(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 397 683**

(51) Int. CI.:

C12Q 1/68 (2006.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(96) Fecha de presentación y número de la solicitud europea: **24.02.2009 E 09714916 (5)**

(97) Fecha y número de publicación de la concesión europea: **17.10.2012 EP 2250289**

(54) Título: **Método y sistemas para el enriquecimiento uniforme de regiones genómicas**

(30) Prioridad:

29.02.2008 US 32594 P

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
08.03.2013

(73) Titular/es:

F. HOFFMANN-LA ROCHE AG (100.0%)
Grenzacherstrasse 124
4070 Basel, CH

(72) Inventor/es:

KITZMAN, JACOB;
ALBERT, THOMAS;
D'ASCENZO, MARK;
JEDDELOH, JEFF;
MIDDLE, CHRISTINA;
RICHMOND, TODD y
RODESCH, MATTHEW

(74) Agente/Representante:

ISERN JARA, Jorge

DESCRIPCIÓN

Métodos y sistemas para el enriquecimiento uniforme de regiones genómicas

Campo de la invención

La presente invención proporciona métodos y composiciones para el enriquecimiento de ácidos nucleicos diana en un sistema de microchips. En particular, la presente invención proporciona métodos y composiciones para el enriquecimiento uniforme de moléculas de ácido nucleico diana en un formato de microchips. La presente descripción también proporciona métodos para el enriquecimiento intencionado no uniforme entre moléculas de ácido nucleico diana.

Antecedentes de la invención

La llegada de la tecnología de microchips de ácido nucleico hace posible la construcción de un chip de millones de secuencias de ácido nucleico en un área muy pequeña, por ejemplo en un portaobjetos de microscopio (por ejemplo, US 6.375.903 y US 5.143.854). En un principio, estos chips se crearon mediante la dispersión de secuencias de DNA presintetizadas en un portaobjetos. Sin embargo, la construcción de sintetizadores de microchips sin máscara (MAS) como se describe en US 6.375.903 permite ahora la síntesis in situ de secuencias de oligonucleótidos directamente sobre el portaobjetos.

En la utilización de un instrumento MAS, la selección de las secuencias de oligonucleótidos que se construirán en el microchip se encuentra bajo el control de un programa informático de tal manera que ahora es posible crear chips personalizados individualmente según las necesidades particulares de un investigador. En general, la tecnología de síntesis de microchips de oligonucleótido basada en MAS permite la síntesis en paralelo de más de 4 millones de oligonucleótidos con características únicas en un área muy pequeña de un portaobjetos de microscopio estándar. Con la disponibilidad de los genomas completos de cientos de organismos, para los que por lo general se ha depositado una secuencia de referencia en una base de datos pública, se han utilizado los microchips para realizar el análisis de secuencias sobre los ácidos nucleicos aislados de una gran variedad de organismos.

La tecnología de microchips de ácidos nucleicos se ha aplicado a muchas áreas de investigación y de diagnóstico, tales como la expresión y el descubrimiento de genes, detección de mutaciones, comparación de la secuencia evolutiva y alélica, mapeado del genoma, descubrimiento de fármacos, y más. Muchas aplicaciones requieren la búsqueda de variantes genéticas y mutaciones a lo largo de todo el genoma humano; variantes y mutaciones que, por ejemplo, pueden ser la base de enfermedades humanas. En el caso de enfermedades complejas, estas búsquedas generalmente resultan en un polimorfismo de un solo nucleótido (SNP) o un conjunto de SNP asociados con una o más enfermedades. La identificación de tales SNP ha demostrado ser una ardua tarea que consume tiempo y dinero en la que se requiere con frecuencia la resecuenciación de grandes regiones de DNA genómico, por lo general mayor de 100 kilobases (Kb) de individuos afectados y / o muestras de tejido para encontrar un solo cambio de base o identificar todas las variantes de la secuencia.

El genoma es normalmente demasiado complejo para ser estudiado como un todo, y las técnicas deben utilizarse para reducir la complejidad del genoma. Para abordar este problema, una solución consiste en reducir ciertos tipos de secuencias abundantes a partir de una muestra de DNA, tal como se encuentra en US 6.013.440. Las alternativas utilizan métodos y composiciones para el enriquecimiento de las secuencias genómicas como se describe, por ejemplo, en Albert, TJ, et al., Nat. Meth., 4 (2007) 903-5, y Okou, DT, et al., Nat. Meth. 4 (11) (2007) 907-9. Albert et al. describe una alternativa que es coste efectiva y rápida en la reducción efectiva de la complejidad de una muestra genómica de una manera definida por el usuario para permitir procesamientos y análisis adicionales.

Sin embargo, es igualmente importante ser capaz de enriquecer secuencias diana de manera uniforme sobre las regiones diana. Si el enriquecimiento no es uniforme, por ejemplo, algunas secuencias diana se capturarán de manera desproporcionada en comparación con otras secuencias diana, negando así las aplicaciones posteriores que dependen de la distribución aproximadamente uniforme de las secuencias diana. Hodges, E., et al., Nature Genetics 39 (12) (2007) 1522-7 observó que un parámetro crítico en la captura del microchip fue la introducción de la captura de la diana sesgada que afecta en gran medida la profundidad de cobertura de la secuencia. Sin embargo, Hodges no ofreció una alternativa a seguir, aparte de decir que la redistribución de la sonda para compensar la captura sesgada necesariamente introducirá otros tipos de sesgos que conllevarán problemas con aplicaciones posteriores, como por ejemplo las aplicaciones de secuenciación.

Como tal, lo que se necesita son métodos y composiciones para proporcionar la captura uniforme, y por lo tanto la representación de las dianas capturadas durante la captura y el enriquecimiento de secuencias diana en un formato de microchip. A la inversa, un investigador podría también requerir una no-uniformidad de captura a propósito, por ejemplo si un investigador prevé buscar exones sobre las regiones intergénicas. Tales métodos proporcionarán una

utilidad de datos máxima a los investigadores en sus esfuerzos por identificar y comprender, por ejemplo, las causas de la enfermedad y los tratamientos terapéuticos asociados.

Resumen de la invención

La presente invención proporciona métodos para el enriquecimiento de ácidos nucleicos diana en un sistema de microchips. En particular, la presente invención proporciona métodos para el enriquecimiento uniforme de moléculas de ácido nucleico diana en un formato de microchips. La presente descripción también proporciona métodos para el enriquecimiento intencionado no uniforme entre moléculas de ácido nucleico diana.

El enriquecimiento del ácido nucleico reduce la complejidad de una muestra grande de ácido nucleico, tal como una muestra de DNA genómico, biblioteca de cDNA o biblioteca de mRNA, para facilitar su posterior procesamiento y análisis genético. Los métodos preexistentes de captura de ácidos nucleicos utilizan sondas de ácido nucleico inmovilizadas para capturar secuencias de ácido nucleico diana (por ejemplo, como las encontradas en el DNA genómico, DNAc, RNAm, etc) mediante la hibridación de la muestra a sondas inmovilizadas sobre un soporte sólido. Los ácidos nucleicos diana capturados, como los encontrados por ejemplo en el DNA genómico, se lavan y se eluyen de las sondas inmovilizadas en el soporte sólido. Las secuencias genómicas eluídas son más susceptibles para el análisis genético detallado que una muestra genómica que no se ha sometido a este procedimiento. El enriquecimiento de las secuencias de ácido nucleico diana lleva la captura de ácidos nucleicos un paso más allá, al reducir la complejidad de una muestra en la que las secuencias de interés se seleccionan por, o se enriquecen, mediante procesos selectivos. Los métodos de enriquecimiento y las composiciones se describen completamente en la solicitud de patente US Números 11/789.135 y 11 / 970.949 y de la Solicitud de la Organización Mundial de la Propiedad Intelectual número PCT/US07/010064.

