(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 690 753**

(51) Int. CI.:

C12Q 1/68 (2008.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(96) Fecha de presentación y número de la solicitud europea: **20.09.2011 E 16170857 (3)**

(97) Fecha y número de publicación de la concesión europea: **25.07.2018 EP 3115468**

(54) Título: **Aumento de la confianza en las identificaciones de alelos con el recuento molecular**

(30) Prioridad:

21.09.2010 US 385001 P
12.01.2011 US 201161432119 P

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
22.11.2018

(73) Titular/es:

AGILENT TECHNOLOGIES, INC. (100.0%)
5301 Stevens Creek Boulevard
Santa Clara, CA 95051, US

(72) Inventor/es:

CASBON, JAMES;
BRENNER, SYDNEY;
OSBORNE, ROBERT;
LICHTENSTEIN, CONRAD y
CLAAS, ANDREAS

(74) Agente/Representante:

ELZABURU, S.L.P

DESCRIPCIÓN

Aumento de la confianza en las identificaciones de alelos con el recuento molecular

El genotipado es una técnica importante de la investigación genética para la cartografía de un genoma y la localización de genes que están ligados a características hereditarias (por ejemplo, enfermedades genéticas). El genotipo de un sujeto generalmente incluye la determinación de los alelos para uno o más locus genómicos en base a los datos de secuenciación obtenidos a partir del ADN del sujeto. Los genomas diploides (por ejemplo, los genomas humanos) se pueden clasificar, por ejemplo, como homocigóticos o heterocigóticos en un locus genómico dependiendo del número de alelos diferentes que tienen para ese locus, donde los individuos heterocigóticos tienen dos alelos diferentes para un locus y los individuos homocigóticos tienen dos copias del mismo alelo para el locus. El genotipado apropiado de las muestras es crucial cuando se realizan estudios en las grandes poblaciones que se necesitan para relacionar el genotipo con el fenotipo con alta confianza estadística.

En el análisis de genotipado de los genomas diploides por secuenciación, se utiliza la cobertura (número de lecturas de secuenciación) para un locus genómico particular, para establecer la confianza de una identificación de alelos. Sin embargo, la confianza en la identificación de alelos se reduce significativamente cuando se introduce un sesgo durante la preparación de la muestra, por ejemplo, cuando la muestra de partida está en cantidades limitantes y/o cuando se emplean una o más reacciones de amplificación para preparar la muestra para la secuenciación. Así, en las muestras que tienen cantidades limitadas de ADN, se puede ver una alta cobertura (es decir, un alto número de lecturas de secuenciación) para un alelo de un cromosoma por encima del alelo de un cromosoma diferente debido al sesgo de la amplificación (por ejemplo, amplificación de solamente algunas, o incluso de solamente una molécula de polinucleótido). En este caso, la cobertura por sí sola puede ser engañosa cuando se mide la confianza en una identificación de alelos.

Ogino y col., Journal of Molecular Diagnostics 4: 185-190 (2002) dan a conocer la cuantificación del sesgo de PCR causado por un polimorfismo de un solo nucleótido en el análisis SMN de dosificación génica.

Wojdacz et al, Epigenetics 4: 231-234 (2009)) dan a conocer el diseño del cebador frente al sesgo de PCR en las amplificaciones por PCR independientes de la metilación.

Tomaz et al, Genetic Testing and Molecular Biomarkers 4: 455-460 (2010) dan a conocer la metilación diferencial como causa del abandono del alelo en el locus *GNAS* impreso.

Grant et al, Nucleic Acids Research 30: e125 (2002) dan a conocer el genotipado de SNP en una plantilla de DOP-PCR amplificada en todo el genoma.

Makrigiorgos y col., Nature Biotechnology 20: 936-939 (2002) dan a conocer un método de amplificación basado en PCR que conserva la diferencia cuantitativa entre dos genomas de comlex.

Weber y col., Analytical Biochemistry 320: 252-258 (2003) dan a conocer un ensayo de reacción en cadena de la polimerasa en tiempo real para la cuantificación de las ratios de los alelos y la corrección del sesgo de amplificación.

La presente invención encuentra utilidad para aumentar la confianza en la identificación de alelos así como en otras aplicaciones basadas en el análisis de secuencias de ácido nucleico, especialmente en el contexto de estudio de genotipos en una población grande de muestras.

De acuerdo con un aspecto de la presente invención se proporciona un método para determinar el número mínimo de moléculas de polinucleótidos individuales que se originan a partir de la misma región genómica de la misma muestra original que han sido secuenciadas en una configuración o procedimiento particular de análisis de secuencias como se indica en la reivindicación 1. Una región de base degenerada (DBR) se une a las moléculas de polinucleótidos de partida que se secuencian posteriormente (por ejemplo, después de haber realizado ciertas etapas del proceso, por ejemplo, amplificación y/o enriquecimiento). Las DBR se pueden usar para mejorar el análisis de muchas aplicaciones diferentes de secuenciación de ácidos nucleicos. Por ejemplo, las DBR permiten la determinación de un valor estadístico para una identificación de alelo en ensayos de genotipado que no pueden obtenerse solo a partir del número de lectura.

En ciertas realizaciones, los aspectos de la presente invención se refieren a métodos para determinar el número de moléculas de polinucleótidos de partida secuenciadas a partir de varias muestras diferentes. En ciertas realizaciones, el método incluye: (1) unir un adaptador a moléculas de polinucleótidos en varias muestras diferentes, donde el adaptador para cada muestra incluye: un MID único y específico para la muestra; y una región de bases degenerada (DBR) (por ejemplo, una DBR con al menos una base de nucleótidos seleccionada entre: R, Y, S, W, K, M, B, D, H, V, N, y versiones modificadas de la misma); (2) reunión de las múltiples muestras unidas a adaptadores diferentes para generar una muestra colectiva; (3) amplificar los polinucleótidos unidos a adaptadores en la muestra colectiva; (4) secuenciación de una pluralidad de los polinucleótidos unidos a adaptadores amplificados, donde la

secuencia de MID, la DBR y al menos una porción del polinucleótido se obtiene para cada uno de la pluralidad de polinucleótidos unidos a adaptadores; y (5) determinar el número de secuencias DBR distintas presentes en la pluralidad de polinucleótidos unidos a adaptadores a partir de cada muestra para determinar o estimar el número de polinucleótidos de partida a partir de cada muestra que se secuenciaron en la etapa de secuenciación.

La invención se entenderá mejor a partir de la siguiente descripción detallada cuando se lea conjuntamente con los dibujos que la acompañan. En los dibujos se incluyen las siguientes Figuras:

Figura 1 que muestra la relación de alelos para cada MID en muestras preparadas a partir de la cantidad indicada de material de partida (parte superior de cada panel; en nanogramos).
Figura 2 que muestra la fracción de secuencias de DBR para cada MID asociadas con cada alelo en una posición polimórfica sintética. Se prepararon muestras a partir de la cantidad indicada de material de partida (parte superior de cada panel; en nanogramos).
Figura 3 que muestra los productos producidos en los dos primeros ciclos de la PCR utilizando cebadores que tienen secuencias de DBR.

A menos que se defina otra cosa, todos los términos técnicos y científicos usados en este documento tienen el mismo significado que es entendido comúnmente por los expertos en la técnica a la que pertenece esta invención. Sin embargo, ciertos elementos se definen en aras de la claridad y facilidad de referencia.

Los términos y símbolos de la química del ácido nucleico, bioquímica, genética y biología molecular que se utilizan aquí siguen los de los tratados y textos estándar de la materia, por ejemplo, Kornberg and Baker, DNA Replication, Second Edition (W.H. Freeman, New York, 1992); Lehninger, Biochemistry, Second Edition (Worth Publishers, New York, 1975); Strachan and Read, Human Molecular Genetics, Second Edition (Wiley-Liss, New York, 1999); Eckstein, editor, Oligonucleotides and Analogs: A practical Approach (Oxford University Press, New York, 1991); Gait, editor, Oligonucleotide Synthesis: A Practical Approach (IRL Press, Oxford, 1984); y similares.

"Amplicón", significa el producto de una reacción de amplificación de polinucleótidos. Es decir, es una población de polinucleótidos, normalmente de doble cadena, que se replican a partir de una o más secuencias de partida. Las una o más secuencias de partida pueden ser una o más copias de la misma secuencia, o pueden ser una mezcla de diferentes secuencias. Los amplicones se pueden producir por una variedad de reacciones de amplificación cuyos productos son múltiples replicados de uno o más ácidos nucleicos diana. En general, las reacciones de amplificación que producen amplicones están "basadas en moldes" porque el apareamiento de bases de los reactantes, ya sean nucleótidos u oligonucleótidos, tienen en un polinucleótido molde los complementos que se requieren para la creación de los productos de reacción. En un aspecto, las reacciones basadas en moldes son extensiones de cebadores con una polimerasa de ácido nucleico o ligamientos de oligonucleótidos con una ligasa de ácido nucleico. Dichas reacciones incluyen, pero no se limitan a, reacciones en cadena de la polimerasa (PCR), reacciones lineales de la polimerasa, amplificación basada en secuencias de ácido nucleico (NASBA), amplificaciones por círculo rodante, y similares, descritas en las siguientes referencias: Mullis *et al.*, patentes de Estados Unidos 4.683.195; 4.965.188; 4.683.202; 4.800.159 (PCR); Gelfand *et al.*, patente de Estados Unidos 5.210.015 (PCR en tiempo real con sondas "TAQMAN™"); Wittwer *et al.*, patente de Estados Unidos 6.174.670; Kacian *et al.*, patente de Estados Unidos 5.399.491 ("NASBA"); Lizardi, patente de Estados Unidos 5.854.033; Aono *et al.*, publicación de patente japonesa JP 4-262799 (amplificación por círculo rodante); y similares. En un aspecto, los amplicones de la invención son producidos por las PCR. Una reacción de amplificación puede ser una amplificación "en tiempo real" si existe disponible una química de detección que permita medir un producto de reacción mientras progresa la reacción de amplificación, por ejemplo, "PCR en tiempo real" descrita a continuación, o "NASBA en tiempo real", tal como se describe en Leone *et al.*, Nucleic Acids Research, 26: 2150-2155 (1998), y referencias similares. Tal como se utiliza aquí, el término "amplificar" significa realizar una reacción de amplificación. Una "mezcla de reacción" significa una solución que contiene todos los reactantes necesarios para realizar una reacción, que pueden incluir, pero no se limita a agentes tampones para mantener el pH a un nivel seleccionado durante una reacción, sales, cofactores, limpiadores, y similares.

El término "evaluar" incluye cualquier forma de medida, e incluye determinar si un elemento está presente o no. Los términos "determinar", "medir", "evaluar", "estimar", "calcular" y "ensayar" se usan de modo intercambiable, e incluyen las determinaciones cuantitativas y cualitativas. La evaluación puede ser relativa o absoluta. "Evaluar la presencia de" incluye determinar la cantidad de algo presente, y/o determinar si algo está presente o ausente.

Los polinucleótidos que son "marcados asimétricamente" tienen dominios de adaptadores a izquierda y derecha que no son idénticos. Este procedimiento se denomina genéricamente como de unión de adaptadores que marcan asimétrica o asimétricamente un polinucleótido, por ejemplo, un fragmento de polinucleótido. Se puede conseguir de cualquier manera conveniente la producción de polinucleótidos que tengan terminales de adaptadores asimétricos. Ejemplos de adaptadores asimétricos están descritos en: las patentes de Estados Unidos 5.712.126 y 6.372.434; publicaciones de patentes de Estados Unidos 2007/0128624 y 2007/0172839; y la publicación PCT WO 2009/032167. En ciertas realizaciones, los adaptadores asimétricos empleados son los descritos en la Patente de EE.UU. nº de publicación 2009/0275087.

Como ejemplo, un usuario de la presente invención puede utilizar un adaptador asimétrico para marcar los polinucleótidos. Un "adaptador asimétrico" es uno que, cuando se liga a ambos extremos de un fragmento de ácido nucleico de doble cadena, dará lugar a la producción de productos de extensión o amplificación de cebadores que tienen secuencias no idénticas que flanquean el inserto genómico de interés. El ligamiento va generalmente seguido por etapas de procedimiento posteriores de modo que se generan las secuencias de adaptadores de terminales no idénticos. Por ejemplo, la replicación de un fragmento o fragmentos unidos a un adaptador asimétrico da como resultado productos polinucleótidos en los que hay al menos una diferencia en la secuencia de ácido nucleico, o modificación de nucleótido/nucleósido, entre las secuencias terminales del adaptador. La unión de adaptadores asimétricamente a los polinucleótidos (por ejemplo, fragmentos de polinucleótidos) da como resultado polinucleótidos que tienen una o más secuencias de adaptador en un extremo (por ejemplo, una o más regiones o dominios, por ejemplo, un sitio de unión del cebador) que están o no presentes o que tienen una secuencia de ácido nucleico diferente en comparación con la secuencia del adaptador en el otro extremo. Se observa que un adaptador que se denomina "adaptador asimétrico" no es necesariamente estructuralmente asimétrico por sí mismo, ni el mero acto de unir un adaptador asimétrico a un fragmento de polinucleótido le vuelve inmediatamente asimétrico. Más bien, un polinucleótido unido a un adaptador asimétrico, que tiene un adaptador asimétrico idéntico en cada extremo, produce productos de replicación (o polinucleótidos aislados de cadena simple) que son asimétricos con respecto a las secuencias del adaptador de los extremos opuestos (por ejemplo, después de una ronda al menos de amplificación/extensión del cebador).

En la práctica de la presente invención se puede emplear, cualquier adaptador asimétrico conveniente, o cualquier procedimiento para unir adaptadores asimétricamente. Ejemplos de adaptadores asimétricos se describen en: patentes de Estados Unidos 5.712.126 y 6.372.434; publicaciones de patentes de Estados Unidos 2007/0128624 y 2007/0172839; y la publicación PCT WO 2009/032167. En ciertas realizaciones, los adaptadores asimétricos empleados son los descritos en la Patente de EE.UU. nº de publicación 2009/0275087.

"Complementario" o "sustancialmente complementario" se refiere a la hibridación o apareamiento de bases o a la formación de un dúplex entre nucleótidos o ácidos nucleicos, tales como, por ejemplo, entre las dos cadenas de una molécula de ADN bicatenario o entre un cebador de oligonucleótidos y un sitio de unión del cebador en un ácido nucleico monocatenario. Los nucleótidos complementarios son, en general, A y T (o A y U), o C y G. Se dice que dos moléculas de ARN o ADN monocatenarios son sustancialmente complementarias cuando los nucleótidos de una cadena, alineados óptimamente y comparados con inserciones o deleciones de nucleótidos apropiados, se aparean con al menos aproximadamente el 80 % de los nucleótidos de la otra cadena, normalmente con al menos aproximadamente el 90 % al 95 %, y más preferiblemente con aproximadamente el 98 al 100 %. Alternativamente, existe complementariedad sustancial cuando una cadena de ARN o ADN se pueda hibridar en condiciones de hibridación selectivas con su complemento. Típicamente, la hibridación selectiva se producirá cuando exista al menos aproximadamente el 65 % de complementariedad en un tramo de al menos 14 a 25 nucleótidos, preferiblemente al menos aproximadamente el 75 %, más preferiblemente al menos aproximadamente el 90 % de complementariedad. Véase, M. Kanehisa Nucleic Acids Res. 12: 203 (1984).

