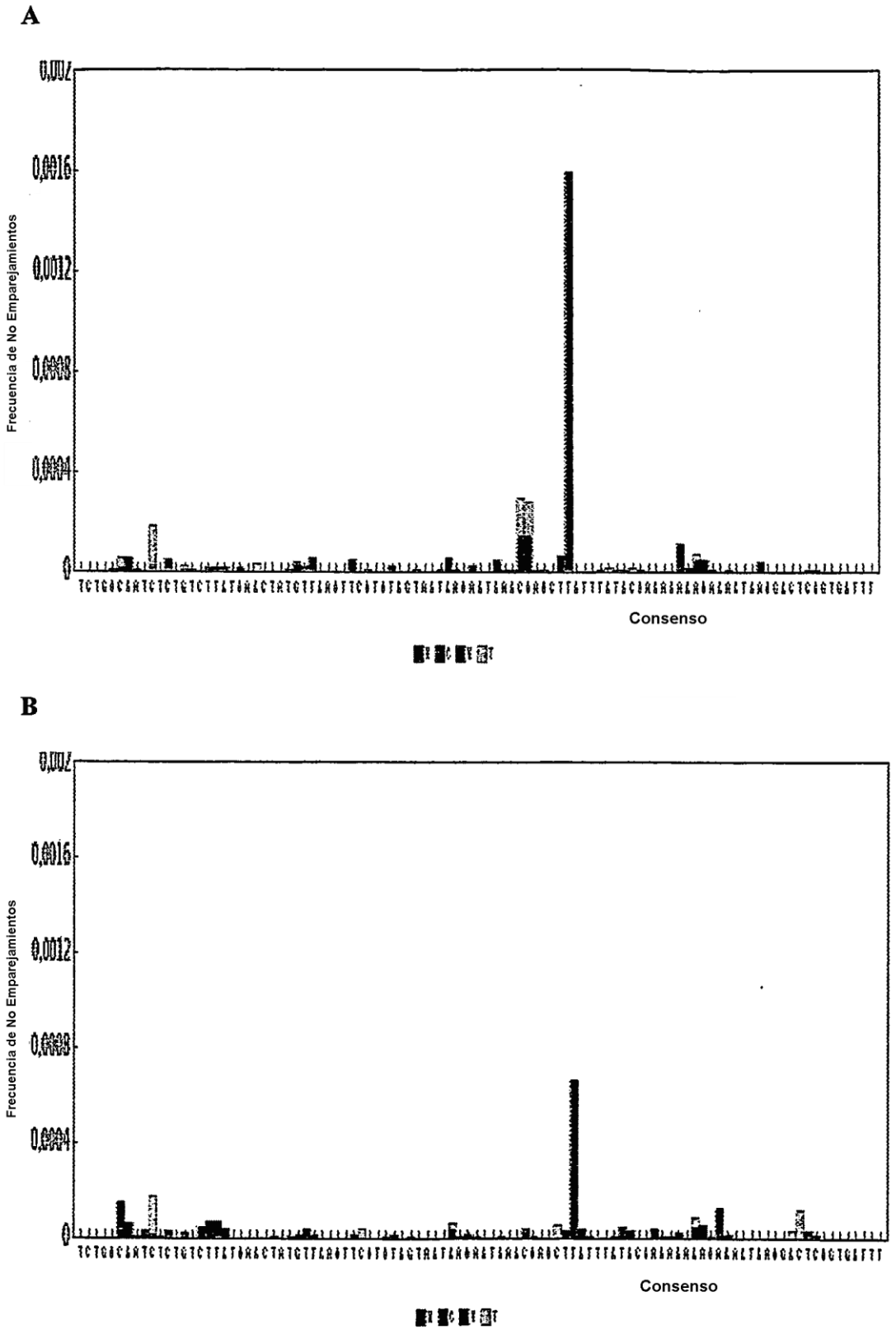

A
Frecuencia de No Emparejamientos
Consenso
B
Frecuencia de No Emparejamientos
Consenso