El enriquecimiento de ácidos nucleicos diana en un formato de microchip es importante para la reducción de la complejidad de una muestra de ácido nucleico antes de, por ejemplo, la secuenciación u otras aplicaciones posteriores. Sin embargo, muchas aplicaciones posteriores dependen fuertemente de la lectura de secuencias resultantes que tiene una distribución aproximadamente uniforme en las regiones diana, como representación desproporcionadamente alta de algunas dianas que necesariamente agota las demás. Aunque el enriquecimiento basado en chips enriquece fuertemente fragmentos específicos, se contempla que ciertas dianas están más fuertemente enriquecidas que otras produciendo de ese modo la dianas o capturas sesgadas.

Como tal, la presente invención proporciona métodos para abordar esta captura sesgada de ácidos nucleicos diana. Por ejemplo, las realizaciones de la presente invención proporcionan para el diseño de los chips que se modifican para redistribuir sondas de las dianas con enriquecimiento superior a la media para aquellas que tienen enriquecimiento por debajo de la media. En el desarrollo de formas de realización de la presente invención, se determinó que esta redistribución de las sondas mejora de manera significativa la uniformidad de enriquecimiento entre las dianas capturadas. Por el contrario, la presente descripción también proporciona para el diseño del chip que se modifica para redistribuir sondas que introducen intencionalmente captura parcial de dianas en un chip. Por ejemplo, si un investigador está interesado en la captura de regiones genómicas específicas sobre otras regiones genómicas, los métodos descritos en este documento pueden ser utilizados para crear capturas sesgadas.

Ciertas realizaciones ilustrativas de la invención se describen a continuación. La presente invención no se limita a estas realizaciones.

En algunas realizaciones, la presente descripción comprende un microchip de soporte sólido, que comprende en general sondas de ácido nucleico inmovilizadas en un soporte para capturar y para enriquecer secuencias específicas de ácidos nucleicos (ácidos nucleicos diana) de una muestra (por ejemplo, DNA genómico, DNAc, RNAm, RNAt, etc.). En algunas realizaciones, las sondas que están inmovilizadas sobre un soporte representan un conjunto de sondas redistribuido. Por ejemplo, las sondas redistribuidas están diseñadas para proporcionar la captura uniforme de regiones diana, de tal manera que la captura de dianas no está sesgada. En algunas realizaciones, las sondas que están inmovilizadas sobre un soporte son sondas redistribuidas, en las que dichas sondas están diseñadas para proporcionar una captura no uniforme de regiones diana, de tal manera que la captura de dianas está sesgada intencionadamente.

En algunas realizaciones, el enriquecimiento de ácido nucleico diana es a través de la hibridación de una muestra de ácido nucleico, por ejemplo, una muestra de DNA genómico, que puede contener una o más secuencias de ácido nucleico diana, en contra de un microchip que comprende sondas de ácido nucleico redistribuidas dirigidas a una región o regiones específicas del genoma. Después de la hibridación, las secuencias de ácidos nucleicos diana presentes en la muestra se enriquecen con el lavado del chip y eluyendo del chip los ácidos nucleicos genómicos hibridados. Después de la elución, las muestras enriquecidas se analizan para determinar el nivel o cantidad de enriquecimiento sobre un control. En algunas realizaciones, la secuencia de ácido nucleico diana se amplifica aún más utilizando, por ejemplo, PCR mediada por ligación no específica (LM-PCR), resultando en un grupo de

productos de PCR amplificados de complejidad reducida en comparación con la muestra original para secuenciación, construcción de librerías, y otras aplicaciones. En algunas formas de realización, el ensayo que comprende sondas redistribuidas para la captura de secuencias diana demuestra una captura uniforme, no sesgada sobre la región diana tal como se ejemplifica en la figura 1.

En algunas realizaciones, la presente descripción comprende un soporte sólido, que comprende en general sondas de ácido nucleico inmovilizadas en un soporte para capturar secuencias específicas de ácidos nucleicos (ácidos nucleicos diana) de una muestra (por ejemplo, DNA genómico, DNAc, RNAm, RNAt, etc .). En algunas realizaciones, el soporte sólido es un portaobjetos, por ejemplo un portaobjetos de microchip. En algunas realizaciones, el soporte sólido comprende cuentas, mientras que las cuentas están en solución, por ejemplo en un tubo u otro tipo de recipiente, o por ejemplo en alícuotas en pocillos de una placa de ensayo (por ejemplo, 12 pocillos, 24 pocillos, 96 pocillos, 384 pocillos, y similares). En algunas realizaciones, las sondas que están inmovilizadas sobre un soporte representan un conjunto de sondas redistribuido. Por ejemplo, las sondas redistribuidas están diseñadas para proporcionar la captura uniforme de las regiones de la molécula de ácido nucleico diana, de tal manera que la captura de dianas no está sesgada, y de tal manera que la frecuencia de cada secuencia individual de las sondas inmovilizadas corresponde a la frecuencia de la secuencia de ácido nucleico diana correspondiente dentro de la población de las moléculas de ácidos nucleicos diana. En algunas realizaciones, las sondas que están inmovilizadas sobre un soporte son sondas redistribuidas, en las que dichas sondas están diseñadas para proporcionar una captura no uniforme de las regiones diana, de tal manera que la captura de dianas está sesgada intencionalmente. En algunas realizaciones, la muestra está fragmentada, por ejemplo mediante sonicación u otros métodos capaces de fragmentar ácidos nucleicos. En algunas realizaciones, la muestra fragmentada (por ejemplo, el DNA genómico fragmentado, DNAc, etc) está modificada mediante la ligación de enlazantes en uno o ambos extremos 5 'y 3'. En algunas realizaciones, los extremos 5 'y 3' de una muestra fragmentada se preparan en primer lugar para la ligación con un enlazante, por ejemplo mediante la realización de una reacción de "relleno" con la enzima Klenow.

La preparación de extremos de ácido nucleico para la posterior ligación a enlazantes es bien conocida en la técnica, y se pueden encontrar en cualquier manual de clonaje molecular tal como "Molecular Cloning: A Laboratory Manual, Sambrook, et al Eds, Cold Spring Harbor. Laboratory Press". En efecto, los ejemplos de métodos para la realización de todo el clonaje molecular, hibridación, lavado, y técnicas de elución como se utiliza aquí se puede encontrar en "Molecular Cloning: A Laboratory Manual", Sambrook, et al, Eds, Cold Spring Harbor Press, así como en "A Molecular Cloning Manual: DNA Microchips", Bowtell, et al, Eds, Cold Spring Harbor Press, así como otros manuales técnicos y guías de referencia conocidos por los expertos en la materia. En algunas formas de realización, la muestra fragmentada y ácido nucleico adaptado - enlazador se hibrida con un chip que comprende sondas redistribuidas diseñadas para capturar secuencias diana de una forma no sesgada, y las secuencias diana son capturadas. En otras realizaciones, la muestra fragmentada y ácido nucleico adaptado - enlazador se hibrida con un chip que comprende las sondas redistribuidas diseñadas intencionalmente para capturar las secuencias diana de manera sesgada, y las secuencias diana son capturadas. El uso de enlazantes para los métodos de enriquecimiento y métodos de enriquecimiento en general son bien conocidos y se describen completamente en la US nº 11/789.135 y 11 /970.949 y de la solicitud de patente de la Organización Mundial de la Propiedad Intelectual Número de solicitud PCT/US07/010064, y aún más en Albert, TJ, et al., Nat. Meth. 4 (2007) 903-5, Okou, D.T., Nat. Meth. 4 (11) (2007) 907-9 y Hodges, E., et al., Nature Genetics 39 (12) (2007) 1522-7.