"Dúplex" significa al menos dos oligonucleótidos y/o polinucleótidos que son total o parcialmente complementarios, sometidos a apareamiento de bases de tipo Watson-Crick entre todos o la mayor parte de sus nucleótidos de manera que se forma un complejo estable. Los términos "anillamiento" e "hibridación" se usan de modo intercambiable para referirse a la formación de un dúplex estable. "Perfectamente ajustado" en referencia a un dúplex significa que las cadenas de polinucleótidos u oligonucleótidos que constituyen el dúplex forman entre sí una estructura de doble cadena de tal manera que todos los nucleótidos de cada cadena experimentan el apareamiento de bases de Watson-Crick con un nucleótido de la otra cadena. Un dúplex estable puede incluir apareamiento de bases de Watson-Crick y/o apareamiento de bases que no es de Watson-Crick entre las cadenas del dúplex (donde el apareamiento de bases significa la formación de enlaces de hidrógeno). En ciertas realizaciones, un par de bases que no es de Watson-Crick incluye un análogo de nucleósido, tal como desoxiinosina, 2,6-diaminopurina, los PNA, LNA y similares. En ciertas realizaciones, un par de bases que no es de Watson-Crick incluye una "base oscilante", tal como desoxiinosina, 8-oxo-dA, 8-oxo-dG y similares, donde por "base oscilante" se entiende una base de ácido nucleico que puede formar un par de bases con una primera base de nucleótido en una cadena complementaria de ácido nucleico, pero que, cuando se emplea como una cadena molde para la síntesis de ácido nucleico, lleva a la incorporación de una segunda base de nucleótido diferente a la cadena de síntesis (las bases oscilantes se describen en más detalle a continuación). Un "desajuste" en un dúplex entre dos oligonucleótidos o polinucleótidos significa que un par de nucleótidos en el dúplex no experimenta la unión de Watson-Crick.

"Locus genético", "locus", o "locus de interés", en referencia a un genoma o polinucleótido diana, significa una sub-región o segmento contiguos del genoma o del polinucleótido diana. Como se usa en este documento, locus genético, locus o locus de interés se puede referir a la posición de un nucleótido, un gen o una porción de un gen en un genoma, incluyendo el ADN mitocondrial u otro ADN no cromosómico (por ejemplo, plásmido bacteriano), o se puede referir a cualquier porción contigua de la secuencia genómica lo mismo que esté o no dentro o asociada con un gen. Un locus genético, locus o locus de interés puede tener desde un solo nucleótido hasta un segmento de unos cientos o unos miles de nucleótidos de longitud o más. En general, un locus de interés tendrá una secuencia

de referencia asociada con el mismo (véase la descripción de "secuencia de referencia" más adelante).

"Kit" se refiere a cualquier sistema de suministro para suministrar materiales o reactivos para llevar a cabo un método de la invención. En el contexto de los ensayos de reacción, dichos sistemas de suministro incluyen sistemas que permiten el almacenaje, transporte o suministro de los reactivos de reacción (por ejemplo, sondas, enzimas, etc., en los recipientes apropiados) y/o de los materiales de soporte (por ejemplo, tampones, instrucciones escritas para realizar el ensayo, etc.) de un lugar a otro. Por ejemplo, los kits incluyen uno o más dispositivos cerrados (por ejemplo, cajas) que contienen los reactivos de reacción y/o los materiales de soporte pertinentes. Dichos contenidos se pueden suministrar al receptor propuesto juntos o por separado. Por ejemplo, un primer recipiente puede contener una enzima para su uso en un ensayo, mientras que un segundo recipiente contiene sondas.

"Ligamiento" significa formar un enlace o unión covalente entre los extremos de dos o más ácidos nucleicos, por ejemplo, oligonucleótidos y/o polinucleótidos. La naturaleza del enlace o unión puede variar ampliamente y el ligamiento se puede llevar a cabo por medios enzimáticos o químicos. Como se usa en este documento, los ligamientos se realizan normalmente por medios enzimáticos para formar un enlace fosfodiéster entre el carbono 5' de un nucleótido terminal de un oligonucleótido con el carbono 3' de otro oligonucleótido. En las siguientes referencias se describen diversas reacciones de ligamiento basadas en moldes: Whiteley *et al.*, patente de Estados Unidos 4.883.750; Letsinger *et al.*, patente de Estados Unidos 5.476.930; Fung *et al.*, patente de Estados Unidos 5.593.826; Kool, patente de Estados Unidos 5.426.180; Landegren *et al.*, patente de Estados Unidos 5.871.921; Xu and Kool, Nucleic Acids Research, 27: 875-881 (1999); Higgins *et al.*, Methods in Enzymology, 68: 50-71 (1979); Engler *et al.*, The Enzymes, 15: 3-29 (1982); y Namsaraev, publicación de patente de Estados Unidos 2004/0110213.

"Identificador multiplex" (MID) tal como se utiliza aquí se refiere a una marca o combinación de marcas asociadas con un polinucleótido cuya identidad (por ejemplo, la marca de una secuencia de ADN) se puede utilizar para diferenciar polinucleótidos en una muestra. En ciertas realizaciones, se utiliza el MID sobre un polinucleótido para identificar la fuente de la cual se deriva el polinucleótido. Por ejemplo, una muestra de ácido nucleico puede ser una agrupación de polinucleótidos derivados de diferentes fuentes, (por ejemplo, polinucleótidos derivados de diferentes individuos, diferentes tejidos o células, o polinucleótidos aislados a diferentes tiempos), donde los polinucleótidos de cada fuente diferente se marcan con un único MID. Como tal, un MID proporciona una correlación entre un polinucleótido y su fuente. En ciertas realizaciones, los MID se emplean para marcar de forma única cada polinucleótido individual en una muestra. La identificación del número de los MID únicos en una muestra puede proporcionar una lectura de cuántos polinucleótidos individuales están presentes en la muestra (o de cuántos polinucleótidos originales se ha derivado una muestra de polinucleótido manipulado; véase, por ejemplo, la patente de Estados Unidos 7.537.897, expedida el 26 de mayo de 2009). Los MID comprenden típicamente bases nucleotídicas y pueden variar en longitud de 2 a 100 bases de nucleótidos o más, y pueden incluir múltiples subunidades, donde cada MID diferente tiene una identidad y/o un orden de subunidades definidos. Ejemplos de marcas de ácido nucleico que encuentran utilidad como MID se describen en la patente de Estados Unidos 7.544.473, expedida el 6 de junio 2009, y titulada "Nucleic Acid Analysis Using Sequence Tokens", así como la patente de Estados Unidos 7.393.665, expedida el 1 de julio 2008, y titulada "Methods and Compositions for Tagging and Identifying Polynucleotides" que describen etiquetas de ácido nucleico y su uso en la identificación de polinucleótidos. En ciertas realizaciones, un conjunto de MID empleados para marcar una pluralidad de muestras no necesita tener ninguna propiedad particular común (por ejemplo, T_m, longitud, composición de bases, etc.), ya que los métodos descritos en el presente documento se pueden adaptar a una amplia variedad de conjuntos de MID únicos. Se destaca aquí que los MID solamente necesitan ser únicos dentro de un experimento dado. Por lo tanto, el mismo MID se puede utilizar para marcar una muestra diferente que está siendo procesada en un experimento diferente. Además, en ciertos experimentos, el usuario puede utilizar el mismo MID para marcar un subconjunto de muestras diferentes dentro del mismo experimento. Por ejemplo, todas las muestras derivadas de individuos que tienen un fenotipo específico pueden ser marcadas con el mismo MID, por ejemplo, todas las muestras derivadas de sujetos testigos (o de tipo natural) se pueden marcar con un primer MID mientras que los sujetos que tienen una enfermedad se pueden marcar con un segundo MID (diferente del primer MID). Como otro ejemplo, puede ser deseable marcar diferentes muestras derivadas de la misma fuente con diferentes MID (por ejemplo, muestras derivadas a lo largo del tiempo o derivadas de diferentes sitios dentro de un tejido). Además, los MID se pueden generar en una variedad de diferentes maneras, por ejemplo, mediante un método combinatorio de marcado en el que un MID se une mediante ligamiento y un segundo MID se une por extensión del cebador. Por lo tanto, los MID se pueden diseñar e implementar en una variedad de maneras diferentes para hacer un seguimiento de los fragmentos de polinucleótidos durante el procesado y el análisis, y por lo tanto no se pretende ninguna limitación a este respecto.

"Secuenciación de nueva generación" (NGS), como se usa aquí se refiere a las tecnologías de secuenciación que tienen la capacidad de secuenciar polinucleótidos a velocidades que no tienen precedentes utilizando los métodos de secuenciación convencionales (por ejemplo, métodos estándar de secuenciación de Sanger o Maxam-Gilbert). Estas velocidades sin precedentes se alcanzan mediante la realización y lectura de miles a millones de reacciones de secuenciación en paralelo. Las plataformas de secuenciación por NGS incluyen, pero no se limitan a las siguientes: Massively Parallel Signature Sequencing (Lynx Therapeutics); 454 pyro-sequencing (454 Life

Sciences/Roche Diagnostics); secuenciación con terminador de tinte reversible, en fase sólida (Solexa/Illumina); Tecnología SOLiD (Applied Biosystems); secuenciación de semiconductor iónico (Ion Torrent); y la secuenciación del ADN Nanoball (Complete Genomics). Descripciones de ciertas plataformas de NGS se pueden encontrar en lo siguiente: Shendure, *et al.*, "Next-generation DNA sequencing", Nature, 2008, vol. 26, No. 10, 1135-1145; Mardis, "The impact of next-generation sequencing technology on genetics" Trends in Genetics, 2007, vol. 24, No. 3, pp 133-141; Su, *et al.*, "Next-generation sequencing and its applications in molecular diagnostics" Expert Rev Mol Diagn, 2011, 11(3): 333-43; y Zhang *et al.*, "The impact of next-generation sequencing on genomics", J Genet Genomics, 2011, 38(3): 95-109.

"Nucleósido" tal como se usa en el presente documento incluye los nucleósidos naturales, incluyendo las formas 2'-desoxi y 2'-hidroxilo, por ejemplo, como se describe en Kornberg and Baker, DNA Replication, 2nd Ed. (Freeman, San Francisco, 1992). "Análogos" en referencia a los nucleósidos incluye nucleósidos sintéticos que tienen restos de bases modificadas y/o restos de azúcar modificado, por ejemplo, descritos por Scheit, Nucleotide Analogs (John Wiley, New York, 1980); Uhlman and Peyman, Chemical Reviews, 90: 543-584 (1990), o similares, con la condición de que sean capaces de hibridación específica. Dichos análogos incluyen nucleósidos sintéticos diseñados para mejorar las propiedades de unión, reducir la complejidad, aumentar la especificidad, y similares. Polinucleótidos que comprenden análogos con mejores propiedades de hibridación o de resistencia a la nucleasa están descritos en Uhlman and Peyman (citado anteriormente); Crooke *et al.*, Exp. Opin. Ther. Patents, 6: 855-870 (1996); Mesmaeker *et al.*, Current Opinion in Structural Biology, 5: 343-355 (1995); y similares. Los ejemplos de tipos de polinucleótidos que son capaces de mejorar la estabilidad del dúplex incluyen N3 '→P5' fosforamidatos de oligonucleótidos (denominados aquí "amidatos"), ácidos nucleicos peptídicos (denominados aquí "PNA"), oligo-2'-O-alquilrribonucleótidos, polinucleótidos que contienen propinilpirimidinas C-5, ácidos nucleicos bloqueados ("LNA"), y compuestos similares. Tales oligonucleótidos o bien están disponibles comercialmente o bien se pueden sintetizar utilizando métodos descritos en la literatura.

"Reacción en cadena de la polimerasa" o "PCR", significa una reacción para la amplificación *in vitro* de secuencias específicas de ADN mediante la extensión simultánea del cebador de cadenas complementarias de ADN. En otras palabras, la PCR es una reacción para fabricar múltiples copias o replicados de un ácido nucleico diana flanqueado por sitios de unión de cebadores, comprendiendo dicha reacción una o más repeticiones de las siguientes etapas: (i) desnaturalización del ácido nucleico diana, (ii) hibridación de los cebadores a los sitios de unión del cebador, y (iii) extensión de los cebadores por medio de una polimerasa de ácido nucleico en presencia de trifosfatos de nucleósidos. Normalmente, la reacción se hace en ciclos a través de diferentes temperaturas optimizadas para cada etapa en un instrumento termociclador. Las temperaturas concretas, las duraciones de cada etapa, y las velocidades de cambio entre etapas dependen de muchos factores bien conocidos por los expertos en la técnica, por ejemplo, ejemplificados por las referencias: McPherson *et al.*, editors, PCR: A practical Approach and PCR2: A practical Approach (IRL Press, Oxford, 1991 y 1995, respectivamente). Por ejemplo, en una PCR convencional que utiliza la ADN polimerasa Taq, se puede desnaturalizar un ácido nucleico diana de doble cadena a una temperatura >90 ºC, los cebadores se pueden hibridar a una temperatura en el intervalo de 50-75 ºC, y los cebadores se extienden a una temperatura en el intervalo 72-78 ºC. El término "PCR" engloba formas derivadas de la reacción, que incluyen pero no se limitan a, RT-PCR, PCR en tiempo real, PCR anidada, PCR cuantitativa, PCR multiplexada, y similares. Los volúmenes de reacción varían desde algunos nanolitros, por ejemplo, 2 nL, a unos centenares de µL, por ejemplo 200 µL. "PCR de transcripción inversa," o "RT-PCR", significa una PCR que va precedida por una reacción de transcripción inversa que convierte un ARN diana en un ADN complementario de cadena simple, que se amplifica entonces, por ejemplo, Tecott *et al.*, patente de Estados Unidos 5.168.038. "PCR en tiempo real" significa una PCR para la cual la cantidad del producto de reacción, es decir, el amplicón, se monitoriza conforme avanza la reacción. Hay muchas formas de PCR en tiempo real que difieren principalmente en la química de detección utilizada para la monitorización del producto de reacción, por ejemplo, Gelfand *et al.*, patente de Estados Unidos 5.210.015 ("TAQMAN™"); Wittwer *et al.*, patentes de Estados Unidos 6.174.670 y 6.569.627 (tintes intercalantes); Tyagi *et al.*, patente de Estados Unidos 5.925.517 (balizas moleculares). La química de detección para la PCR en tiempo real se revisa en Mackay *et al.*, Nucleic Acids Research, 30: 1292-1305 (2002). "PCR anidada" significa una PCR de dos fases en donde el amplicón de una primera PCR se convierte en la muestra para una segunda PCR que utiliza un nuevo conjunto de cebadores, uno de los cuales al menos se une en una localización interior del primer amplicón. Como se usa aquí, "cebadores iniciales" en referencia a una reacción de amplificación anidada significan los cebadores usados para generar un primer amplicón, y "cebadores secundarios" significan los uno o más cebadores utilizados para generar un segundo amplicón, o amplicón anidado. “PCR multiplexada" significa una PCR en la que múltiples secuencias diana (o una sola secuencia diana y una o más secuencias de referencia) se llevan a cabo simultáneamente en la misma mezcla de reacción, por ejemplo, Bernard *et al.*, Anal. Biochem., 273: 221-228 (1999) (PCR de dos colores en tiempo real). Usualmente, se emplean distintos conjuntos de cebadores para cada secuencia que se amplifica.