Figura 6

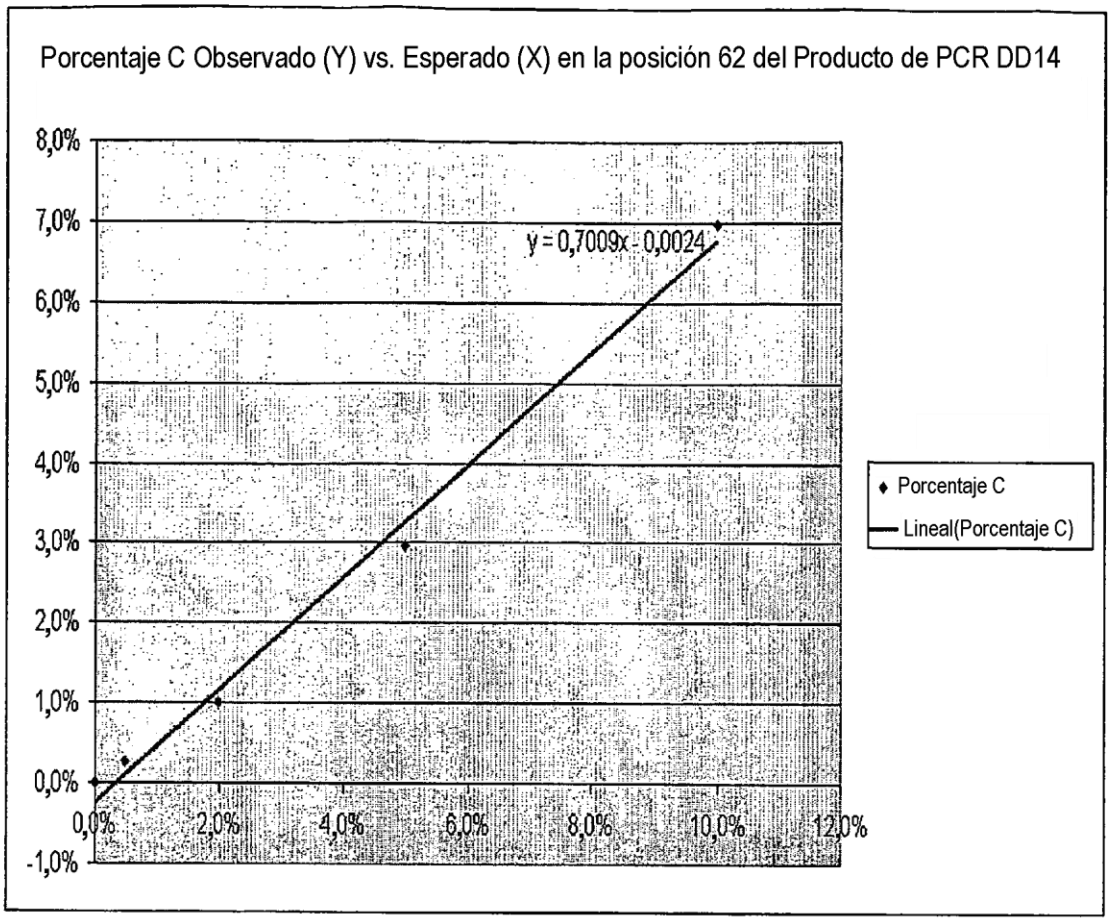
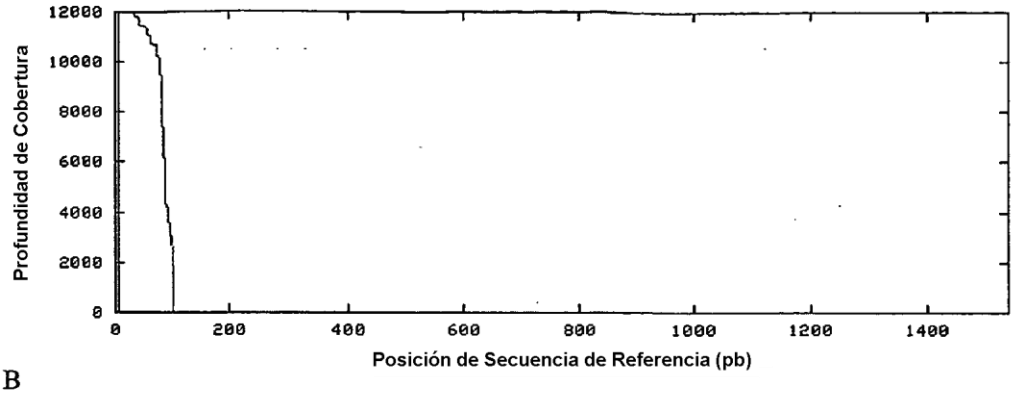
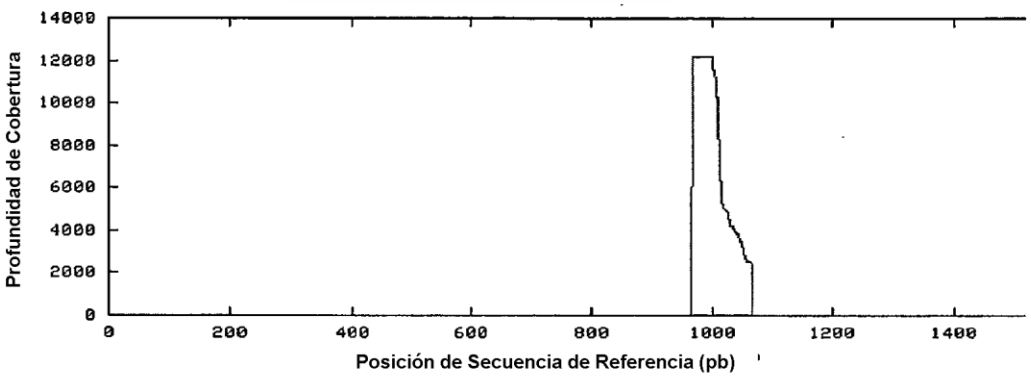