Después de la hibridación, los ácidos nucleicos no diana se lavan del microchip y los ácidos nucleicos diana unidos, se eluyen del chip siguiendo los protocolos conocidos en la técnica. La calidad de la muestra enriquecida se calcula y la cantidad de enriquecimiento se determina y comunica al usuario. En algunas realizaciones, el cálculo de enriquecimiento comprende el número de veces de enriquecimiento en comparación con una muestra de control de enriquecimiento. Las muestras de suficiente calidad se utilizan para aplicaciones posteriores, tales como la secuenciación, la clonación, la construcción de librerías, etc.

La presente invención no está limitada por ningún uso posterior de los ácidos nucleicos enriquecidos, y un experto en la técnica entenderá la multitud de usos que puede proporcionar una muestra incluyendo, pero sin limitarse a, la secuenciación, la detección de SNP para el descubrimiento y la correlación con estados de enfermedad y factores de riesgo, el uso de secuencias específicas en aplicaciones de descubrimiento de medicamentos, etc.

Las secuencias diana enriquecidas pueden analizarse para, por ejemplo, la calidad de ácidos nucleicos diana enriquecidos basados en microchip (por ejemplo, nivel de efectividad de los métodos de enriquecimiento no sesgados (o intencionadamente sesgados) como se describe en el presente documento). Dicho análisis no sólo da una idea de la eficacia general de la tecnología de enriquecimiento, sino que también proporciona al investigador un método de acceso a la calidad de los ácidos nucleicos enriquecidos antes de gastar tiempo y recursos en las aplicaciones posteriores con una muestra que no está debidamente enriquecida. En algunas realizaciones, la evaluación de la calidad de los ácidos nucleicos diana se efectúa analizando el enriquecimiento de un subconjunto de secuencias de referencia, por ejemplo, regiones conservadas en un genoma, tal como se encuentra en la solicitud de patente provisional US 61/026.596.

En una realización, la presente invención comprende un método para el enriquecimiento uniforme de una población de moléculas de ácido nucleico en una muestra, que comprende proporcionar una muestra de moléculas de ácido nucleico que comprenden una pluralidad de secuencias de ácido nucleico diana, hibridar la muestra a un soporte que comprende sondas inmovilizadas de ácido nucleico en condiciones para apoyar la hibridación entre las sondas de ácido nucleico inmovilizado y la pluralidad de secuencias de ácidos nucleicos diana, en el que dichas sondas inmovilizadas de ácido nucleico son complementarias a dicha pluralidad de secuencias de ácidos nucleicos diana, en la que la densidad de dichas sondas de ácido nucleico inmovilizadas para producir de manera óptima una profundidad de lectura uniforme se predijo usando un modelo de regresión lineal ajustado empíricamente, que ajusta la profundidad de lectura con la densidad de sondas de ácidos nucleicos inmovilizadas, y en las que dichas sondas de hibridación de ácidos nucleicos inmovilizadas proporcionan una hibridación uniforme entre dicha pluralidad de secuencias de ácidos nucleicos diana, y la separación de secuencias no hibridadas de ácidos nucleicos a partir de secuencias diana de ácido nucleico hibridado y así enriquecer una población de moléculas de ácido nucleico en una muestra. En algunas realizaciones, la separación de las secuencias hibridadas de las no hibridadas comprende lavar el soporte de tal manera que las secuencias de ácidos nucleicos no hibridadas se retiran del soporte. En algunas realizaciones, las moléculas de ácido nucleico están fragmentadas antes de la hibridación y en otras realizaciones los fragmentos se ligan a moléculas adaptadoras en uno o ambos extremos. En algunas realizaciones, las moléculas de ácidos nucleicos fragmentados adaptadas al enlazante se desnaturalizan antes de la hibridación. En algunas realizaciones, las secuencias de ácidos nucleicos diana hibridadas se eluyen del soporte y muchas veces se secuencian después de la elución. En algunas realizaciones, el soporte es un soporte sólido, en el que dicho soporte sólido es un portaobjetos de microchip o una cuenta. En realizaciones preferidas, las moléculas de ácidos nucleicos son moléculas de DNA genómico o moléculas de DNA genómico amplificados. En realizaciones preferidas, las sondas de ácidos nucleicos se caracterizan en que la frecuencia de las secuencias individuales de las sondas de ácido nucleico inmovilizado corresponden a la frecuencia de la correspondiente pluralidad de secuencias de ácido nucleico dentro de una población de moléculas de ácido nucleico, en el que la determinación de la frecuencia comprende la utilización de un modelo de regresión lineal ajustado empíricamente.

En una realización, la presente descripción comprende un soporte sólido y una pluralidad de sondas de ácido nucleico inmovilizadas sobre dicho soporte sólido, en el que cada una de dicha pluralidad de sondas de ácido nucleico inmovilizadas proporciona una hibridación uniforme entre una pluralidad de secuencias de ácidos nucleicos diana.

Puede proporcionarse un equipo para realizar un enriquecimiento uniforme de secuencias de ácido nucleico que comprende uno o más recipientes, en el que dichos recipientes comprenden un soporte sólido que comprende sondas de ácido nucleico inmovilizadas, en el que dichas sondas se seleccionan de un grupo que consiste de una pluralidad de sondas capaces de hibridar con una pluralidad de secuencias de ácido nucleico y en el que dichas sondas proporcionan el enriquecimiento uniforme de dicha pluralidad de secuencias de ácidos nucleicos diana, y uno o más reactivos para realizar hibridaciones, lavados, y elución de las secuencias de ácido nucleico diana.

Puede proporcionarse un proceso para el enriquecimiento uniforme de una población de secuencias de ácidos nucleicos en una muestra que comprende una pluralidad de secuencias de hibridación de sondas inmovilizadas en el que la frecuencia de las secuencias individuales de las sondas de hibridación inmovilizadas corresponde a la frecuencia de una pluralidad de secuencias de ácido nucleico diana correspondientes dentro de una población de moléculas de ácido nucleico, y en el que dicho proceso para el enriquecimiento uniforme comprende hibridar dichas sondas a las secuencias de ácidos nucleicos diana correspondientes y separar las secuencias de ácidos nucleicos no hibridadas de las secuencias hibridadas de ácido nucleico diana. En algunas realizaciones, el procedimiento comprende además la elución de las secuencias de ácido nucleico diana hibridadas. En realizaciones preferidas, la hibridación de la sonda y las secuencias diana en el proceso se realiza en un soporte sólido tal como un portaobjetos de microchip o una cuenta. En realizaciones preferidas, la determinación de la frecuencia de las secuencias de la sonda comprende la utilización de un modelo de regresión lineal ajustado empíricamente. En algunas realizaciones, la muestra utilizada en el proceso es una muestra de DNA genómico.