"Polinucleótido" u "oligonucleótido" se utilizan de modo intercambiable y cada uno significa un polímero lineal de monómeros nucleotídicos. Los monómeros que constituyen los polinucleótidos y oligonucleótidos son capaces de unirse específicamente a un polinucleótido natural, por medio de un patrón regular de interacciones monómero a monómero, tales como el apareamiento de bases de tipo Watson-Crick, apilamiento de bases, apareamiento de bases de tipos Hoogsteen o Hoogsteen inverso, apareamiento de bases oscilantes, o similares. Como se describe

en detalle a continuación, por "base oscilante" se entiende una base de ácido nucleico que puede formar un par de bases con una primera base de nucleótido en una cadena complementaria de ácido nucleico, pero que, cuando se emplea como una cadena molde para la síntesis de ácido nucleico, lleva a la incorporación de una segunda base de nucleótidos diferente a la cadena de síntesis. Dichos monómeros y sus uniones internucleosídicas pueden ser de origen natural o pueden ser análogos de los mismos, por ejemplo, análogos de origen natural o no natural. Los análogos que no son naturales pueden incluir ácidos nucleicos peptídicos (los PNA, por ejemplo, como se describen en la patente de Estados Unidos 5.539.082), ácidos nucleicos bloqueados (los LNA, por ejemplo, como se describe en la patente de Estados Unidos 6.670.461), uniones internucleosídicas de fosforotioato, bases que contienen grupos de unión que permiten la fijación de etiquetas, tales como fluoróforos, o haptenos y similares. Siempre que el uso de un oligonucleótido o polinucleótido requiera un procesamiento enzimático, tal como extensión por una polimerasa, ligamiento por una ligasa, o similares, un experto debe entender que los oligonucleótidos o polinucleótidos en estos casos no deben contener ciertos análogos de uniones internucleosídicas, restos de azúcar, o bases en todas o algunas posiciones. Los polinucleótidos normalmente varían de tamaño desde algunas unidades monoméricas, por ejemplo, 5-40, cuando se denominan usualmente "oligonucleótidos", hasta varios miles de unidades monoméricas. Siempre que un polinucleótido u oligonucleótido esté representado por una secuencia de letras (mayúsculas o minúsculas), tal como "ATGCCTG", se entenderá que los nucleótidos están en orden 5 '→ 3', de izquierda a derecha y que "A" indica desoxiadenosina, "C" indica desoxicitidina, "G" indica desoxiguanosina, y "T" indica timidina, "I" indica desoxiinosina, "U" indica uridina, salvo que se indique otra cosa o se deduzca del contexto. A menos que se indique otra cosa los convenios de la terminología y de la numeración de átomos seguirán los descritos en Strachan and Read, Human Molecular Genetics 2 (Wiley-Liss, New York, 1999). Normalmente, los polinucleótidos comprenden los cuatro nucleósidos naturales (por ejemplo, desoxiadenosina, desoxicitidina, desoxiguanosina, desoxitimidina para el ADN o sus análogos de ribosa para el ARN) unidos por enlaces de fosfodiéster; sin embargo, también pueden comprender análogos de nucleótidos no naturales, por ejemplo, que incluye bases modificadas, azúcares, o uniones internucleosídicas. Es claro para los expertos en la técnica que, cuando una enzima tiene requisitos específicos de sustrato de oligonucleótidos o polinucleótidos para su actividad, por ejemplo, ADN monocatenario, dúplex de ARN/ADN, o similares, entonces la selección de la composición apropiada para los sustratos de oligonucleótidos o polinucleótidos está efectivamente dentro de los conocimientos de un experto, especialmente con la guía de tratados tales como Sambrook *et al.*, Molecular Cloning, Second Edition (Cold Spring Harbor Laboratory, New York, 1989), y referencias similares.

"Cebador" significa un oligonucleótido, ya sea natural o sintético, que, al formar un dúplex con un molde de polinucleótido, es capaz de actuar como un punto de iniciación de la síntesis de ácido nucleico y de extenderse desde su extremo 3' a lo largo del molde de modo que se forma un dúplex extendido. La secuencia de nucleótidos añadida durante el proceso de extensión está determinada por la secuencia del polinucleótido de molde. Normalmente, los cebadores se extienden por una ADN polimerasa. Los cebadores tienen generalmente una longitud compatible con su uso en la síntesis de productos de extensión del cebador, y normalmente están en el intervalo de entre 8 a 100 nucleótidos de longitud, tal como 10 a 75, 15 a 60, 15 a 40, 18 a 30, 20 a 40, 21 a 50, 22 a 45, 25 a 40, y así sucesivamente, más típicamente en el intervalo de entre 18-40, 20-35, 21-30 nucleótidos de largo, y cualquier longitud entre los intervalos indicados. Los cebadores típicos pueden estar en el intervalo de entre 10-50 nucleótidos de longitud, tal como 15-45, 18-40, 20-30, 21-25 y así sucesivamente, y cualquier longitud entre los intervalos indicados. En algunas realizaciones, los cebadores normalmente no tienen más de aproximadamente 10, 12, 15, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 35, 40, 45, 50, 55, 60, 65 o 70 nucleótidos de longitud.

Los cebadores son normalmente de una sola cadena para una máxima eficiencia en la amplificación, pero pueden ser alternativamente de doble cadena. Si es de doble cadena, el cebador generalmente se trata en primer lugar para separar sus cadenas antes de ser utilizado para preparar productos de extensión. Esta etapa de desnaturalización se efectúa típicamente con calor, pero alternativamente se puede llevar a cabo utilizando un álcali, seguido por neutralización. Por lo tanto, un "cebador" es complementario de un molde, y se compleja mediante enlace de hidrógeno o hibridación con el molde para dar un complejo cebador/molde para la iniciación de la síntesis por una polimerasa, que se extiende por la adición de bases unidas covalentemente ligadas en su extremo 3' complementario al molde en el proceso de síntesis de ADN.

Un "par de cebadores" como se usa aquí, se refiere a los cebadores primero y segundo que tienen la secuencia de ácido nucleico adecuada para la amplificación a base de ácido nucleico de un ácido nucleico diana. Dichos pares de cebadores generalmente incluyen un primer cebador que tiene una secuencia que es la misma o similar a la de una primera porción de un ácido nucleico diana, y un segundo cebador que tiene una secuencia que es complementaria a una segunda porción de un ácido nucleico diana para proporcionar la amplificación del ácido nucleico diana o de un fragmento del mismo. La referencia a cebadores "primero" y "segundo" en este documento es arbitraria, a menos que se indique específicamente otra cosa. Por ejemplo, el primer cebador puede ser diseñado como un "cebador directo" (que inicia la síntesis de ácido nucleico a partir de un extremo 5' del ácido nucleico diana) o como un "cebador inverso" (que inicia la síntesis de ácido nucleico a partir de un extremo 5' del producto de extensión producido a partir de la síntesis iniciada a partir del cebador directo). Asimismo, el segundo cebador puede ser diseñado como un cebador directo o como un cebador inverso.

"Lectura" significa un parámetro o parámetros, que se miden y/o se detectan, que se pueden convertir en un número

o valor. En algunos contextos, la lectura se puede referir a una representación numérica real de dichos datos recogidos o registrados. Por ejemplo, una lectura de las señales de intensidad fluorescente procedentes de una micromatriz es la dirección y la intensidad de fluorescencia de una señal que se genera en cada sitio de hibridación de la micromatriz; por lo tanto, una lectura de este tipo puede ser registrada o almacenada de diversas formas, por ejemplo, como una imagen de la micromatriz, como una tabla de números, o similares.

"Sitio reflejo", "secuencia refleja" y equivalentes se utilizan para indicar una o más secuencias presentes en un polinucleótido que se emplean para mover un dominio intra-molecularmente desde su localización inicial a una localización diferente en el polinucleótido. El uso de secuencias reflejas se describe en detalle en la solicitud PCT número de serie PCT/IB2010/02243 titulada "Compositions and Methods for Intramolecular Nucleic Acid Rearrangement", publicada el 24 de febrero de 2011 como WO/2011/021102. En ciertas realizaciones, se elige una secuencia refleja de manera que sea distinta de otras secuencias del polinucleótido (es decir, con poca homología de secuencia con otras secuencias que puedan estar presentes en el polinucleótido, por ejemplo, secuencias genómicas o sub-genómicas a procesar). Como tal, una secuencia refleja se debe seleccionar de manera que no se hibride con ninguna secuencia excepto su complemento en las condiciones empleadas en los procedimientos reflejos. La secuencia refleja puede ser una secuencia sintética o generada artificialmente (por ejemplo, añadida a un polinucleótido en un dominio de adaptador) o una secuencia presente normalmente en un polinucleótido a ensayar (por ejemplo, una secuencia presente dentro de una región de interés en un polinucleótido a ensayar). En el sistema reflejo, está presente un complemento de la secuencia refleja (por ejemplo, insertado en un dominio de adaptador) en la misma cadena del polinucleótido que la secuencia refleja (por ejemplo, la misma cadena de un polinucleótido de doble cadena o en el mismo polinucleótido de cadena simple), donde el complemento se coloca en una localización particular a fin de facilitar un suceso de unión y polimerización intramolecular sobre dicha cadena particular. Las secuencias reflejas empleadas en el procedimiento reflejo descrito en este documento pueden tener por tanto un amplio intervalo de longitudes y secuencias. Las secuencias reflejas pueden variar de 5 a 200 bases nucleotídicas de longitud.

"Soporte sólido", "soporte ", y "soporte en fase sólida" se usan de modo intercambiable y se refieren a un material o grupo de materiales que tienen una superficie o superficies rígidas o semi-rígidas. En muchas realizaciones, al menos una superficie del soporte sólido será sustancialmente plana, aunque en algunas realizaciones puede ser deseable separar físicamente regiones de síntesis para diferentes compuestos, por ejemplo, con pocillos, regiones elevadas, pernos, zanjas grabadas, o similares. Según otras realizaciones, el soporte o soportes sólidos tendrán la forma de perlas, resinas, geles, microesferas, u otras configuraciones geométricas. Las micromatrices comprenden usualmente al menos un soporte en fase sólida planar, tal como un portaobjetos de microscopio de vidrio.

"Específica" o "especificidad", en referencia a la unión de una molécula con otra molécula, tal como una secuencia diana marcada para una sonda, significa el reconocimiento, contacto y formación de un complejo estable entre las dos moléculas, junto con sustancialmente menos reconocimiento, contacto o formación de complejos de esa molécula con otras moléculas. En un aspecto, "específica" en referencia a la unión de una primera molécula con una segunda molécula significa que en la medida en que la primera molécula reconoce y forma un complejo con otra molécula en una reacción o muestra, ella forma el mayor número de los complejos con la segunda molécula. Preferiblemente, este mayor número es al menos el cincuenta por ciento. Generalmente, las moléculas implicadas en un suceso de unión específica tienen zonas en sus superficies o en cavidades que dan lugar al reconocimiento específico entre las moléculas que se unen entre sí. Los ejemplos de unión específica incluyen interacciones anticuerpo-antígeno, interacciones enzima-sustrato, formación de dúplex o tríplex entre polinucleótidos y/o oligonucleótidos, interacciones de biotina-avidina o biotina-estreptavidina, interacciones receptor-ligando, y similares. Como se usa en la presente memoria, "contacto", en referencia a la especificidad o unión específica significa que dos moléculas están lo suficientemente próximas para que interacciones químicas no covalentes débiles, tales como las fuerzas de Van der Waals, enlaces de hidrógeno, interacciones en apilamiento de bases, interacciones iónicas e hidrófobas, y similares, dominen en la interacción de las moléculas.

Tal como se utiliza aquí, el término "T_m" se utiliza en referencia a la "temperatura de fusión". La temperatura de fusión es la temperatura (por ejemplo, medida en ºC) a la que una población de moléculas de ácido nucleico de doble cadena llega a ser semi-disociada en cadenas simples. Son conocidas en la técnica varias ecuaciones para calcular la Tm de ácidos nucleicos (véase, por ejemplo, Anderson and Young, Quantitative Filter Hybridization, in Nucleic Acid Hybridization (1985). Otras referencias (por ejemplo, Allawi, H.T. & SantaLucia, J., Jr., Biochemistry 36, 10581-94 (1997)) incluyen métodos alternativos de cálculo que tienen en cuenta, para el cálculo de T_m, las características estructurales y ambientales, así como las características de las secuencias.

"Muestra" significa una cantidad de material procedente de una fuente biológica, ambiental, médica, o de pacientes en la que se busca la detección, medida, o marcado de los ácidos nucleicos diana. Por un lado, se pretende incluir un espécimen o cultivo (por ejemplo, cultivos microbiológicos). Por otra parte, se pretende incluir tanto las muestras biológicas como las ambientales. Una muestra puede incluir un espécimen de origen sintético. Las muestras biológicas pueden ser animales, incluyendo de seres humanos, fluidos, sólidos (por ejemplo, heces) o tejidos, así como alimentos líquidos y sólidos y productos e ingredientes alimentarios tales como productos lácteos, verduras, carne y subproductos cárnicos y residuos. Las muestras biológicas pueden incluir materiales tomados de un

paciente, que incluyen pero no se limitan a los cultivos, sangre, saliva, líquido cefalorraquídeo, fluido pleural, leche, linfa, esputos, semen, aspirados por punción con aguja, y similares. Las muestras biológicas se pueden obtener de todas las diversas familias de animales domésticos, así como de animales silvestres o salvajes, que incluyen, pero no se limitan a animales tales como ungulados, osos, peces, roedores, etc. Las muestras ambientales incluyen material ambiental tal como muestras de materia superficial, suelo, agua y muestras industriales, así como las muestras obtenidas de los instrumentos, aparatos, equipo, utensilios, artículos desechables y no desechables, para el procesamiento de alimentos y productos lácteos. Estos ejemplos no se deben interpretar como limitantes de los tipos de muestra aplicables a la presente invención.

Los términos "aguas arriba" y "aguas abajo" en la descripción de la orientación y/o polimerización de la molécula de ácido nucleico se utilizan en el presente documento como son entendidos por los expertos en la técnica. Como tal, "aguas abajo" generalmente significa proceder en la dirección 5' a 3', es decir, la dirección en la que una polimerasa de nucleótido normalmente extiende una secuencia, y "aguas arriba" generalmente significa lo contrario. Por ejemplo, un primer cebador que se hibrida "aguas arriba" de un segundo cebador sobre la misma molécula de ácido nucleico diana está localizado en el lado 5' del segundo cebador (y por lo tanto la polimerización del ácido nucleico a partir del primer cebador se dirige hacia el segundo cebador).