Porcentaje C Observado (Y) vs. Esperado (X) en la posición 62 del Producto de PCR DD14
8,0%
7,0%
6,0%
5,0%
4,0%
3,0%
2,0%
1,0%
0,0%
-1,0%
0,0%
2,0%
4,0%
6,0%
8,0%
10,0%
12,0%
y = 0,7009x - 0,0024
Porcentaje C
Lineal(Porcentaje C)

Figura 7

A


Profundidad de Cobertura
12000
10000
8000
6000
4000
2000
0
0
200
400
600
800
1000
1200
1400
Posición de Secuencia de Referencia (pb)

B


Profundidad de Cobertura
14000
12000
10000
8000
6000
4000
2000
0
0
200
400
600
800
1000
1200
1400
Posición de Secuencia de Referencia (pb)

C

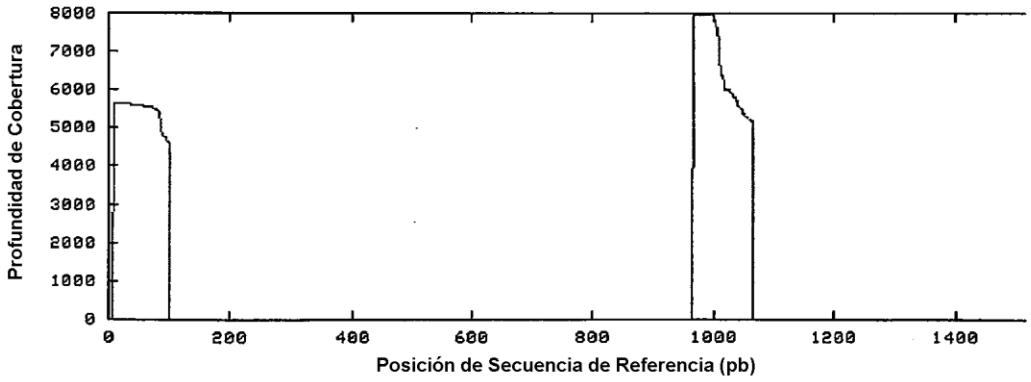

Profundidad de Cobertura
8000
7000
6000
5000
4000
3000
2000
1000
0
0
200
400
600
800
1000
1200
1400
Posición de Secuencia de Referencia (pb)