Definiciones

Tal como se usa en este documento, el término "muestra" se utiliza en su sentido más amplio. En un sentido, se entiende que incluye una muestra de ácido nucleico obtenida de cualquier fuente. Las muestras biológicas de ácidos nucleicos se pueden obtener a partir de animales (incluidos los humanos) y abarcan los ácidos nucleicos aislados a partir de líquidos, sólidos, tejidos, etc. Las muestras biológicas de ácidos nucleicos también pueden provenir de animales no humanos, que incluye, pero sin limitarse a, vertebrados tales como roedores, primates no humanos, ovinos, bovinos, rumiantes, lagomorfos, porcinos, caprinos, equinos, caninos, felinos, aves, etc. Los ácidos nucleicos biológicos también se pueden obtener a partir de procariotas, como las bacterias y otros eucariotas no animales tales como las plantas. Se contempla que la presente invención no está limitada por la fuente de la muestra de ácidos nucleicos, y cualquier ácido nucleico a partir de cualquier reino biológico encuentra utilidad en los métodos

descritos en el presente documento.

Tal como se utiliza aquí, el término "molécula de ácido nucleico" se refiere a cualquier molécula que contiene ácido nucleico, que incluye pero no se limita a, DNA o RNA. El término abarca secuencias que incluyen cualquiera de los análogos de base conocidos de DNA y RNA que incluye pero no se limita a, 4-acetilcitosina, 8-hidroxi-N6-metiladenosina, aziridinilcitosina, pseudoisocitosina, 5-(carboxihidroxilmetil) uracilo, 5-fluorouracilo, 5-bromouracilo, 5-carboximetilaminometil-2-tiouracilo, 5-carboximetilaminometiluracilo, dihidrouracilo, inosina, N6-isopenteniladenina, 1-metiladenina, 1-metilpseudouracilo, 1-metilguanina, 1-metilinosina, 2,2-dimetilguanina, 2-metiladenina, 2-metilguanina, 3-metilcitosina, 5-metilcitosina, N6-metiladenina, 7-metilguanina, 5-metilaminometiluracilo, 5-metoxiaminometil-2-tiouracilo, beta-D-manosilqueosina, 5'-metoxicarbonilmetiluracilo, 5-metoxiuracili, 2-metiltio -N6-isopenteniladenina, uracilo-5oxiacetato de metilo, ácido uracilo-5-oxiacético, oxibutoxosina, pseudouracilo, queosina, 2-tiocitosina, 5-metil-2tiouracilo, 2-tiouracilo, 4-tiouracilo, 5-metiluracilo, uracil-N-5-oxiacetato de metilo, ácido uracilo-5-oxiacético, pseudouracilo, queosina, 2-tiocitosina, y 2,6 diaminopurina.

Tal como se utiliza aquí, el término "oligonucleótido" se refiere a una molécula compuesta de dos o más desoxirribonucleótidos o ribonucleótidos, preferiblemente más de tres, y por lo general más de diez. El tamaño exacto dependerá de muchos factores, que a su vez dependen de la función final o uso del oligonucleótido. El oligonucleótido puede generarse de cualquier manera, incluyendo la síntesis química, replicación de DNA, transcripción reversa, o una combinación de los mismos. El término oligonucleótido también se puede utilizar indistintamente con el término "polinucleótido".

Tal como se utiliza aquí, los términos "complementario" o "complementariedad" se utilizan en referencia a polinucleótidos relacionados con las reglas de emparejamiento de bases. Por ejemplo, la secuencia "5'-A-G-T-3'," es complementaria a la secuencia "3'-T-C-A-5' ". La complementariedad puede ser "parcial", en la que sólo algunas de las bases de los ácidos nucleicos "se comparan de acuerdo con las reglas de emparejamiento de bases. O, puede haber complementariedad "completa" o "total" entre los ácidos nucleicos. El grado de complementariedad entre cadenas de ácido nucleico tiene efectos significativos en, por ejemplo, la eficiencia y fuerza de hibridación entre las cadenas de ácido nucleico, la especificidad de amplificación, etc.

Tal como se utiliza aquí, el término "hibridación" se usa en referencia al emparejamiento de ácidos nucleicos complementarios. La hibridación y la fuerza de hibridación (es decir, la fuerza de la asociación entre los ácidos nucleicos) se ve afectada por factores tales como el grado de complementariedad entre los ácidos nucleicos, astringencia de las condiciones implicadas, la Tm del híbrido formado, y la proporción de G:C dentro de los ácidos nucleicos. Aunque la invención no se limita a un conjunto particular de condiciones de hibridación, se emplean preferiblemente las condiciones de hibridación astringentes. Las condiciones de hibridación astringentes dependen de la secuencia y serán diferentes al variar los parámetros ambientales (por ejemplo, concentraciones de sal, y la presencia de compuestos orgánicos). Generalmente, las "condiciones astringentes" se seleccionan para ser aproximadamente de 5 °C a 20 °C más bajas que el punto de fusión térmico (Tm) para la secuencia de ácido nucleico específica a una fuerza iónica y pH definidos. Preferiblemente, las condiciones astringentes son aproximadamente de 5 °C a 10 °C más bajas que el punto de fusión térmico para un ácido nucleico específico unido a un ácido nucleico complementario. La Tm es la temperatura (bajo una fuerza iónica y pH definidos) a la que el 50% de un ácido nucleico (por ejemplo, el ácido nucleico marcador) se hibrida con una sonda emparejándose perfectamente.

Tal como se utiliza aquí, el término "astringencia" se utiliza en referencia a las condiciones de temperatura, fuerza iónica, y la presencia de otros compuestos tales como disolventes orgánicos, bajo las cuales se llevan a cabo las hibridaciones de ácidos nucleicos. En "condiciones de baja astringencia" una secuencia de ácido nucleico de interés se hibridará con su complementario exacto, las secuencias con desemparejamientos de una sola base, secuencias estrechamente relacionadas (por ejemplo, las secuencias con homología del 90% o más), y las secuencias que tienen una homología parcial (por ejemplo, secuencias con homología de 50-90%). En "condiciones de astringencia media," una secuencia de ácido nucleico de interés se hibridará sólo con su complementario exacto, secuencias con desemparejamientos de una sola base, y secuencias en estrecha relación (por ejemplo, homología del 90% o más). Bajo "condiciones de alta astringencia", una secuencia de ácido nucleico de interés se hibridará sólo con su complementario exacto, y (dependiendo de las condiciones como la temperatura) las secuencias con desemparejamientos de una sola base. En otras palabras, bajo condiciones de alta astringencia, se puede elevar la temperatura a fin de excluir la hibridación con secuencias con desemparejamientos de una sola base.

A modo de ejemplo, "condiciones astringentes" o "condiciones de alta astringencia", comprenden la hibridación en formamida al 50%, 5 x SSC (0,75 M NaCl, citrato sódico 0,075 M), fosfato sódico 50 mM (pH 6,8), pirofosfato de sodio 0,1%, solución de Denhardt 5x, DNA de esperma de salmón sonicado (50 mg / ml), SDS 0,1%, y sulfato de dextrano al 10% a 42 °C, con lavados a 42 °C en 0,2 x SSC (cloruro sódico / citrato sódico) y formamida al 50% a 55 °C, seguido de un lavado con 0,1 x SSC que contiene EDTA a 55 °C. Para las condiciones moderadamente astringentes, se contempla que los tampones que contienen formamida al 35%, 5 x SSC, y dodecil sulfato de sodio

0,1% (p / v) son adecuados para la hibridación a 45 °C durante 16-72 horas. Además, se contempla que la concentración de formamida puede ajustarse adecuadamente entre un rango de 20 a 45% dependiendo de la longitud de la sonda y el nivel de astringencia deseado. En algunas realizaciones de la presente invención, la optimización de la sonda se obtiene para sondas más largas (por ejemplo, mayor de 50 meros) mediante el aumento de la temperatura de hibridación o la concentración de formamida para compensar un cambio en la longitud de la sonda. Ejemplos adicionales de condiciones de hibridación se proporcionan en muchos manuales de referencia, por ejemplo, en "Molecular Cloning: A Laboratory Manual".