Se observa, además, que las reivindicaciones se pueden redactar para excluir cualquier elemento opcional. Por lo tanto, esta declaración tiene por objeto servir como base antecedente para el uso de una terminología tan exclusiva como "únicamente", "solamente" y similares en relación con el detalle de los elementos reivindicados, o el uso de una limitación "negativa".

Antes de que se describa la presente invención, se debe entender que esta invención no se limita a las realizaciones particulares descritas, ya que estas pueden, por supuesto, variar. Se debe entender también que la terminología usada en este documento tiene el propósito de describir realizaciones particulares solamente, y no se pretende que sea limitante, ya que el alcance de la presente invención estará únicamente limitado por las reivindicaciones adjuntas.

Cuando se proporciona un intervalo de valores, se entiende que cada valor intermedio, hasta la décima de la unidad del límite inferior a menos que el contexto dicte claramente otra cosa, entre los límites superior e inferior de ese intervalo está también específicamente descrito. Cada intervalo más pequeño entre cualquier valor establecido o valor intermedio en un intervalo establecido y cualquier otro valor establecido o valor intermedio en dicho intervalo establecido está incluido dentro de la invención. Los límites superior e inferior de estos intervalos más pequeños pueden estar independientemente incluidos o excluidos del intervalo, y cada intervalo en el que cualquiera, ninguno o ambos límites están incluidos en los intervalos más pequeños está también incluido dentro de la invención, sujeto a cualquier límite específicamente excluido en el intervalo establecido. Cuando el intervalo establecido incluye uno o ambos límites, los intervalos que excluyen cualquiera o ambos de dichos límites incluidos, están incluidos también en la invención.

A menos que se defina otra cosa, todos los términos técnicos y científicos usados en este documento tienen el mismo significado que es entendido comúnmente por los expertos en la técnica a la que pertenece esta invención. Aunque en la práctica o ensayo de la presente invención se pueden utilizar cualquier método y materiales similares o equivalentes a los descritos aquí, se describen ahora algunos métodos y materiales potenciales y preferidos. Todas las publicaciones mencionadas en el presente documento se citan como referencia para dar a conocer y describir los métodos y/o los materiales en relación a los cuales se citan las publicaciones. Se entiende que la presente descripción reemplaza cualquier descripción de una publicación citada en la medida en que haya una contradicción.

Se debe observar que, como se usa aquí y en las reivindicaciones adjuntas, las formas singulares "un", "uno", “una” y "el", “la” incluyen los plurales referentes a menos que el contexto dicte claramente otra cosa. Así, por ejemplo, la referencia a "un ácido nucleico" incluye una pluralidad de dichos ácidos nucleicos y la referencia a "el compuesto" incluye la referencia a uno o más compuestos y equivalentes de los mismos conocidos por los expertos en la técnica, y así sucesivamente.

La práctica de la presente invención puede emplear, a menos que se indique otra cosa, técnicas y descripciones convencionales de la química orgánica, tecnología de polímeros, biología molecular (incluyendo técnicas recombinantes), biología celular, bioquímica e inmunología, que están dentro de la experiencia de la técnica. Tales técnicas convencionales incluyen la síntesis de matrices de polímeros, hibridación, ligamiento, y detección de la hibridación utilizando una etiqueta. Ilustraciones específicas de técnicas adecuadas se pueden haber tomado como referencia para el ejemplo que sigue. Sin embargo, también se pueden utilizar por supuesto, otros procedimientos convencionales equivalentes. Tales técnicas y descripciones convencionales se pueden encontrar en manuales de laboratorio estándar, tales como *Genome Analysis: A Laboratory Manual Series* (Vols I-IV.), *Using Antibodies: A Laboratory Manual, Cells: A Laboratory Manual, PCR Primer: A Laboratory Manual, and Molecular Cloning: A Laboratory Manual* (todos de Cold Spring Harbor Laboratory Press), Stryer, L. (1995) Biochemistry (4th Ed.) Freeman, New York, Gait, "Oligonucleotide Synthesis: A Practical Approach" 1984, IRL Press, London, Nelson and Cox (2000), Lehninger, A., Principles of Biochemistry 3rd Ed., W. H. Freeman Pub., New York, N.Y. y Berg *et al.*

(2002) Biochemistry, 5th Ed., W.H. Freeman Pub., New York, N. Y.

Las publicaciones expuestas en el presente documento se proporcionan únicamente para su descripción antes de la fecha de presentación de la presente solicitud. Nada en este documento se debe interpretar como una admisión de que la presente invención no tiene derecho a anteceder a dicha publicación en virtud de una invención anterior. Además, las fechas de publicación proporcionadas pueden ser diferentes de las fechas de publicación reales que puede ser necesario que sean confirmadas de forma independiente.

Como se ha resumido antes, algunos aspectos de la presente invención se dirigen al uso de bases de nucleótidos degeneradas (por ejemplo, en una región de bases degeneradas, o DBR) añadidas a los polinucleótidos sometidos a análisis de secuencias que se utilizan en el establecimiento del número de moléculas de polinucleótidos individuales procedentes de la misma región genómica de la misma muestra original que han sido secuenciadas en una configuración o proceso particular de análisis de secuencias. La inclusión de una DBR en los polinucleótidos sometidos a análisis de secuenciación se utiliza en una variedad de análisis genéticos, incluyendo el aumento de la confianza en la identificación de alelos al proporcionar un mecanismo para determinar un valor estadístico para una identificación de alelos, un valor que no se puede derivar del número de lecturas solo. La DBR se puede añadir a un polinucleótido de cualquier manera conveniente, incluyéndola como parte de un adaptador (o conjunto de adaptadores) unido a los polinucleótidos a secuenciar, por ejemplo, la DBR puede estar en un adaptador que incluye también un sitio del cebador de secuenciación, o la DBR puede estar presente en un cebador de síntesis de ácido nucleico, por ejemplo, un cebador de PCR, de tal manera que la DBR se añade a un polinucleótido diana cuando se utiliza el cebador en una reacción de polimerización.

Las DBR también encuentran utilidad en la realización de análisis genéticos en muestras agrupadas de polinucleótidos en las que cada polinucleótido de la muestra conjunta incluye un MID específico para su muestra de origen (descrito en detalle más adelante). Esto permite al usuario determinar la cobertura de secuencia de una especie específica de polinucleótidos (o múltiples especies) de cada una de las muestras de origen que se reunieron para generar la muestra conjunta. Por lo tanto, las realizaciones de la presente invención incluyen análisis de secuencia de polinucleótidos en una muestra conjunta, donde cada polinucleótido contiene un MID y una DBR.

Ácidos nucleicos

La presente invención (como se describe en detalle más adelante) se puede emplear para la manipulación y análisis de secuencias de ácido nucleico de interés (o polinucleótidos) procedentes de prácticamente cualquier fuente de ácido nucleico, que incluyen pero no se limitan a ADN genómico, ADN complementario (ADNc), ARN (por ejemplo, ARN mensajero, ARN ribosomial, ARN interferente corto, microARN, etc.), ADN plasmídico, ADN mitocondrial, ADN sintético, etc. Además, cualquier organismo, material orgánico o sustancia que contiene ácido nucleico se puede utilizar como una fuente de ácidos nucleicos para ser procesada de acuerdo con la presente invención, que incluyen, pero no se limitan a, plantas, animales (por ejemplo, reptiles, mamíferos, insectos, gusanos, peces, etc.), muestras de tejidos, bacterias, hongos (por ejemplo, levaduras), fagos, virus, tejido cadavérico, muestras arqueológicas/antiguas, etc. En ciertas realizaciones, los ácidos nucleicos en la muestra de ácido nucleico se derivan de un mamífero, y en determinadas realizaciones, el mamífero es un ser humano.

En ciertas realizaciones, las secuencias de ácidos nucleicos están enriquecidas. Por enriquecido se entiende que los ácidos nucleicos (por ejemplo, en una muestra de polinucleótidos) se someten a un proceso que reduce la complejidad de los ácidos nucleicos, generalmente aumentando la concentración relativa de una especie particular de ácido nucleico en la muestra (por ejemplo, que tiene un locus específico de interés, que incluye una secuencia específica de ácido nucleico, que carece de un locus o secuencia, que está dentro de un intervalo de tamaño específico, etc.). Hay una amplia variedad de formas de enriquecer los ácidos nucleicos que tienen una característica o características de secuencia específicas, y por tanto se puede emplear cualquier método conveniente para conseguir esto. El enriquecimiento (o reducción de la complejidad) puede tener lugar en cualquiera de una serie de etapas del proceso, y será determinado por deseo del usuario. Por ejemplo, el enriquecimiento puede tener lugar en muestras parentales individuales (por ejemplo, ácidos nucleicos sin marcar antes del ligamiento al adaptador) o en muestras multiplexadas (por ejemplo, ácidos nucleicos marcados con secuencias de adaptadores que codifican MID; los MID se describen con más detalle más adelante).

En ciertas realizaciones, los ácidos nucleicos de la muestra de ácido nucleico se amplifican antes del análisis. En algunas de estas realizaciones, la reacción de amplificación sirve también para enriquecer una muestra de ácido nucleico de partida en una secuencia o locus de interés. Por ejemplo, una muestra de ácido nucleico de partida se puede someter a una reacción en cadena de la polimerasa (PCR) que amplifica una o más regiones de interés. En ciertas realizaciones, la reacción de amplificación es una reacción de amplificación exponencial, mientras que en otras realizaciones particulares, la reacción de amplificación es una reacción de amplificación lineal. En la práctica de la presente invención, se puede utilizar cualquier método conveniente para llevar a cabo las reacciones de amplificación en una muestra de ácido nucleico de partida. En ciertas realizaciones, la polimerasa de ácido nucleico empleada en la reacción de amplificación es una polimerasa que tiene la capacidad de corrección de pruebas (por ejemplo, ADN polimerasa phi29, ADN polimerasa de Thermococcus litoralis, ADN polimerasa de Pyrococcus furiosus, etc.).

En ciertas realizaciones, la muestra de ácido nucleico a analizar se deriva de una única fuente (por ejemplo, un único organismo, virus, tejido, célula, sujeto, etc.), mientras que en otras realizaciones, la muestra de ácido nucleico es un conjunto de ácidos nucleicos extraídos de una pluralidad de fuentes (por ejemplo, un conjunto de ácidos nucleicos procedentes de una pluralidad de organismos, tejidos, células, sujetos, etc.), donde por "pluralidad" se entiende dos o más. Como tal, en ciertas realizaciones, una muestra de ácido nucleico puede contener ácidos nucleicos procedentes de 2 o más fuentes, 3 o más fuentes, 5 o más fuentes, 10 o más fuentes, 50 o más fuentes, 100 o más fuentes, 500 o más fuentes, 1000 o más fuentes, 5000 o más fuentes, 10.000 o más fuentes, 25.000 o más fuentes, etc.

En ciertas realizaciones, los fragmentos de ácido nucleico se van a reunir con fragmentos de ácido nucleico se derivan de una pluralidad de fuentes (por ejemplo, una pluralidad de organismos, tejidos, células, sujetos, etc.), donde por "pluralidad" se entiende dos o más. En tales realizaciones, los ácidos nucleicos derivados de cada fuente incluyen un identificador multiplex (MID) de tal modo que se puede determinar la fuente de la que se deriva cada fragmento de ácido nucleico marcado. En tales realizaciones, cada fuente de la muestra de ácido nucleico se correlaciona con un MID único, donde por MID único se entiende que cada MID diferente empleado se puede diferenciar de cualquier otro MID empleado en virtud de al menos una característica, por ejemplo, la secuencia de ácido nucleico del MID. Se puede utilizar cualquier tipo de MID, incluyendo pero sin limitarse a los descritos en la publicación de patente de EE.UU. 2007/0259357, también en tramitación, presentada el 22 de enero de 2007, y titulada "Nucleic Acid Analysis Using Sequence Tokens", así como en la patente de Estados Unidos 7,393,665, expedida el 1 de julio de 2008, y titulada "Methods and Compositions for Tagging and Identifying Polynucleotides" que describen etiquetas de ácidos nucleicos y su uso en la identificación de polinucleótidos. En ciertas realizaciones, un conjunto de MIDs empleados para marcar una pluralidad de muestras no necesita tener ninguna propiedad particular común (por ejemplo, T_m, longitud, composición de bases, etc.), ya que los métodos de marcado asimétrico (y muchos métodos de lectura de la marca, que incluyen pero no se limitan a la secuenciación de la marca o la medida de la longitud de la marca) pueden contener una amplia variedad de conjuntos de MID únicos.

Región de bases degeneradas (DBR)

Algunos aspectos de la presente invención incluyen métodos y composiciones para determinar o estimar el número de moléculas de polinucleótidos individuales procedentes de la misma región genómica de la misma muestra original que han sido secuenciadas en una configuración o proceso particular de análisis de secuencias. En estos aspectos de la invención, una región de bases degeneradas (DBR) se une a las moléculas de polinucleótidos de partida que se secuencian posteriormente (por ejemplo, después de que se realicen ciertas etapas del proceso, por ejemplo, la amplificación y/o el enriquecimiento, por ejemplo, la PCR). Como se detalla a continuación, la evaluación del número (y en algunos casos, la combinación) de diferentes secuencias de DBR presentes en una ronda de secuenciación permite el establecimiento del número (o número mínimo) de diferentes polinucleótidos de partida que han sido secuenciados para un polinucleótido particular (o región de interés; ROI). Este número se puede utilizar, por ejemplo, para dar una medida estadística de la confianza en las identificaciones de alelos, aumentando de este modo la confianza en la realización de tales determinaciones de alelos (por ejemplo, cuando se identifican alelos homocigóticos). Las DBR permiten también la identificación de errores potenciales de secuenciación o amplificación que impactan negativamente en el análisis genético si no son detectados.

La secuenciación del ADN incluye típicamente un etapa de unión de un adaptador a los polinucleótidos en una muestra a secuenciar, donde el adaptador contiene un sitio del cebador de la secuenciación (por ejemplo, mediante ligamiento). Como se usa en el presente documento, un "sitio del cebador de secuenciación" es una región de un polinucleótido que es o bien idéntica o bien complementaria a la secuencia de un cebador de secuenciación (cuando está en la forma de cadena simple) o una región de cadena doble formada entre una secuencia del cebador de secuenciación y su complemento. La orientación específica de un sitio del cebador de secuenciación puede ser deducida por los expertos en la técnica a partir de las características estructurales del polinucleótido específico que contiene el sitio del cebador de secuenciación.