Figura 8

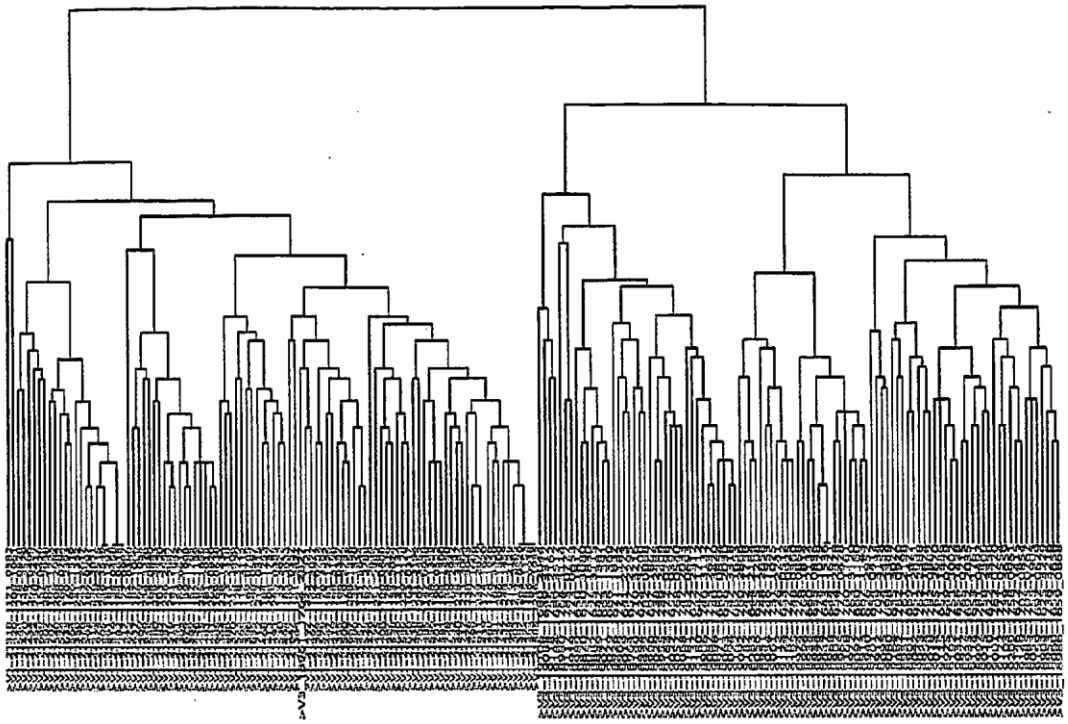


Figura 9