De forma similar, las condiciones de lavado "astringentes" se determinan normalmente de forma empírica para la hibridación de secuencias diana a un chip de sondas correspondiente. Por ejemplo, los chips se hibridan primero y después se lavan con tampones de lavado que contienen concentraciones sucesivamente más bajas de sales, o mayores concentraciones de detergentes, o a temperaturas crecientes hasta que la proporción de señal-ruido para que la hibridación específica respecto a la no específica sea suficientemente alta para facilitar la detección de la hibridación específica. A modo de ejemplo, las condiciones astringentes de temperatura normalmente incluirán temperaturas superiores a aproximadamente 30 °C, más generalmente por encima de aproximadamente 37 °C, y ocasionalmente por encima de alrededor de 45 °C. Las condiciones astringentes de sal serán normalmente inferiores a aproximadamente 1000 mM, habitualmente inferiores a aproximadamente 500 mM, más habitualmente inferiores a aproximadamente 150 mM. Las condiciones de hibridación y de lavado astringentes son conocidas por los expertos en la materia, y se pueden encontrar en, por ejemplo, Crit Rev Wetmur, JG, y Davidson, N., J Mol Biol. 31 (1966) 349-70 y Wetmur, JG, Bio Mol Biol. 26 (3/4) (1991) 227-59.

Se conoce bien en la técnica que pueden emplearse numerosas condiciones equivalentes para ajustar y regular las condiciones de astringencia; se tienen en consideración factores como la longitud y naturaleza (DNA, RNA, composición de las bases) de la sonda y la naturaleza de la diana (DNA, RNA, composición de las bases, presente en solución o inmovilizado, etc.) y la concentración de las sales y otros componentes (por ejemplo, la presencia o ausencia de formamida, sulfato de dextrano, polietilenglicol). Como tal, los componentes y concentraciones de hibridación y las soluciones de de lavado pueden variar para generar las condiciones de astringencia. En realizaciones preferidas de la presente invención, las soluciones de hibridación y de lavado se utilizan tal como se encuentra comercialmente disponible a través de Roche-NimbleGen (por ejemplo, NimbleChip™ CGH chips, equipos de hibridación NimbleGen, etc).

Tal como se utiliza aquí, el término "cebador" se refiere a un oligonucleótido, tanto si se produce naturalmente en una digestión de restricción purificada o producido sintéticamente, que es capaz de actuar como un punto de iniciación de la síntesis cuando se coloca bajo condiciones en las que se induce la síntesis de un producto de extensión del cebador que es complementario a una cadena de ácido nucleico, (es decir, en presencia de nucleótidos y un agente inductor tal como una polimerasa de DNA y a una temperatura y pH adecuados). El cebador es preferiblemente de cadena sencilla para una máxima eficiencia en la amplificación, pero alternativamente puede ser de doble cadena. Si es de doble cadena, el cebador se trata primero para separar sus cadenas antes de ser utilizado para preparar productos de extensión. Preferiblemente, el cebador es un oligodesoxirribonucleótido. El cebador debe ser suficientemente largo para cebar la síntesis de productos de extensión en presencia del agente inductor. Las longitudes exactas de los cebadores dependerán de muchos factores, incluyendo la temperatura, fuente del cebador y la utilización del método.

El término "reacción en cadena de la polimerasa" ("PCR") se refiere a un método para aumentar la concentración de un segmento de una secuencia diana en una mezcla de DNA genómico sin clonación o purificación. Este proceso para amplificar la secuencia diana consiste en introducir un gran exceso de dos cebadores de oligonucleótidos a la mezcla de DNA que contiene la secuencia diana deseada, seguida por una secuencia precisa de ciclos térmicos en presencia de una polimerasa de DNA. Los dos cebadores son complementarios a sus respectivas cadenas de la secuencia diana de doble cadena. Para efectuar la amplificación, la mezcla se desnaturaliza y los cebadores a continuación se hibridan a sus secuencias complementarias dentro de la molécula diana. Tras la hibridación, los cebadores se extienden con una polimerasa para formar un nuevo par de cadenas complementarias. Los pasos de desnaturalización, hibridación del cebador, y extensión con polimerasa se pueden repetir muchas veces (es decir, desnaturalización, hibridación y extensión constituyen un "ciclo"; puede haber numerosos "ciclos") para obtener una concentración alta de un segmento amplificado de la secuencia diana deseada. La longitud del segmento amplificado de la secuencia diana deseada está determinada por las posiciones relativas de los cebadores con respecto a ellos, y por lo tanto, esta longitud es un parámetro controlable. Con la PCR, es posible amplificar una sola copia de una secuencia diana específica en DNA genómico a un nivel detectable mediante varias metodologías diferentes conocidas por los expertos en la materia. Además del DNA genómico, cualquier secuencia de oligonucleótido o polinucleótido puede ser amplificado con el conjunto apropiado de moléculas de cebador. En particular, los segmentos amplificados creados mediante el procedimiento de PCR, son ellos mismos moldes eficientes para posteriores amplificaciones mediante PCR. La PCR mediada por ligación se refiere a la PCR que se lleva a cabo, cuando los cebadores son homólogos (por ejemplo, complementarios) a enlazantes que se ligan a los extremos de DNA (por ejemplo, fragmentos de DNA).

Tal como se utiliza aquí, el término "sonda" se refiere a un oligonucleótido (es decir, una secuencia de nucleótidos), ya sea de origen natural como en una digestión por restricción purificada o producida sintéticamente, por recombinación o por amplificación mediante PCR, que es capaz de hibridarse al menos a una parte de otro oligonucleótido de interés. Una sonda puede ser de cadena sencilla o de cadena doble, sin embargo, en la presente invención, las sondas pretenden ser de cadena sencilla. Las sondas son útiles en la detección, identificación y aislamiento de secuencias de genes particulares.

Tal como se utiliza aquí, el término "porción" cuando, en referencia a una secuencia de nucleótidos (como en "una porción de una secuencia de nucleótidos determinada") se refiere a fragmentos de esa secuencia. Los fragmentos pueden variar en tamaño desde cuatro nucleótidos hasta la secuencia de nucleótidos completa menos un nucleótido (10 nucleótidos, 20, 30, 40, 50, 100, 200, etc.)

Tal como se utiliza aquí, el término "purificado" o "purificar" se refiere a la eliminación de componentes (por ejemplo, contaminantes) y / o los contaminantes de una muestra. El término "purificado" se refiere a moléculas, ya sean secuencias de ácido nucleico o de aminoácidos que se eliminan, aíslan o separan de su entorno natural. Una "una secuencia o muestra de ácido nucleico aislada" es por lo tanto una secuencia o muestra de ácido nucleico purificada. Las moléculas "sustancialmente purificadas" están libres en al menos un 60%, preferiblemente al menos un 75%, y más preferiblemente al menos un 90% libres de otros componentes con los que están asociados de forma natural.

Descripción de las figuras

Figura 1 muestra ejemplos de datos de redistribución de sondas utilizando la optimización empírica para mitigar la captura de diana sesgada localmente: a) profundidades de lectura a lo largo de la región central de 200kbp de los intervalos diana anidados en cinco experimentos de captura por separado demuestran grandes regiones diana que se correlacionan con profundidades de secuenciación inferior, b) respuesta de captura calculada dentro de una ventana localizada mediante ajuste de profundidad de lectura a la ventana para capturar la densidad de la sonda a través de los experimentos de captura; c) respuesta de captura a lo largo de la región en estudio muestra áreas de sesgo con captura excesiva o insuficiente; d) un chip de control muestra sondas distribuidas uniformemente a lo largo de la diana; e) mientras que en las sondas de chip optimizadas están redistribuidas de manera no uniforme a fin de lograr una distribución uniforme de lectura después de la captura y la secuenciación.