En adición al sitio del cebador de secuenciación, también está unida a los polinucleótidos una región de bases degeneradas (DBR), ya sea como parte del adaptador que contiene el sitio del cebador de secuenciación o de forma independiente (por ejemplo, en un segundo adaptador unido al polinucleótido). Se puede emplear cualquier método conveniente para la unión o adición de una DBR a los polinucleótidos. Una DBR es una región que puede tener una composición o secuencia de bases variable (que se puede considerar como "aleatoria") en comparación con otros polinucleótidos marcados de la muestra. El número de DBR diferentes en una población de polinucleótidos de una muestra dependerá del número de bases en la DBR, así como del número potencial de diferentes bases que pueden estar presentes en cada posición. Por lo tanto, una población de polinucleótidos que tiene unidas DBR con dos posiciones de bases, donde cada posición puede ser una cualquiera de A, C, G y T, tendrá potencialmente 16 DBR diferentes (AA, AC, AG, etc.). Las DBR pueden incluir por lo tanto 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 o más bases, incluyendo 15 o más, 20 o más, etc. En ciertas realizaciones, la DBR tiene de 3 a 10 bases de longitud. Además, cada posición en una DBR puede tener una composición de bases diferente. Por ejemplo, una DBR de 4 bases puede tener cualquiera de las siguientes composiciones: NNNN; NRSN; SWSW; BDHV (véase la Tabla 1 a continuación para el código de nucleótidos según la IUPAC). Se observa además que en ciertas realizaciones, una base en una DBR

puede variar en virtud de tener una modificación detectable u otro resto unido a la misma. Por ejemplo, se pueden utilizar ciertas plataformas de secuenciación de nueva generación (por ejemplo, Pacific Biosciences™) para detectar diferencias de metilación en las bases durante el proceso de secuenciación. Como tal, una base no metilada en una DBR se podría distinguir de una base metilada en una DBR. Por lo tanto, no se pretende ninguna limitación con respecto a la longitud o composición de las bases de una DBR.

Código de nucleótidos según IUPAC	Base
A	Adenina
C	Citosina
G	Guanina
T (o U)	Timina (o uracilo)
R	A o G
Y	C o T
S	G o C
W	A o T
K	G o T
M	A o C
B	C o G o T
D	A o G o T
H	A o C o T
V	A o C o G
N	cualquier base

Se observa aquí que una DBR puede ser una única región (es decir, que tiene todas las bases de nucleótidos adyacentes entre sí) o puede estar presente en diferentes localizaciones en un polinucleótido (es decir, las bases de la DBR están separadas por secuencias no DBR, llamada también una DBR dividida). Por ejemplo, una DBR puede tener una o más bases en un primer adaptador en una primera localización en un polinucleótido y una o más bases en un segundo adaptador en una segunda localización en el mismo polinucleótido (por ejemplo, la DBR puede tener bases presentes en ambos extremos de un polinucleótido marcado asimétricamente, es decir, un polinucleótido que tiene adaptadores asimétricos). No se pretende ninguna limitación a este respecto.

Las DBR se pueden diseñar para facilitar la detección de los errores que se producen en las DBR durante los procesos de amplificación que se llevan a cabo antes del análisis de secuencia y/o los errores que se producen en la propia reacción de secuenciación. En tales realizaciones, las secuencias de DBR empleadas se diseñan de tal manera que un error en una secuencia de DBR no lleva necesariamente a la generación de otra posible secuencia de DBR (produciendo de ese modo la identificación incorrecta de replicones derivados del mismo molde como si fueran de un molde diferente debido a una mutación de la DBR). Considérese, por ejemplo, el uso de una DBR con la secuencia N. Un error en N convertiría una DBR en otra, lo que podría llevar a sobreestimar la probabilidad de que se estaba asignando correctamente un genotipo. Compárese esto con una DBR con secuencia Y. Si se ve una R en esta posición se sabe que ha habido un error. Aunque la DBR correcta no puede ser necesariamente asignada a esta DBR que contiene errores, se puede detectar que esto es debido a un error (por ejemplo, en la secuenciación o amplificación).

En algunas realizaciones, se pueden utilizar las secuencias de bases degeneradas como una combinación de MID-DBR que puede tanto (1) asignar la identidad de la muestra como (2) rastrear/contar las moléculas. Considérese, por ejemplo, dos muestras, una marcada con YYY y la otra marcada con RRR. En la reacción de secuenciación se observa una MID-DBR con la secuencia TAT, que no se ajusta a ninguna de las estructuras de la secuencia combinada de MID-DBR. Se requiere una mutación para convertir YYY en TAT. Se necesitan dos mutaciones para convertir RRR en TAT. Por lo tanto se podría decir que hay una mayor probabilidad de que la MID-DBR sea YYY en lugar de RRR.

La descripción de ejemplos de secuencias que identifican errores (o que corrigen errores) se puede encontrar en la literatura (por ejemplo, se describen en las publicaciones de la solicitud de patente de Estados Unidos 2010/0323348, titulada “Method and compositions for using error-detecting and/or error-correcting barcodes in nucleic acid amplification process”, y de la 2009/0105959, titulada “System and method for identification of individual samples from a multiplex mixture".

En ciertas realizaciones en las que está presente la DBR dentro de una población de adaptadores que incluye otros dominios funcionales (por ejemplo, sitio del cebador de secuenciación, MID, secuencia refleja), los dominios funcionales en la población de adaptadores serán idénticos entre sí mientras que la DBR variará. En otras palabras, a diferencia de los otros dominios en una población de adaptadores, la DBR tiene una composición de bases variable (o aleatoria). Por "población adaptadora", "población de adaptadores", y similares, se entiende una muestra de moléculas de adaptadores que están diseñadas para unirse a los polinucleótidos en una muestra.

La generación de adaptadores que tienen una DBR se puede conseguir de cualquier manera conveniente, por

ejemplo, utilizando métodos de síntesis de ADN bien conocidos en la técnica (véase las citas en la sección de definiciones anterior).

Una vez unidos a los polinucleótidos en la muestra parental, los polinucleótidos se pueden someter a un proceso adicional y, finalmente, se pueden secuenciar. Las etapas de proceso que se pueden realizar incluyen cualquier etapa de proceso que desee el usuario, por ejemplo, enriquecimiento, amplificación, y similares. En la etapa de secuenciación, se obtiene la secuencia de la DBR, así como la de una porción del polinucleótido (por ejemplo, que contiene una región de interés). Una vez que se obtienen las secuencias, se determina el número de diferentes DBR unidas a un polinucleótido de interés. Este número se puede emplear para determinar, o estimar, el número de diferentes polinucleótidos de interés procedentes de la muestra parental de partida, que se representan en los resultados de la secuenciación, donde en algunas realizaciones, el número determinado es el número mínimo de diferentes polinucleótidos de interés procedentes de la muestra parental de partida que se representan en los resultados de la secuenciación.

Considérese, por ejemplo, una DBR de dos bases que tiene una composición de bases NN (donde N es cualquier base de desoxinucleótido, es decir, A, G, C o T) empleada en la secuenciación de un locus para realizar una identificación de alelos para una muestra particular del sujeto (es decir, si un sujeto es homocigótico o heterocigótico en el locus). Aunque puede haber algunos sesgos desde la síntesis de oligonucleótidos, se puede esperar que haya 16 DBR diferentes en la población de adaptadores con aproximadamente la misma probabilidad (como se ha descrito antes). Cuando se identifica la posibilidad de identificación de un alelo homocigótico, la determinación del número de las DBR presentes en el ciclo de secuenciación se puede utilizar para determinar/estimar el número (o número mínimo) de moléculas de polinucleótidos que fueron secuenciadas en realidad (y por tanto el número de las que fueron amplificadas durante la etapas de proceso).

Para genomas diploides, la identificación de alelos (en el caso ideal o teórico) puede ser modelada por la distribución binomial. Dado que las dos copias de alelos (X e Y) difieren en algún sitio, la probabilidad de observar todos X o todos Y viene dada por la fórmula $(½)^c$, donde c es el número de observaciones (lecturas) del sitio. Si se observa X diez veces en un sitio (y no se observa Y), se puede decir que probablemente la muestra es homocigótica para el tipo X. Por lo tanto, la probabilidad de un error en esta determinación es $(½)^{10}$ (algo menos de uno en mil).

Los experimentos de esta invención demuestran que bajas cantidades de ADN en la fase inicial de preparación de la muestra pueden dar lugar a una alta cobertura de lecturas que corresponden todas a un alelo, y que esto puede ocurrir muchas más veces de lo que se debería esperar de acuerdo con la distribución binomial. Esto es debido a la amplificación de algunas moléculas de ADN (o incluso de una sola molécula) que da como resultado un gran número de lecturas derivadas de un locus genético en un único cromosoma (es decir, sólo uno de los dos cromosomas diploides realmente presentes en la muestra de interés). El resultado de esto es que el error como una función de la cobertura se desvía totalmente del error binomial pronosticado.

El uso de una DBR como se describe aquí aumentará la confianza en la realización de identificaciones de alelos procedentes de muestras que tienen cantidades limitadas de ADN. Por ejemplo, si 16 lecturas de secuenciación de un alelo de un locus genético contienen todas la secuencia GA de DBR, entonces es probable que todas estas lecturas procedan de la misma molécula de polinucleótido parental (y por lo tanto no está justificada una identificación de un alelo homocigótico). Sin embargo, si las 16 lecturas de secuenciación tienen cada una, una secuencia de DBR diferente, se puede realizar una identificación homocigótica con más confianza, ya que cada lectura proviene de una molécula de polinucleótido parental diferente.

Se observa aquí que en muchas realizaciones, no es posible llegar a la conclusión de que los polinucleótidos que tienen secuencias de DBR idénticas se derivan de la misma molécula de polinucleótido parental, ya que múltiples DBR idénticas pueden estar presentes en los polinucleótidos unidos a la DBR. Por ejemplo, si una población de adaptadores que contiene una DBR de dos bases N se utiliza para marcar una muestra que contiene más de 16 polinucleótidos, un subconjunto de los polinucleótidos marcados tendrá idénticas DBR, y por lo tanto no será posible determinar que sus secuencias fueron derivadas de diferentes moléculas del polinucleótido parental.

Una manera a modo de ejemplo para determinar más exactamente el número real de moléculas de partida o moléculas parentales debería ser aumentar la degeneración de las DBR (es decir, aumentar el número de secuencias únicas en la DBR utilizadas para marcar la muestra particular de interés) de modo que es probable que cada molécula única tenga una DBR diferente. En cualquier caso, en métodos a modo de ejemplo, se puede utilizar el número de DBR observadas o también la distribución de probabilidad del número esperado de lecturas para producir muy probablemente el número observado de las DBR.

Cuando se calculan estimaciones de si una identificación de alelo particular es un heterocigoto o un homocigoto, se puede crear/emplear una función apropiada $L(r,v)$ que restituye la probabilidad de un genotipo con las lecturas de referencia r y variante v. Cuando se emplean las DBR como se describe aquí, se puede utilizar una función modificada para el cálculo de los estimados $L(r',v')$, donde r' es el número de DBR únicas para la lectura de referencia y v' es el número de DBR únicas para la lectura variante. Se puede emplear cualquier función conveniente

para realizar identificaciones de alelos y se puede modificar para emplear los datos relativos a las lecturas de DBR como se describe aquí.

Se observa aquí, que se pueden utilizar aspectos de la invención para aumentar la confianza en la identificación de las variaciones del número de copias en una muestra de polinucleótidos, por ejemplo, una muestra genómica. Las variaciones del número de copias pueden incluir reordenamientos genómicos tales como deleciones, duplicaciones, inversiones y translocaciones o aneuploidías cromosómicas enteras tales como monosomías, disomías, trisomías, tetrasomías y pentasomías. Considérese, por ejemplo, un suceso de duplicación donde los padres tienen genotipos CA y CC en un SNP dado, y el probando tiene genotipo ACC. En el padre con el genotipo AC, con suficiente profundidad de cobertura de secuenciación (es decir, suficientes lecturas de secuenciación), se espera que el número de las DBR asociadas con el alelo C y con el alelo A sea similar. En el probando, con suficiente profundidad de cobertura de secuenciación, se espera que el número de las DBR asociadas con el alelo C sea 2 veces el número de las DBR asociadas con el alelo A, lo que proporciona la prueba de un caso de duplicación que engloba el alelo C. El uso de las DBR, más que el número de lecturas de secuenciación, proporciona más confianza en la identificación de una variación del número de copias ya que la DBR se puede utilizar para identificar lecturas que se derivan de diferentes moléculas de polinucleótidos.

Ejemplos de aplicaciones de las DBR

Como se ha detallado antes, las DBR permiten la validación estadística de las variantes de secuencia en una muestra heterogénea, incluyendo genomas o conjuntos de genomas complejos. Por ejemplo, las DBR encuentran uso en el análisis de genomas complejos en muestras de tumores, muestras microbianas, muestras ambientales, etc.

A continuación se proporcionan ejemplos de métodos estadísticos y ejemplos de aplicaciones de las DBR. Las descripciones que siguen se indican sólo a modo de ejemplos y no pretenden limitar el alcance del empleo de las DBR en los análisis de polinucleótidos.

Métodos Estadísticos

Como se ha descrito antes, en aspectos de la presente invención, las rondas de bases degeneradas (las DBR) se utilizan para estimar, o para obtener una medida cuantitativa del número real de moléculas molde secuenciadas o analizadas en un proceso dado. Dos lecturas pueden tener la misma DBR ya sea porque las lecturas proceden de la misma molécula molde o porque las moléculas han recibido la misma DBR por casualidad. El número potencial de distintas moléculas molde secuenciadas varía del número de las DBR al número de lecturas. La distribución de las DBR desde un número de moléculas de partida viene dada por la distribución de ocupación [véase C.A. Charalambides and C.A. Charalambides. Combinatorial methods in discrete distributions. John Wiley and Sons, 2005]. Dado un número observado de DBR, el número probable de moléculas de partida se puede calcular utilizando la estimación de máxima probabilidad, u otras técnicas adecuadas. Alternativamente, para cada DBR, la molécula molde más probable se puede estimar utilizando la secuencia de consenso de todas las lecturas con esa particular DBR. Los métodos se pueden combinar para generar estimados exactos del número de moléculas molde asociadas con variantes particulares.