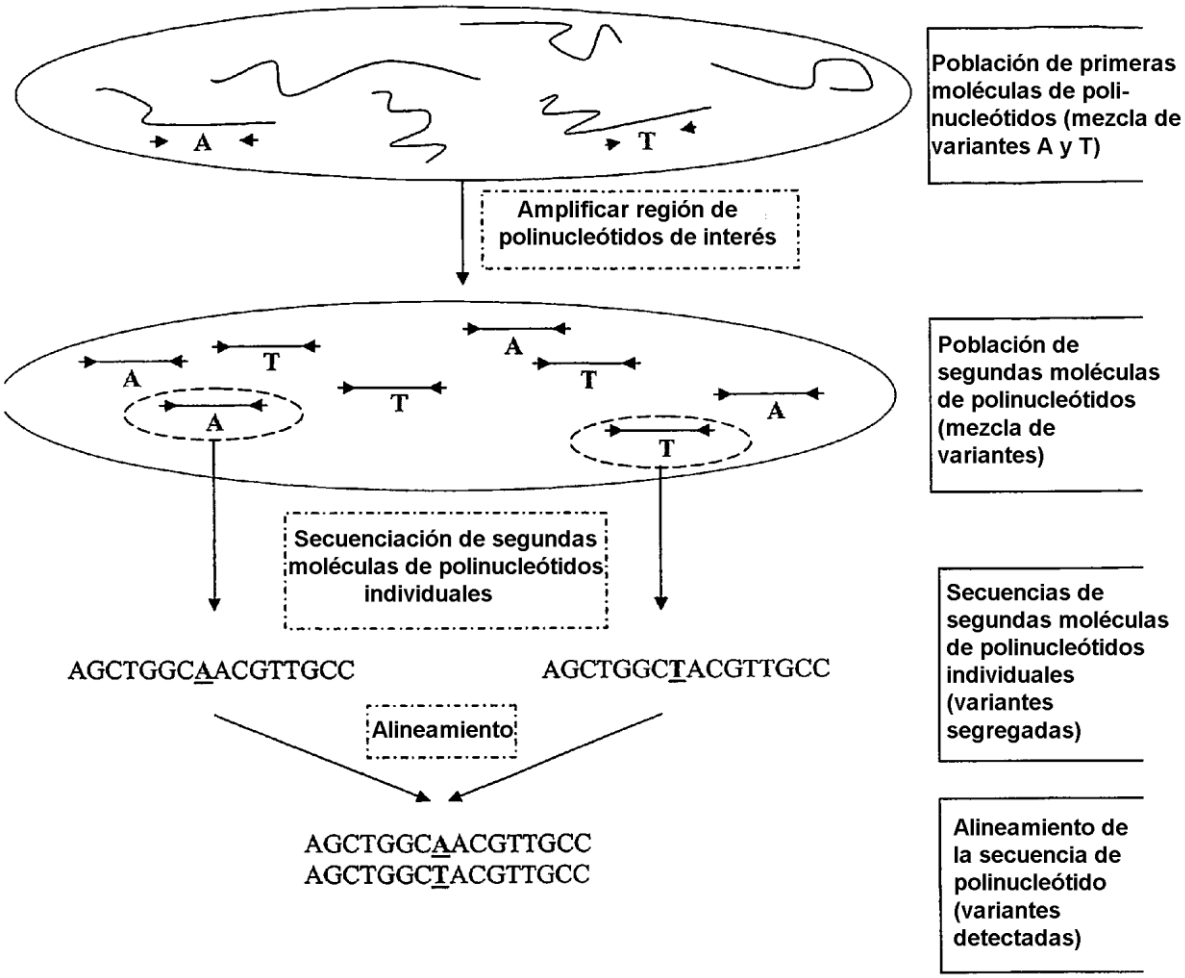
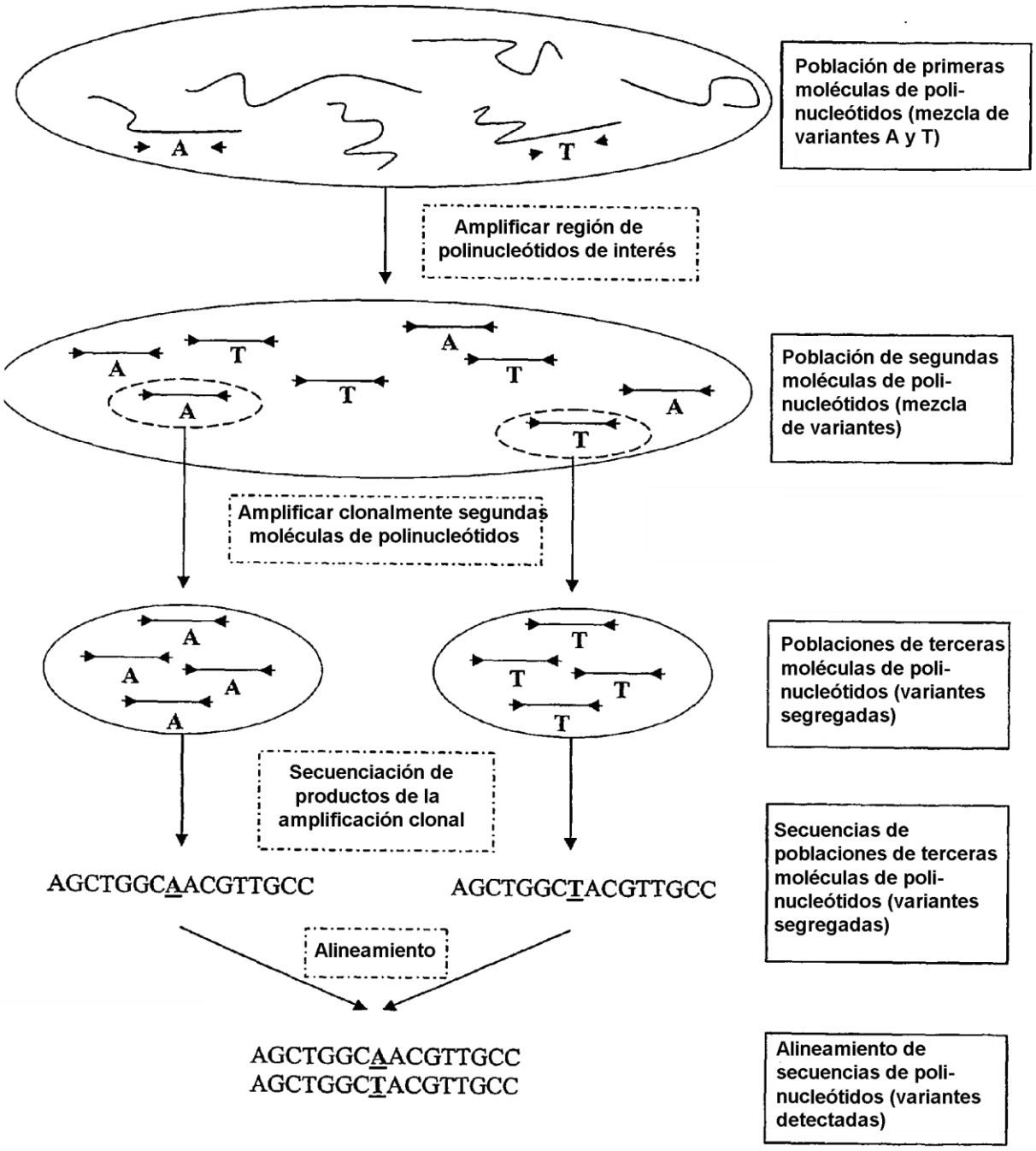