Figura 2 muestra la profundidad de la cobertura relativa representada a lo largo de la región diana de captura de un ejemplo de experimento de redistribución de la sonda control (Control, línea clara) y de reequilibrio (Rebal, línea oscura). La varianza en la cobertura es menos grave para el chip redistribuido cuando se compara con el chip control.

Figura 3 demuestra esquemas ejemplares para el enriquecimiento de la diana: a) una representación esquemática de moléculas de ácido nucleico y la utilización de la sonda como se encuentra en una realización de la presente invención, antes de la redistribución de la sonda, y b) un esquema de un ejemplo de estrategia de enriquecimiento de microchip de diana genómica de la presente invención.

Figura 4 muestra el efecto de la longitud del exón en la densidad de la sonda.

Figura 5 demuestra la falta de cualquier efecto agregado de nivel de longitud del exón en respuesta de la sonda de captura.

Figura 6 representa una comparación de las desviaciones estándar de las distribuciones de cobertura de la secuencia del locus diana de experimentos que utilizan cinco diseños de reequilibrado diferentes (RebalA hasta RebalE) y una línea de base (HumanExon7Chip) siguiendo un diseño de embaldosado estándar.

Descripción detallada de la invención

La secuenciación genómica dirigida es una de las aplicaciones biomédicas más importantes de las tecnologías de secuenciación de nueva generación. Un método revolucionario para dirigir la secuenciación de próxima generación utiliza microchips de DNA como los dispositivos de preparación de muestras. Estos chips capturan regiones del genoma definidas por las sondas del chip, que se eluyen a continuación y, por ejemplo, se secuencian. Debido al coste relativamente alto de ejecución de la secuenciación de nueva generación, es importante tener métricas sólidas de control de calidad que aseguren que sólo son secuenciadas las muestras que están altamente enriquecidas para las regiones diana. Dos características importantes de las muestras capturadas con éxito son: 1) regiones diana altamente enriquecidas, y 2) enriquecidas uniformemente a lo largo de todas las regiones diana. La presente

invención proporciona ensayos que demuestran el enriquecimiento uniforme a lo largo de las regiones diana de un genoma.

La captura de secuencias en un formato de microchip facilita el enriquecimiento selectivo de los ácidos nucleicos antes de las aplicaciones posteriores, por ejemplo la secuenciación. Al realizar el enriquecimiento selectivo, una muestra de ácido nucleico, por ejemplo una muestra de DNA o RNA, se hibrida con un microchip que comprende sondas de oligonucleótidos complementarias a las secuencias diana deseadas. Los ácidos nucleicos diana, capturados se eluyen del microchip, con la fracción resultante enriquecida varios órdenes de magnitud para los fragmentos específicos cuando se compara con un microchip control. Los métodos de enriquecimiento están más completamente descritos en las solicitudes de patente US Números 11/789.135 y 11 /970.949 y de la solicitud de la Organización Mundial de la Propiedad Intelectual Número PCT/US07/010064, y aún más en Albert, TJ, et al., Nat. Meth., 4 (2007) 903-5, Okou, D. T., et al., Nat. Meth. 4 (11) (2007) 907-9 y Hodges, E., et al., Nature Genetics 39 (12) (2007) 1522-7.

Muchas aplicaciones posteriores dependen fuertemente, por ejemplo, de una distribución aproximadamente uniforme de captura sobre la región diana de captura, y se contempla que la representación desproporcionadamente alta de algunas dianas agotan otras dianas. En el desarrollo de formas de realización de la presente invención, se han desarrollado nuevos métodos para tratar este sesgo, la captura de dianas desproporcionada, en el que las sondas son una redistribución de las dianas que demuestran un enriquecimiento por encima de la media de las sondas que demuestran un enriquecimiento por debajo de la media. Tal como se demuestra en este documento, los métodos de redistribución de sondas del presente método mejoran significativamente la uniformidad de enriquecimiento entre las dianas capturadas.

La presente descripción proporciona métodos para determinar y diseñar microchips que comprenden sondas de oligonucleótidos redistribuidas para permitir la captura uniforme, o intencionadamente no uniforme, de moléculas de ácido nucleico diana. En el desarrollo de formas de realización de la presente invención, se realizaron los experimentos de captura y secuenciación de microchips utilizando un conjunto anidado de regiones diana centradas en el cromosoma humano 17q21.31. Como una medida indirecta de la abundancia relativa de la secuencia diana tras la captura, se calculó la profundidad de la cobertura de secuencia como el número de lecturas que contiene una base diana determinada en promedio sobre el área diana. Se observó la existencia de un sesgo significativo y reproducible entre las regiones diana comunes de los microchips, de tal manera que la profundidad de cobertura abarca casi tres órdenes de magnitud y está altamente correlacionada entre los experimentos (relación a pares 0,85 <p <0,99).

En la realización de la experimentación en apoyo de las realizaciones de la presente invención, se determinó que la longitud del exón tiene un efecto sobre la densidad de la sonda. La Figura 4 es un ejemplo de la experimentación realizada utilizando un diseño de microchip de embaldosado estándar. En un diseño estándar, las sondas de captura están localizadas normalmente de forma desproporcionada en intervalos más largos de la diana. La densidad de sondas por exón diana es mayor para exones más largos que en los más cortos, por lo que se correlaciona con un patrón de cobertura sesgado hacia secuencias diana más largas. Sin embargo, se determinó que la respuesta de de captura de sonda no se correlacionó con la longitud del exón. Por ejemplo, la Figura 5 demuestra los datos de secuencia de un experimento utilizando un diseño de microchips de embaldosado estándar. Las distribuciones de respuesta de captura agregadas se muestran dentro de las dianas agrupadas por la longitud del exón diana. Las distribuciones no son significativamente diferentes entre regiones diana más cortas y más largas. La presente invención no se limita a un mecanismo particular. De hecho, la comprensión del mecanismo no es necesaria para practicar la presente invención. No obstante, se contempla que una falta de diferencias significativas entre las regiones diana cortas y más largas en la distribución de la captura de respuesta indica que una diferencia en la cobertura entre los dos grupos de dianas surge de la densidad de sondas no uniforme.

Los microchips experimentales se diseñaron para capturar secuencialmente regiones diana más amplias, sin embargo cada uno de los microchips representan aproximadamente el mismo número total de sondas y las regiones diana común a todos los chips se embaldosaron a densidad secuencialmente más baja. Al comparar los dianas individuales durante los experimentos, se observó que cada profundidad de secuenciación de cada diana fue linealmente dependiente de la densidad local de las sondas de captura y la pendiente de esta relación lineal en una diana particular se caracteriza por un sesgo hacia o en contra de la captura de esa diana (Figura 1c).

Basado en las observaciones, se contempla que la redistribución de las densidades de las sondas dentro de un solo chip podría utilizarse para mitigar el sesgo entre las dianas. Para ello, se utilizó un modelo lineal generalizado de la abundancia de diana relativa como una función de la densidad relativa de la sonda de captura después de lo cual se aplicó de optimización con restricciones para distribuir un número fijo de sondas de captura total para lograr una distribución de sondas predecibles para proporcionar una profundidad de cobertura uniforme. Esta optimización restringida comprende un algoritmo codicioso para asignar las sondas a las regiones con el fin de alcanzar la distribución de abundancias de diana deseada. El algoritmo tiene varias entradas: el modelo ajustado, la distribución

final deseada de las abundancias relativas de las dianas, y las densidades de sonda mínima y máxima permitidas en cualquier intervalo. La distribución de lectura de dianas se fija inicialmente en cero y se escala proporcionalmente de una manera gradual para llegar a la distribución final deseada.