Las DBR en la amplificación por PCR

Las DBR se pueden utilizar para estimar o para obtener una medida del número de moléculas de partida utilizadas como moldes para una reacción de PCR. Por ejemplo, una muestra de polinucleótido de partida puede ser amplificada por PCR en el primer ciclo, o los primeros ciclos, utilizando un par de cebadores de la PCR en los que uno de los cebadores (o ambos) incluyen una secuencia de cebador genérico y una DBR 5' hasta la secuencia diana específica. Después del ciclo o ciclos iniciales, este par de cebadores de PCR que contienen la DBR puede ser eliminado o inactivado y reemplazado con cebadores de PCR que no tienen una DBR para los ciclos restantes. La eliminación/inactivación de los cebadores que contienen DBR se puede llevar a cabo de cualquier manera conveniente, por ejemplo, por medios físicos o bioquímicos. Por ejemplo, los cebadores que contienen DBR pueden tener unido a los mismos un primer miembro de un par de unión (por ejemplo, biotina), facilitando de este modo la eliminación de estos cebadores al poner en contacto la muestra con el compañero de unión unido a un soporte sólido (por ejemplo, estreptavidina unida a un soporte sólido) y recogiendo la fracción no unida. Alternativamente, los cebadores libres que contienen DBR se pueden eliminar tratando la muestra con una exonucleasa específica de cadena simple (por ejemplo, exonucleasa I), haciendo a los cebadores incapaces de participar en otras etapas de extensión del cebador (por ejemplo, mediante la incorporación de un didesoxinucleótido en el extremo 3'), o por un proceso de inmovilización reversible de fase sólida (SPR1) (por ejemplo, Agencourt AMPure XP-PCR Purification, Beckman Coulter). Los segundos cebadores de PCR se diseñan para comprender las secuencias presentes en el extremo 5' de cada uno del primer conjunto de cebadores de manera que se replique la DBR en moldes generados a partir de los cebadores de PCR que contienen DBR utilizados en el primero o primeros ciclos. Por lo tanto, los ciclos restantes de la PCR amplificarán sólo los productos del primero o primeros ciclos que contienen las DBR. En otra realización, los cebadores que contienen la DBR pueden ser diseñados para tener una T_m más alta que el segundo conjunto de cebadores que no contienen la DBR (es decir, la T_m de la secuencia específica diana de los primeros cebadores de la PCR es más alta que la de los segundos cebadores de la PCR específica para las secuencias del cebador genérico en 5'). En este escenario a modo de ejemplo, los cebadores que contienen DBR pueden estar

presentes en cantidades limitantes y el primero o primeros ciclos de la PCR se llevan a cabo a la T_m más alta de tal modo que sólo los cebadores que contienen DBR se hibridan y participan en la síntesis del ácido nucleico. Debido a que los cebadores que contienen DBR están presentes en cantidades limitantes, se utilizarán sólo en el primero o primeros ciclos de la PCR. La realización de los ciclos de PCR restantes a una T_m más baja permitirá una amplificación adicional mediante el segundo conjunto de cebadores de la PCR que no incluyen las DBR pero que replicarán las DBR procedentes de los productos del primero o primeros ciclos (como se ha descrito antes).

Se observa que hay muchas combinaciones diferentes de cebadores de la PCR y condiciones de amplificación que se pueden emplear para llevar a cabo la amplificación por PCR con las DBR descrita anteriormente. Por ejemplo, tales reacciones pueden incluir 3 cebadores, en donde el cebador 1 (cebador directo específico para el polinucleótido diana y que contiene una DBR y una secuencia de cebado genérico en 5') y el cebador 2 (cebador específico inverso para el polinucleótido diana y sin ninguna DBR) se utilizan para amplificar la diana en el primero o primeros ciclos, y el cebador 3 (cebador directo específico para la secuencia de cebado genérico en 5' del cebador 1) y el cebador 2 se utilizan para los ciclos restantes.

Se observa, además, que el marcado de los dos extremos de un producto de la PCR con una DBR (es decir, donde ambos cebadores utilizados en el primero/primeros ciclos incluyen una DBR) puede proporcionar una mayor confianza en la estimación del número de polinucleótidos de partida amplificados. Se observa que si se utilizan más de 2 ciclos de PCR para unir las DBR, entonces es necesario tomar precauciones adicionales durante el análisis de los datos cuando se utilizan las DBR para rastrear la molécula molde inicial (o de partida) a partir de la cual se amplificaron los productos. Esto es debido a la posibilidad de que en el tercer ciclo de la PCR, un cebador de PCR que tiene una DBR se puede unir a un sitio de DBR existente en un producto de la PCR generado previamente, introduciendo de este modo una nueva secuencia de DBR. Como se indica a continuación, el análisis teórico de los tres primeros ciclos de PCR demuestra que es posible rastrear el linaje de una molécula. Se observa que el análisis que sigue podría ser utilizado teóricamente para cualquier número de ciclos de PCR para la adición de las secuencias de DBR, aunque la profundidad de la secuenciación tendría que ser suficiente.

El método descrito a continuación permite lecturas de una a un grupo de secuencias después de >1 ciclo de adición de DBR utilizando cebadores de PCR que contienen DBR. La Tabla I muestra cada uno de los productos de la PCR generados en cada uno de tres ciclos de PCR a partir de un único molde de doble cadena (el molde que tiene en la parte superior la cadena A y en la parte inferior la cadena B, como se señala en el ciclo 0 de la Figura 3). En la Tabla I, se muestra cada cadena presente en cada ciclo (señalada con las letras A a P) junto con su respectiva cadena molde (es decir, la cadena que sirvió como molde durante la síntesis de la cadena indicada), y la DBR 5' y la DBR 3' presentes en la cadena, si hay alguna (indicadas con los números 1 a 14). Por "DBR 5'" se entiende una secuencia de DBR que fue incorporada como parte de un cebador de PCR. Por "DBR 3'" se entiende una secuencia complementaria de una secuencia DBR 5' (es decir, generada como resultado de la extensión del cebador a lo largo de una secuencia existente de DBR 5'). En el ciclo 3, se puede ver que se puede producir una sobreescritura de DBR (indicada en la columna más a la derecha; véase, por ejemplo, las cadenas K y N producidas en el ciclo 3).

Tabla I. Marcado de DBR en los ciclos 0 a 3 de una reacción de PCR para un único molde de doble cadena (A y B). Ciclo	Cadena	DBR 5' nº	DBR 3' nº	Cadena molde	Sobreescritura de DBR
Ciclo 0	A	-	-	-	
	B	-	-	-	
Ciclo 1	A	-	-	-	
	B	-	-	-	
	C	1	-	A	
	D	2	-	B	
Ciclo 2	A	-	-	-	
	B	-	-	-	
	C	1	-	A	
	D	2	-	B	
	E	3	-	A	
	F	4	1	C	

	G	5	2	D	
	H	6	-	B	
Ciclo 3	A	-	-	-	
	B	-	-	-	
	C	1	-	A	
	D	2	-	B	
	E	3	-	A	
	F	4	1	C	
	G	5	2	D	
	H	6	-	B	
	I	7	-	A	
	J	8	3	E	
	K	9	4	F	sobreescritura de DBR 1 con DBR 9
	L	10	1	C	
	M	11	2	D	
	N	12	5	G	sobreescritura de DBR 2 con DBR 12
	O	13	6	H	
	P	14	-	B	

La Tabla I anterior y la Figura 3 muestran las cadenas que se han acumulado durante todo el proceso de la PCR. (Debe observarse el arrastre de las cadenas A y B del ciclo 0 a los ciclos 1, 2, y 3; el arrastre de las cadenas C y D del ciclo 1 a los ciclos 2 y 3; etc.).

Dada la suficiente profundidad de secuenciación, las DBR se pueden utilizar para rastrear la molécula de origen, incluso si se ha producido una sobreescritura de la DBR. Por ejemplo, la cadena K tiene DBR 5' nº 9 y DBR 3' nº 4. La DBR nº 4 se comparte con la cadena F, que tiene DBR 5' nº 4 y DBR 3' nº 1. La DBR nº 1 se comparte con la cadena C. Por lo tanto las cadenas K y F se derivan originalmente de la cadena C. Del mismo modo, la cadena N tiene DBR 5' nº 12 y DBR 3' nº 5. La DBR nº 5 se comparte con la cadena G, que tiene DBR 5' nº 5 y DBR 3' nº 2. La DBR nº 2 se comparte con la cadena D. Por lo tanto las cadenas N y G se derivan originalmente de la cadena D.

Como se ha expuesto anteriormente, los cebadores de la PCR que contienen DBR se eliminan después de los primeros ciclos (por ejemplo, después de la finalización del ciclo 3 como se muestra en la Tabla 1).

La Figura 3 muestra un esquema para los 2 primeros ciclos de PCR para un único molde de cadena doble como se muestra en la Tabla I. En el ciclo 0, sólo las cadenas molde de doble cadena están presentes, es decir, la cadena A de la parte superior y la cadena B de la parte inferior. Obsérvese que la dirección de las flechas sobre cada cadena en la Figura 3 indica la dirección 5' a 3'. En el primer ciclo de la PCR (ciclo 1), se producen 2 productos mediante el par de cebadores de la PCR (ambos miembros del par de cebadores incluyen una secuencia de DBR), la primera cadena (C) que tiene DBR nº 1 ("1" como se muestra en la figura) y la segunda (D) que tiene DBR2 ("2" como se muestra en la figura). En el segundo ciclo de la PCR (ciclo 2), los cuatro moldes (A, B, C y D) producen 4 productos (E, F, G y H), que tiene cada uno, una subsiguiente DBR unida (DBR nº 3, nº 4, nº 5 y nº 6, respectivamente). Se debe observar que los productos generados a partir de los moldes C y D (F y G) tienen ahora las DBR en ambos extremos. El ciclo 3 (que no se muestra en la Figura 3) utiliza entonces los 8 moldes del ciclo 2 para producir 8 productos, que tiene cada uno DBR adicionales unidas (véase los productos que se muestran en la Tabla I). El ciclo 3 es el primer ciclo en el que puede ocurrir una sobreescritura de la DBR (es decir, el cebado y la extensión de los moldes F y G con los cebadores de PCR marcados con subsiguientes DBR sobrescribirán la DBR nº 1 y la DBR nº 2; estas se muestran en la Tabla I como cadenas K y N).

En el análisis de las DBR de polinucleótidos en los que es posible la sobreescritura de las DBR, las lecturas se agrupan de acuerdo con las secuencias 5' y 3' de DBR y se rastrea el linaje de la molécula parental.

Como se desprende de la descripción anterior, la DBR es útil (1) para la identificación de los errores de la PCR que se presentan durante los primeros ciclos y (2) para la exactitud de la determinación de la identificación/número de copias de un alelo. Para la identificación de errores, es claramente permisible agrupar sucesos de cebado independientes. La exactitud de los cálculos de la identificación de alelos aumentó ligeramente en complejidad dado que los sucesos de cebado no representan necesariamente moléculas de partida independientes. Sin embargo, es una suposición razonable que los sucesos de cebado son igualmente probables en cualquier alelo, y por lo tanto este análisis es útil para mejorar la exactitud de la identificación de alelos.

Para copias muy bajas del molde inicial, el uso de múltiples ciclos de adición de DBR puede ser ventajoso. Por ejemplo, a concentraciones muy bajas de ADN no se podría recuperar un número suficiente de las DBR para dar un

genotipo exacto utilizando métodos estándar. El permitir múltiples casos de cebado en la misma molécula molde puede dar, en este caso, suficiente confianza para realizar una identificación de alelos al proporcionar más datos.

Se puede utilizar el análisis de las DBR en los productos de amplificación final para estimar el número de moléculas de partida amplificadas en la reacción. Dicho análisis permitirá al usuario determinar si los productos de la reacción de la PCR representan la amplificación selectiva de sólo algunos (o incluso uno) polinucleótidos de partida y/o ayudar en la determinación de los errores de la PCR que se han producido durante la amplificación (por ejemplo, como se ha descrito antes).

Las DBR en muestras de tumores heterogéneos

Las DBR también se utilizan en la evaluación de la heterogeneidad de anomalías cromosómicas en muestras de tumores, por ejemplo, dentro de un único tumor o entre tumores diferentes en un sujeto. Por ejemplo, se pueden obtener una o más muestras tumorales a partir de un único tumor (por ejemplo, en diferentes localizaciones dentro o alrededor del tumor) y/o de diferentes tumores en un sujeto y se pueden analizar en cuanto a la variación genética en una o más localizaciones cromosómicas. En ciertas realizaciones, se pueden obtener muestras de un tumor (o sujeto) a lo largo del tiempo. Tales variaciones pueden incluir cambios específicos de bases, deleciones, inserciones, inversiones, duplicaciones, etc., que son conocidos en la técnica. Las DBR se pueden emplear para marcar los polinucleótidos en la muestra o muestras tumorales antes de identificar las variaciones genéticas específicas, proporcionando así una manera de realizar los análisis estadísticos para validar todas las variantes identificadas. Por ejemplo, se puede utilizar el análisis estadístico para determinar si una variación detectada representa una mutación en un subconjunto de células del tumor, si es una variación que es específica para un tumor particular del sujeto, si es una variación que se encuentra en células no tumorales del individuo, o si es un artefacto del procedimiento por el cual se identificó la variante (por ejemplo, un artefacto de PCR).

Las DBR en la evaluación de la diversidad microbiana

El análisis de las DBR también se puede utilizar en la determinación de la variación/diversidad genética de una población de microbios/virus en una única muestra o entre muestras diferentes (por ejemplo, muestras recogidas a diferentes puntos de tiempo o de diferentes localizaciones). Por ejemplo, las muestras recogidas de un individuo a lo largo de una infección pueden ser analizadas en cuanto a la variación genética durante el proceso de infección utilizando las DBR como se describe aquí. Sin embargo, no se pretende ninguna limitación en el tipo de muestra microbiana/viral, y como tal, la muestra puede ser de cualquier fuente, por ejemplo, de un sujeto con una infección, de una fuente ambiental (suelo, agua, plantas, animales o productos de desechos animales, etc.), de una fuente de alimentos, o cualquier otra muestra para la que se desea la determinación de la diversidad genética de los microorganismos de la muestra, en uno o más locus o regiones genéticas. En la práctica de los métodos, los polinucleótidos derivados de la muestra se marcan con las DBR como se describe en la presente memoria (ya sea antes o después de una etapa de enriquecimiento) y se procesan para identificar variaciones genéticas en uno o más sitios genéticos o locus de interés. El análisis de las DBR se puede llevar a cabo entonces para proporcionar un aumento de la confianza en la determinación de la diversidad genética de los microbios de la muestra en el locus de interés. Dicho análisis se puede realizar sobre muestras recogidas de diversas fuentes y/o en diversos puntos de tiempo desde una fuente. Los ejemplos de locus genéticos que se pueden utilizar en la evaluación de la diversidad microbiana incluyen, pero no se limitan a, ARN ribosómico, por ejemplo, ARN ribosómico 16S, genes de resistencia a antibióticos, genes de enzimas metabólicas, etc.

Las DBR en la evaluación de los niveles de diferentes especies de polinucleótidos en una muestra

El análisis de las DBR se puede utilizar también en la evaluación de los niveles de diferentes especies de polinucleótidos en una muestra. En concreto, debido a que el análisis de las DBR puede determinar (o estimar) el número de polinucleótidos parentales en una muestra, se pueden evaluar la cantidad relativa o cuantitativa de las especies de polinucleótidos específicos y la confianza en la determinación de dichas especies. Por ejemplo, el análisis de una muestra de ADNc utilizando las DBR se puede emplear para evaluar los niveles relativos o cuantitativos de diferentes especies de ADNc en la muestra, proporcionando así una manera de determinar sus niveles relativos de expresión génica.

Las DBR en el análisis de muestras conjuntas

Otra aplicación de las DBR es en la realización de análisis genéticos sobre muestras reunidas de polinucleótidos en las que cada polinucleótido de la muestra conjunta incluye un MID específico para su muestra de origen (descrito en detalle anteriormente). Esto permite al usuario determinar la cobertura de secuencia de una especie de polinucleótidos específicos (o múltiples especies) de cada una de las muestras de origen que se reúnen para generar la muestra conjunta. Esto proporciona un mecanismo para asegurar que los polinucleótidos de cada muestra de partida en la muestra conjunta están representados adecuadamente. Por lo tanto, las realizaciones de la presente invención incluyen análisis de secuencias de polinucleótidos en una muestra conjunta, donde cada polinucleótido contiene un MID y una DBR. Se observa que en estas realizaciones, se puede usar el mismo diseño de DBR en conjunción con todas las muestras parentales/MIDs, ya que esta es la combinación de MID/DBR que se utiliza en el análisis de secuencias de la muestra específica.