A
T
Población de primeras moléculas de poli-nucleótidos (mezcla de variantes A y T)
Amplificar región de polinucleótidos de interés
Población de segundas moléculas de polinucleótidos (mezcla de variantes)
Secuenciación de segundas moléculas de polinucleótidos individuales
Secuencias de segundas moléculas de polinucleótidos individuales (variantes segregadas)
AGCTGGCAACGTTGCC
AGCTGGCTACGTTGCC
Alineamiento
Alineamiento de la secuencia de polinucleótido (variantes detectadas)
AGCTGGCAACGTTGCC
AGCTGGCTACGTTGCC

Figura 10


Población de primeras moléculas de poli-nucleótidos (mezcla de variantes A y T)
Amplificar región de polinucleótidos de interés
Población de segundas moléculas de poli-nucleótidos (mezcla de variantes)
Amplificar clonalmente segundas moléculas de polinucleótidos
Poblaciones de terceras moléculas de poli-nucleótidos (variantes segregadas)
Secuenciación de productos de la amplificación clonal
Secuencias de poblaciones de terceras moléculas de poli-nucleótidos (variantes segregadas)
AGCTGGCAACGTTGCC
AGCTGGCTACGTTGCC
Alineamiento
Alineamiento de secuencias de poli-nucleótidos (variantes detectadas)
AGCTGGCAACGTTGCC
AGCTGGCTACGTTGCC