En cada paso, el recuento de sonda necesaria para alcanzar la distribución de lectura de diana se calcula en cada intervalo sujeto a las restricciones de densidad máxima y mínima de la sonda. El algoritmo termina cuando se ha asignado el recuento completo de sondas disponibles. Se contempla que cualquier modelo relacionado con la abundancia de diana con la densidad de la sonda es modificable para practicar los métodos de la presente invención en el diseño de chips con sondas redistribuidas para la captura uniforme de secuencias diana. Como tal, la presente invención proporciona métodos y sistemas para la redistribución de sondas, caracterizado en que la frecuencia de cada secuencia individual de las sondas redistribuidas corresponde a la frecuencia de la secuencia de ácido nucleico diana correspondiente con la población de secuencias de moléculas de ácidos nucleicos diana. Una vez se determina el grado de respuesta de captura para cada sonda de oligonucleótido (por ejemplo, basado en la secuencia de la sonda y los cálculos como se definen aquí), y la abundancia de cada sonda de ácido nucleico, la práctica de los métodos de redistribución de la sonda de la presente invención, por lo tanto, corresponderá de forma recíproca a la respuesta de captura predeterminada de las secuencias diana (por ejemplo, más secuencias diana, menos sondas para esa región y viceversa).

Utilizando el modelo, los chips experimentales se diseñaron para lograr la captura sin sesgos y profundidad de secuenciación uniforme a lo largo de una región de aproximadamente 200 kb compartida entre los chips analizados. Por ejemplo, se sintetizaron dos chips de captura; uno que es un chip con una densidad de sonda aproximadamente uniforme (Figura 1d), y el segundo un chip con sondas redistribuidas (Figura 1e). La muestra de DNA se hibridó con los chips de captura, se eluyó, se amplificó, y se secuenció. Después de alinear las lecturas de los experimentos se compararon con los datos de cobertura del genoma de referencia. La estadística reveló un aumento significativo en la cobertura media a lo largo de las regiones diana (Tabla 1) después de la normalización de la variación en el número total de lecturas entre las secuenciaciones. Cuando se representó la profundidad de la cobertura a lo largo de las regiones diana (Figura 2), se reveló una mejora significativa en la uniformidad de la cobertura entre las regiones diana en los chips redistribuidos. Como tal, se demuestra aquí que los métodos de la presente invención proporcionan para los chips de captura una variedad de distribuciones de cobertura, tanto uniforme e intencionadamente no uniforme (por ejemplo, para enriquecer las dianas exónicas frente a regiones diana intrónicas / intergénicas).

El efecto de la práctica de los métodos de la presente invención para reequilibrar la distribución de la sonda en un soporte sólido para la hibridación de la secuencia diana se ejemplifica en la Figura 6. Se muestra una comparación de las desviaciones estándar de distribución de cobertura del locus diana entre cinco diseños diferentes de microchips reequilibrados (Rebal A hasta Rebal E) y un diseño basal de microchip de embaldosado estándar (HumanExon7Chip). Los datos demostraron una marcada reducción en la puntuación de la falta de uniformidad en los chips reequilibrados respecto al basal. Se eligió un conjunto de loci diana comunes en los seis diseños y se tomaron 140.000 lecturas al azar de los datos de la secuencia tras la captura con cada diseño. Se calculó la profundidad de cobertura en cada región y se representaron los datos calculados para cada diseño como una función de distribución acumulativa que indica el porcentaje de loci diana con una cobertura mayor o igual a un determinado nivel. Aunque se seleccionó un número igual de lecturas de cada captura, el diseño basal tenía significativamente más dianas con cobertura aberrantemente elevada y por consiguiente una alta proporción de las dianas con cobertura cero en comparación con los diseños reequilibrados (aproximadamente 80% frente a <20%).

Ciertas realizaciones ilustrativas de la invención se describen a continuación. La presente invención no se limita a estas realizaciones.

La presente invención permite la captura y el enriquecimiento de las moléculas de ácido nucleico diana o regiones genómicas diana a partir de una muestra biológica compleja mediante selección genómica directa. En algunas realizaciones, las formas de realización preferidas encuentran utilidad en la búsqueda de variantes y mutaciones genéticas, por ejemplo polimorfismos de un solo nucleótido (SNP), o conjunto de SNP, que subyacen en las enfermedades humanas. El descubrimiento de variantes genéticas y mutaciones permite, por ejemplo, el estudio y caracterización de enfermedades y otros trastornos genéticos, incluyendo la investigación en el diagnóstico y tratamientos terapéuticos de enfermedades y trastornos.

En algunas realizaciones, la presente descripción proporciona un soporte sólido, en el que el soporte sólido comprende sondas de oligonucleótidos inmovilizadas y en el que dichas sondas se distribuyen de tal manera que proporcionan la captura uniforme de moléculas de ácido nucleico diana enriquecidas. En algunas realizaciones, el soporte sólido es un portaobjetos de microchip, mientras que en otras realizaciones, el soporte sólido es una cuenta (por ejemplo, en solución en un tubo, en un pocillo de una placa, etc.) En algunas realizaciones, el soporte sólido comprende una cuenta sobre la cual se inmoviliza una sonda de oligonucleótido. La cuenta puede estar compuesta de cualquier tipo de material. Por ejemplo, cuentas útiles como soportes sólidos en los métodos de la presente

invención puede comprender, gel de sílice, vidrio, resina (por ejemplo, resina de Wang como se encuentra en la patente US 6.133.436), plástico metálico, celulosa, dextrano (por ejemplo, Sephadex ®), agarosa (por ejemplo, Sepharose ®), y similares. Las cuentas no están limitadas por el tamaño, sin embargo son preferibles las cuentas con un diámetro en el intervalo de aproximadamente 1 a aproximadamente 100 um.

En algunas realizaciones, la presente invención comprende la aplicación de una muestra de moléculas de ácido nucleico, por ejemplo, una muestra de DNA genómico, con el soporte sólido. En algunas realizaciones, la muestra se fragmenta antes de aplicarla al soporte sólido. En algunas realizaciones, las moléculas de ácido nucleico comprenden fragmentos enlazantes ligados a uno o ambos de los extremos del fragmento. En algunas realizaciones, los fragmentos se desnaturalizan para crear una sola cadena de ácido nucleico antes de aplicar dicha muestra al soporte sólido. En algunas realizaciones, la muestra de ácido nucleico desnaturalizada, fragmentada se aplica al soporte sólido bajo condiciones que permiten la hibridación de las secuencias diana en la muestra de ácido nucleico a la sonda de oligonucleótidos que comprende la secuencia diana asociada. En algunas realizaciones, el chip hibridado se lava para eliminar las moléculas de ácido nucleico no unidas y no unidas de forma específica. En algunas realizaciones, las secuencias diana capturadas y enriquecidas de manera uniforme se eluyen del soporte sólido y se realizan aplicaciones posteriores en las secuencias eluidas (Figura 3b).