Los análisis de la muestra conjunta utilizando los MID y DBR se utilizan en numerosos análisis genéticos, incluyendo

la realización de identificaciones de alelos, la corrección de errores de secuencias, los análisis relativos y cuantitativos de la expresión génica, y similares. Se observa que al analizar polinucleótidos en una muestra conjunta de acuerdo con aspectos de la presente invención, es importante mantener tanto los dominios de MID como los dominios de DBR en cada etapa del flujo de trabajo que se emplea, ya que la pérdida de uno u otro dominio tendrá un impacto negativo sobre la confianza en los resultados obtenidos.

Se observa, además, que el uso de los dominios de MID y DBR en el análisis genético es especialmente potente cuando se combina con plataformas de secuenciación de nueva generación (NGS), muchas de las cuales proporcionan datos de secuencias para cada polinucleótido individual presente en la muestra a ser secuenciada. En contraste con los métodos convencionales de secuenciación en los que los clones individuales de los polinucleótidos se secuencian de forma independiente, las plataformas de secuenciación de nueva generación proporcionan secuencias de múltiples polinucleótidos diferentes en una muestra de forma simultánea. Esta diferencia permite que se hagan análisis estadísticos específicos de la muestra que no están limitados por tener que clonar y secuenciar de forma independiente cada polinucleótido. Por lo tanto, los análisis del dominio MID/DBR descritos en la presente memoria son sinérgicos con las plataformas de secuenciación de nueva generación, proporcionando mejores métodos estadísticos para analizar las cantidades muy grandes de datos de las secuencias de las muestras conjuntas.

Kits y sistemas

La descripción proporciona también kits y sistemas para la práctica de los métodos propuestos, es decir, para el uso de las DBR para determinar el número (o mínimo número), de diferentes polinucleótidos de partida que han sido secuenciados para un polinucleótido particular. Como tales, los sistemas y kits pueden incluir polinucleótidos que contienen las DBR (por ejemplo, adaptadores), así como cualquier otro dominio funcional de interés como se describe en el presente documento (por ejemplo, sitios de los cebadores de secuenciación, MID, secuencias reflejas, etc.). Los sistemas y kits pueden incluir también reactivos para llevar a cabo todas las etapas para unir los adaptadores a los polinucleótidos de una muestra parental, preparar una muestra parental para la unión adaptador/DBR, y/o reactivos para llevar a cabo las reacciones de secuenciación (por ejemplo, ligasas, enzimas de restricción, nucleótidos, polimerasas, cebadores, cebadores de secuenciación, dNTPs, ddNTPs, exonucleasas, etc.). Los diversos componentes de los sistemas y kits pueden estar presentes en recipientes separados o ciertos componentes compatibles pueden ser pre-reunidos en un único recipiente, según se desee.

Los sistemas y kits propuestos pueden incluir también uno o más de otros reactivos para preparar o procesar una muestra de ácido nucleico de acuerdo con los métodos propuestos. Los reactivos pueden incluir una o más matrices, disolventes, reactivos de preparación de muestras, tampones, reactivos de desalación, reactivos enzimáticos, reactivos desnaturalizantes, donde los estándares de calibración, tales como los controles positivos y negativos se pueden proporcionar también. Así pues, los kits pueden incluir uno o más recipientes tales como viales o frascos, conteniendo cada recipiente un componente separado para llevar a cabo una etapa de procesamiento o preparación de una muestra según la presente invención.

Además de los componentes mencionados antes, los kits propuestos incluyen típicamente además instrucciones para utilizar los componentes del kit para practicar los métodos propuestos, por ejemplo, para emplear las DBR como se ha descrito anteriormente. Las instrucciones para la práctica de los métodos de la invención generalmente se registran en un medio de registro adecuado. Por ejemplo, las instrucciones se pueden imprimir sobre un sustrato, tal como papel o plástico, etc. Como tales, las instrucciones pueden estar presentes en los kits como un prospecto, en el etiquetado del recipiente del kit o de sus componentes (es decir, asociado con el empaquetado o sub-empaquetado), etc. En otros kits, las instrucciones están presentes como un archivo electrónico de almacenamiento de datos presente en un medio de almacenamiento legible por un ordenador adecuado, por ejemplo, CD-ROM, disquete, etc. Todavía en otras realizaciones, las instrucciones reales no están presentes en el kit, pero se proporcionan medios para obtener las instrucciones de una fuente remota, por ejemplo, a través de Internet. Un ejemplo de esta realización es un kit que incluye una dirección web en la que se pueden ver las instrucciones y/o de la que se pueden descargar las instrucciones. Al igual que con las instrucciones, este medio para obtener las instrucciones se registra en un sustrato adecuado.

Además de la base de datos, programación e instrucciones propuestas, los kits pueden incluir también una o más muestras control y reactivos, por ejemplo, dos o más muestras control para utilizar en el ensayo del kit.

EJEMPLOS

Métodos

Dos muestras idénticas de ADN genómico de ratón se marcaron con adaptadores. Una de las muestras utilizó un adaptador que tiene una región sintetizada de forma redundante que consiste en 7 bases (RYBDHVB), cada una de las cuales podría ser una de dos bases (para las posiciones R e Y) o de tres bases (para las posiciones B, D, H o V) (o un total de 972 secuencias diferentes), seguida por las bases AC**A**; la segunda muestra utilizó un adaptador que tiene una región sintetizada de forma redundante que consiste en 7 bases (RYBDHVB), cada una de las cuales podría ser una de dos bases (para las posiciones R e Y) o de tres bases (para las posiciones B, D, H o V) (o un total

de 972 secuencias diferentes) seguida por las bases AC**G**. Nótese que las bases en negrita y subrayadas corresponden a un sitio polimórfico sintético. En estos adaptadores, la secuencia RYB sirve como la región DBR y la DHVB sirve como el MID. Por lo tanto, había presentes 12 posibles códigos de DBR (2 × 2 × 3) y 81 diferentes MID (3 × 3 × 3 × 3).

Se mezclaron después las dos muestras juntas en cantidades iguales para crear, en efecto, un perfecto heterocigoto 50/50 de tres bases **A** y **G** aguas abajo del MID (es decir, la secuencia DHVB). Diferentes cantidades de la mezcla (100 ng, 300 ng, 600 ng, 2500 ng, 5000 ng, y 10.000 ng) se sometieron a reacciones de interacción (*pull-down*) de hibridación seguidas por la amplificación a través de 10 ciclos de PCR con cebadores TiA y TiB. Las sondas de captura empleadas fueron oligonucleótidos purificados en cartucho en fase inversa de 60 meros 5'-biotinilados (Biosearch). Después de la amplificación con TiA y TiB, se utilizó una reacción secundaria de PCR con TiA y un cebador específico de la secuencia con cola de TiB

(5'-CCTATCCCCTGTGTGCCTTGGCAGTCTCAGGGACACCCAGCCAAGACAGC-3') (SEQ ID NO: 1) para amplificar un fragmento específico. El fragmento de PCR generado a partir de cada muestra en la hibridación *pull-down*/PCR fue enviado para secuenciación con Shot 454 Ti para determinar la DBR, el MID y los alelos A/G.

La secuencia del amplicón procedente de la hibridación *pull down*/PCR se muestra a continuación (SEQ ID NO: 2). La región DBR está subrayada, el MID está en negrita, y el alelo (R, que corresponde a A o G) está en negrita y subrayado.

CCATCTCATCCCTGCGTGTCTCCGACTCAGRYB**DHVB**AC**R**TAGAATGTGCATGGATCGTATG

AGCACCTGTGGGCAGGGCAAGTGGCAGATGCCTTAGTGGATCTCACTGGAAGCCTGGCAGAA

AGGTGGAGCTTGAAGGATGTAACGAAAGCCAGCGGCCAGCAGGACAGACCCAGTGGTGGGGA

GCACAGAACTTGTCGGCAGCTACTCCACCTGAAGGACCGGTGTCTAATCAGCTGCTCTGTGC

TTAGCCCCAGAGCAGGTACAGCTATGGCTACAACTCCCTCCACCATTAGCTTGTTACAGAGA

AGGAAATCGGTCCTTGAGAGGCTGTCTTGGCTGGGTGTCCCTGAGACTGCCAAGGCACACAG

GGGATAGG

Resultados

La Figura 1 muestra la relación de alelos para cada MID en la muestra. Los números en la parte superior de cada uno de los 6 paneles muestran la masa de entrada de ADN genómico utilizada (en nanogramos). El eje de abscisas muestra, para cualquier MID particular, la fracción de las lecturas de secuenciación de A, o *calls* (en inglés), (es decir, el número de lecturas de una de las bases polimórficas en la posición del SNP sintético) frente al número total de lecturas, por ejemplo, la relación de (lecturas de A)/(lecturas de A + lecturas de G). Esto se denomina la relación de alelos. El eje de ordenadas es el número de los MID observados con una relación particular de alelos (el número total de los MID, como se señaló anteriormente, es 81). Debido a que era sabido que el ADN de entrada tenía una relación de A/G 50/50, la relación de alelos para cada muestra debería ser 0,5.

Para una masa de entrada baja, la relación de alelos se distorsiona - demasiado alta o demasiado baja - debido a que las moléculas de entrada a la primera etapa de la PCR eran limitantes y por lo tanto se observó preferentemente uno de los dos alelos. Como la masa aumenta en la primera etapa, se observan ambos alelos y están cada vez más próximos a la relación esperada de 0,5. Este análisis demuestra que hubo una considerable caída de los alelos a 100 ng, 300 ng y 600 ng, mientras que hubo una pequeña o ninguna caída de los alelos en los niveles de entrada más altos.

La Figura 2 muestra la fracción de secuencias de DBR para cada MID asociada con cada alelo. Dado que el material de entrada es nominalmente A y G a 50/50 en la posición polimórfica sintética, es de esperar que cada uno de los 81 MID efectivos esté asociado con las lecturas del 50 % de A y 50 % de G (como se señaló anteriormente). Además, puesto que las 12 DBR son aleatorias y se asocian con cada uno de los 81 MID diferentes, se podría esperar, con suficientes copias de ADN de entrada, que las 12 DBR se observarían para el alelo A y las 12 para el alelo G. Por lo tanto, la fracción de las DBR observada para una base particular en el caso ideal debería ser 12/24 o 0,5 para cada alelo. En la medida en que un número insuficiente de moléculas entre en los etapas de proceso, puede haber menos de 12 DBR para cada alelo de MID, y así la relación se puede desviar de la ideal.

En la Figura 2, el eje de abscisas muestra, para cualquier MID particular, la proporción de las DBR realmente observadas para cada alelo dividida por el número total de las DBR realmente observadas. El eje de ordenadas es el número de los MID observados con una proporción particular o las DBR asociadas con un alelo. Se ven, y se deberían ver, un total de 81 MID diferentes. Se observa que en una masa de entrada baja, se puede ver con

frecuencia una proporción distorsionada, demasiado alta o demasiado baja. Esto es debido probablemente al número limitante de moléculas de entrada en la primera etapa de la PCR, y por lo tanto se observó preferentemente uno de los dos alelos. Como aumenta la masa en la primera etapa, se observan ambos alelos con más frecuencia y se asocia con cada alelo una observación de cada vez más de las DBR, y por lo tanto una proporción más cerca del 0,5.

Una comparación de los datos en las figuras 1 y 2 muestra una característica adicional del empleo de las DBR como se describe aquí. Para un heterocigoto verdadero, se empieza a ver antes (es decir, a una masa más baja) en el análisis de DBR (Figura 2) que los grupos de distribución están alrededor de la relación esperada de 0,5. Esto tiene sentido porque, por ejemplo, la observación de 6 de las 12 DRB para el alelo A y 4 de las 12 DBR del alelo G se traduciría en una proporción de (6/[6 + 4]) = 0,60 que, de hecho, está bastante cerca de la esperada 0,5. El efecto neto es que el uso de las DBR da mucha mayor confianza en presencia de un verdadero heterocigoto o un verdadero homocigoto.

Aunque la invención precedente se ha descrita en algún detalle a modo de ilustración y ejemplo con fines de claridad de comprensión, es fácilmente evidente para los expertos en la técnica a la luz de las enseñanzas de esta invención que se pueden hacer ciertos cambios y modificaciones de la misma sin apartarse del alcance de las reivindicaciones adjuntas.

Por consiguiente, el texto precedente simplemente ilustra los principios de la invención. Se debe apreciar que los expertos en la técnica serán capaces de idear diversos arreglos que, aunque no se describan ni se muestren explícitamente en el presente documento, incorporan los principios de la invención y están incluidos dentro de su alcance. Además, todos los ejemplos y el lenguaje condicional presentados en este documento tienen como principal objetivo ayudar al lector en la comprensión de los principios de la invención y los conceptos aportados por la invención para la promoción de la técnica, y se deben interpretar como no limitados a dichos ejemplos y condiciones específicamente indicados. Por lo tanto, no se pretende que el alcance de la presente invención se limite a las realizaciones mostradas como ejemplos y descritas en el presente documento. Más bien, el alcance de la presente invención se materializa en las reivindicaciones adjuntas.

Aunque la invención anterior se ha descrito con cierto detalle a modo de ilustración y ejemplo a los efectos de la claridad de la comprensión, es evidente para los expertos en la técnica a la luz de las enseñanzas de esta invención que se pueden hacer ciertos cambios y modificaciones en la misma sin apartarse del alcance de las reivindicaciones adjuntas.

En consecuencia, lo anterior simplemente ilustra los principios de la invención. Se apreciará que los expertos en la materia podrán diseñar diversas disposiciones que, aunque no se describen explícitamente o se muestran en este documento, incorporan los principios de la invención y se incluyen dentro de su alcance. Además, todos los ejemplos y el lenguaje condicional enumerados en la presente memoria están destinados principalmente a ayudar al lector a comprender los principios de la invención y los conceptos aportados por los inventores para mejorar la técnica, y deben interpretarse sin limitación a dichos ejemplos y condiciones específicamente citados. El alcance de la presente invención, por lo tanto, no está destinado a limitarse a las realizaciones ejemplares que se muestran y describen en este documento. Por el contrario, el alcance de la presente invención está determinado por las reivindicaciones adjuntas.

Listado de secuencias

<110> Population Genetics Technologies Ltd.

<120> Aumento de la confianza en las identificaciones de alelos con el recuento molecular

<130> KE/N30061

<150> 61/385,001
<151> 2010-09-21

<150> 61/432,119
<151> 2011-01-12

<160> 2

<170> FastSEQ para Windows Versión 4.0

<210> 1
<211> 50
<212> DNA
<213> Secuencia Artificial

<220>
<223> secuencia sintética

<400> 1

cctatcccct gtgtgccttg gcagtctcag ggacacccag ccaagacagc 50

<210> 2
<211> 380
<212> DNA
<213> Secuencia Artificial

<220>
<223> secuencia sintética

<400> 2

ccatctcatc cctgcgtgtc tccgactcag rybdhvbacr tagaatgtgc atggatcgta 60
tgagcacctg tgggcagggc aagtggcaga tgccttagtg gatctcactg gaagcctggc 120
agaaaggtgg agcttgaagg atgtaacgaa agccagcggc cagcaggaca gacccagtgg 180
tggggagcac agaacttgtc ggcagctact ccacctgaag gaccggtgtc taatcagctg 240
ctctgtgctt agccccagag caggtacagc tatggctaca actccctcca ccattagctt 300
gttacagaga aggaaatcgg tccttgagag gctgtcttgg ctgggtgtcc ctgagactgc 360
caaggcacac aggggatagg 380

REIVINDICACIONES

1. Un método para determinar el número mínimo de moléculas de polinucleótidos individuales que se originan a partir de la misma región genómica de la misma muestra original que han sido secuenciadas en una configuración o procedimiento de análisis de secuencia particular, que incluye:

unir una región de bases degeneradas (DBR) a moléculas de polinucleótidos de partida;
amplificar las moléculas de polinucleótidos de partida unidos a DBR;
secuenciar las moléculas de polinucleótidos amplificadas, en donde se obtiene la secuencia de la DBR así como una porción del polinucleótido;
determinar el número de DBR diferentes unidas a un polinucleótido de interés; y
usar el número de secuencias de DBR diferentes presentes en la etapa de secuenciación para determinar el número mínimo de moléculas de polinucleótidos individuales que se originan a partir de la misma región genómica de la misma muestra original que han sido secuenciadas en la configuración o el procedimiento de análisis de secuencias particular;
donde el método incluye acervar moléculas de polinucleótidos de una pluralidad de muestras originales, donde las moléculas de polinucleótidos derivadas de cada muestra original incluyen una etiqueta de identificador múltiplex (MID), en donde cada muestra original se correlaciona con una única MID de modo que se puede determinar la muestra original a partir de la cual se obtuvo cada molécula de polinucleótido etiquetada.

2. Un método según la reivindicación 1, en donde la DBR se añade a los polinucleótidos de partida como parte de un adaptador.

3. Un método según la reivindicación 2, en donde la DBR está en un adaptador que también incluye un sitio para el cebador de la secuenciación.

4. Un método según la reivindicación 1, 2 o 3, que incluye determinar un valor estadístico para una determinación de alelo en un ensayo de genotipado que no se puede obtener a partir del número de lectura a solas.

5. Un método según la reivindicación 1, en donde la DBR tiene una longitud de 3 a 10 bases.

6. Un método según la reivindicación 1, en donde los polinucleótidos son de ADN genémico.

7. Un método según la reivindicación 6, en donde los polinucleótidos en la muestra de ácido nucleico se derivan de un ser humano.

8. Un método según la reivindicación 2, en donde antes del ligamiento del adaptador se enriquecen las moléculas de polinucleótidos de muestras parentales individuales para reducir su complejidad.

9. Un método según la reivindicación 1, que incluye enriquecer los polinucleótidos unidos.

10. Un método según la reivindicación 1, en donde la DBR está presente en diferentes sitios en un polinucleótido.

11. Un método según la reivindicación 1, en donde la DBR está presente en un cebador de síntesis de ácidos nucleicos, tal que la DBR se añade a un polinucleótido diana cuando se usa el cebador en una reacción de polimerización.

12. Un método según la reivindicación 11, en donde el cebador de la síntesis de ácidos nucleicos es un cebador de PCR.

13. Un método según la reivindicación 12, que incluye determinar el número de moléculas usadas como moldes para una reacción PCR.

14. Un método según la reivindicación 1, en donde el método es utilizado para determinar la heterogeneidad genética de las moléculas de polinucleótidos en la muestra, en donde la muestra comprende moléculas de polinucleótidos derivadas de tumores, microorganismos y/o virus.

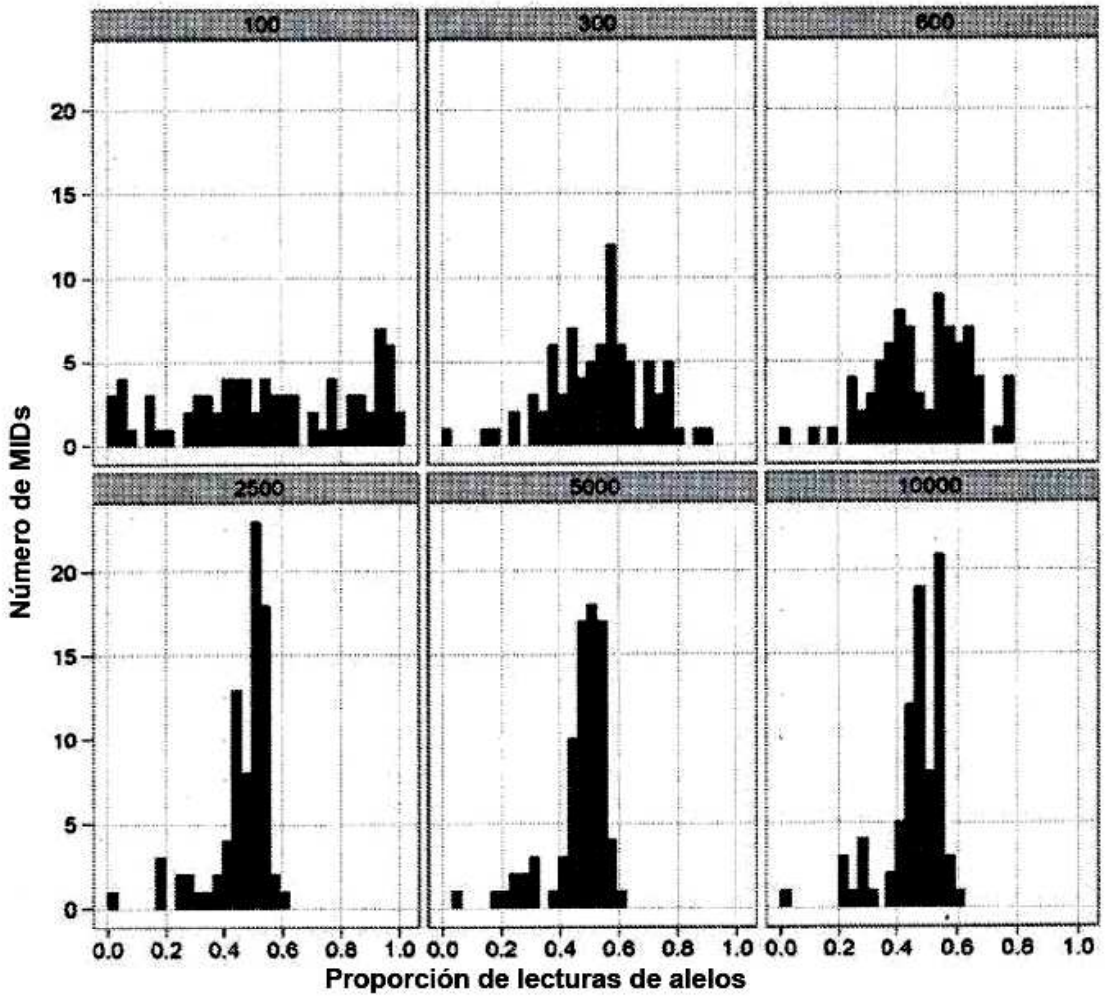
100
300
600
2500
5000
10000
Número de MIDs
0
5
10
15
20
0.0
0.2
0.4
0.6
0.8
1.0
Proporción de lecturas de alelos

Fig. 1

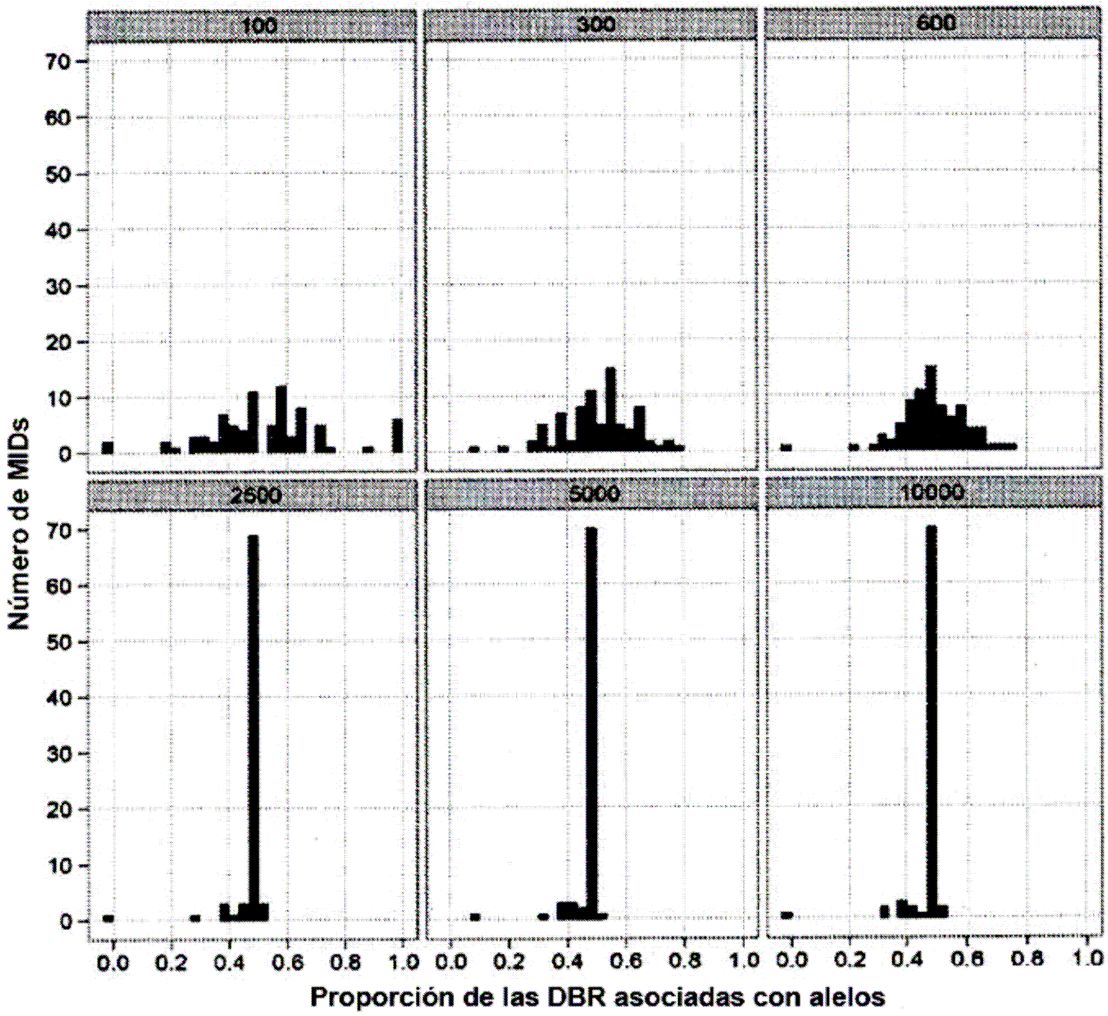
100
300
600
2500
5000
10000
Número de MIDs
Proporción de las DBR asociadas con alelos

Fig. 2

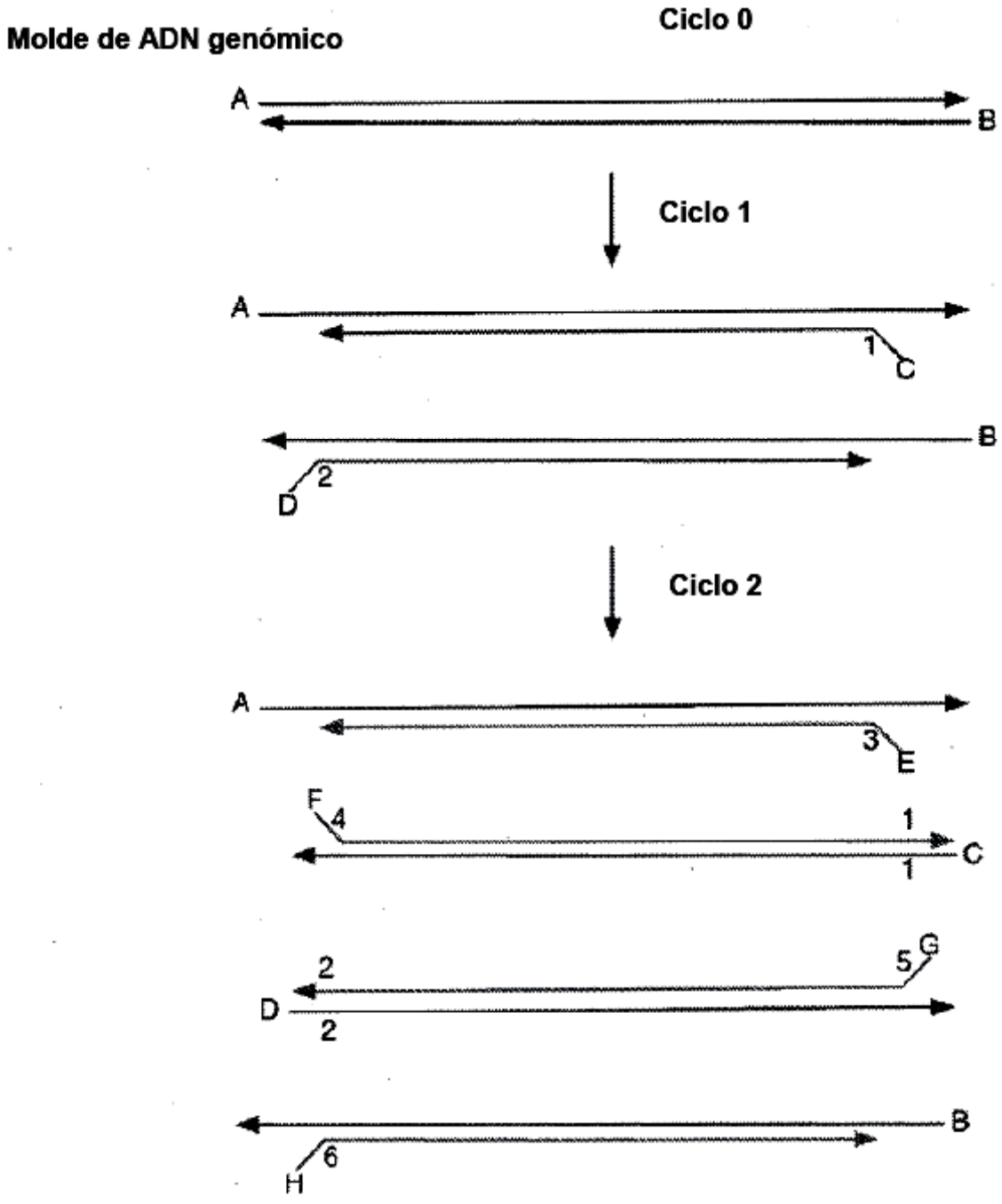
Molde de ADN genómico
Ciclo 0
A
B
Ciclo 1
A
1
C
B
D
2
Ciclo 2
A
3
E
F
4
1
C
1
2
5
G
D
2
B
H
6

Fig: 3