En general, los microchips de oligonucleótidos están diseñados para dirigirse a una región o regiones de un genoma. En algunas realizaciones, las sondas se diseñan para ser sondas superpuestas, por ejemplo los nucleótidos de inicio de las sondas adyacentes están separados en el genoma por menos de la longitud de una sonda, o sondas que no se solapan, donde la distancia entre las sondas adyacentes son mayores que la longitud de una sonda. La superposición de las sondas es a menudo denominada "embaldosado" de sondas, creando de este modo un chip de embaldosado. En los chips de embaldosado, la distancia entre sondas adyacentes en general se superpone, variando el espacio entre el nucleótido de partida de dos sondas entre, por ejemplo, 1 y 100 bases. Se contempla que se analicen las sondas para su singularidad en el genoma. Por ejemplo, para evitar la unión no específica de elementos genómicos a los chips de captura, se excluyen elementos altamente repetitivos del genoma de los microchips de selección. El proceso comparó el conjunto de sondas contra un histograma de frecuencia pre-calculada de todas los sondas 15-mero posibles en el genoma humano. Para cada sonda, las frecuencias de los 15-meros que comprenden la sonda se utilizan entonces para calcular la frecuencia media de 15-mero de la sonda.

Las sondas inmovilizadas corresponden en secuencia a una o más regiones del genoma y se proporcionan, en una realización, sobre un soporte sólido en paralelo utilizando la tecnología de síntesis de chip sin máscara (MAS) como se ha descrito previamente. En algunas realizaciones, las sondas se obtuvieron en serie utilizando un sintetizador de DNA estándar y después se aplicaron a un soporte sólido. En algunas realizaciones, las sondas se obtuvieron de un organismo y se inmovilizaron en el soporte sólido. En realizaciones preferidas, las sondas inmovilizadas representan la redistribución de la sonda de tal manera que las sondas proporcionan la captura uniforme de secuencias diana. En otras realizaciones, las sondas inmovilizadas representan la redistribución de la sonda que no es uniforme y definida como tal por un investigador. La redistribución de la sonda se determina mediante la práctica de los métodos como se describe en la presente invención, por ejemplo como se demuestra en los ejemplos y figuras de este documento. Los ácidos nucleicos fragmentados se hibridan con las sondas inmovilizadas, y los ácidos nucleicos que no se hibridan, o que se hibridan de forma no específica a las sondas se separadas de las sondas unidas al soporte mediante un lavado. Las moléculas de ácidos nucleicos restantes que se hibridan específicamente con las sondas se eluyen del soporte sólido (por ejemplo, con agua caliente, mediante un tampón de elución de ácido nucleico que comprende por ejemplo tampón TRIS y / o EDTA) para producir una elución enriquecida de moléculas de ácido nucleico diana capturadas de forma uniforme.

En los métodos de la presente invención, la naturaleza y el rendimiento de las sondas seleccionadas se varían para normalizar ventajosamente y / o ajustar la distribución de las moléculas diana capturadas y enriquecidas. En algunas realizaciones, la normalización de la sonda proporciona un gen expresado por lectura. La normalización se puede aplicar, por ejemplo, a las poblaciones de moléculas de cDNA antes de la construcción de la librería, como la distribución de las moléculas en la población refleja los diferentes niveles de expresión de los genes expresados de las poblaciones de moléculas de cDNA que se producen. Por ejemplo, el número de reacciones de secuenciación necesarias para analizar de forma eficaz cada región diana se puede reducir mediante la normalización del número de copias de cada secuencia diana en la población enriquecida de tal manera que a lo largo del conjunto de sondas se normaliza el rendimiento de la captura de las diferentes sondas, en base a la combinación de capacidad y otros atributos de la sonda. La capacidad, que se caracteriza por una "métrica de captura", se determina ya sea informáticamente o empíricamente. Por ejemplo, la capacidad de unión de las moléculas diana se ajusta para proporcionar la denominada sondas de oligonucleótidos isotérmicas (Tm-equilibrado), como se describe en la solicitud de patente publicada Número US 2005/0282209, que permiten un rendimiento uniforme de la sonda, elimina los artefactos de hibridación y / o sesgos y proporcionar resultados de mayor calidad. Las longitudes de sonda se ajustan (normalmente, alrededor de 20 a aproximadamente 100 nucleótidos, preferiblemente de aproximadamente 40 a aproximadamente 85 nucleótidos, en particular entre 45 y aproximadamente 75 nucleótidos, opcionalmente, más de 100 nucleótidos hasta aproximadamente 250 nucleótidos) para igualar la temperatura de

fusión (por ejemplo, Tm = 76 ° C, normalmente de aproximadamente 55 ºC a aproximadamente 76 °C, en particular aproximadamente de 72 ºC a aproximadamente 76 °C) a través de todo el conjunto. En algunas realizaciones, las sondas están optimizadas para realizar de manera equivalente a una astringencia determinada en las regiones genómicas de interés, incluyendo regiones ricas en AT y en GC. En algunas realizaciones, la secuencia de las sondas individuales se ajusta, usando bases naturales o análogos sintéticos de base tales como inositol, o una combinación de los mismos para lograr una capacidad de captura deseada de estas sondas.

En algunas realizaciones, se utilizan sondas de ácidos nucleicos bloqueadas, sondas de ácidos nucleicos peptídicos y similares que tienen estructuras que proporcionan un rendimiento de captura deseado. Un experto en la materia apreciará que la longitud de la sonda, temperatura de fusión y la secuencia se pueden ajustar de forma coordinada para cualquier sonda dada para llegar a un rendimiento de captura deseada para la sonda. La temperatura de fusión (Tm) de la sonda se puede calcular utilizando, por ejemplo, la fórmula: Tm= 5x (Gn+ Cn) + lx (An + Tn), donde n es el número de cada base específica (A, T , G o C) presente en la sonda.

En algunas realizaciones, la eficiencia de la captura se normalizó determinando la idoneidad de la captura de las sondas en el conjunto de sondas y ajustando la cantidad de cada sonda individual en el soporte sólido de acuerdo con ello. Por ejemplo, si una primera sonda captura veinte veces más ácido nucleico que una segunda sonda, entonces la eficiencia de la captura de ambas sondas puede igualarse proporcionando veinte veces más de copias de la segunda sonda, por ejemplo aumentando en veinte veces el número de puntos que contienen la segunda sonda. Si las sondas se preparan de forma seriada y se aplican en el soporte sólido, la concentración de cada sonda individual en el conjunto también puede hacerse variar del mismo modo. Además, otra estrategia para normalizar la captura de ácidos nucleicos diana es someter a las moléculas diana eluidas a una segunda ronda de hibridación con las sondas bajo condiciones menos astringentes que las utilizadas para la primera ronda de hibridación. Aparte de un enriquecimiento sustancial en la primera hibridación, que reduce la complejidad relativa del ácido nucleico genómico original, la segunda hibridación se realiza bajo condiciones de hibridación que saturan todas las sondas de captura. Se contempla que al proporcionarse cantidades idénticas de sondas de captura sobre el soporte sólido, la saturación de las sondas asegurará que se eluyen cantidades sustancialmente iguales de cada diana tras la segunda hibridación y lavado.

Otra estrategia de normalización se realiza tras la elución y amplificación de las moléculas diana capturadas en el soporte sólido. Las moléculas diana en el material eluido se desnaturalizan utilizando, por ejemplo, un proceso de desnaturalización química o térmica, hasta un estado de cadena sencilla y se rehibridan. Las consideraciones cinéticas indican que las especies más abundantes rehibridarán antes que las especies menos abundantes. Por lo tanto, eliminando la fracción inicial de especies rehibridadas, las especies de cadena sencilla restantes se equilibran en el eluido en relación a la población inicial. El tiempo necesario para una eliminación óptima de las especies abundantes se determinó de forma empírica. Las moléculas de ácido nucleico desnaturalizadas y fragmentadas que se proporcionan comprenden un tamaño promedio de alrededor de 100 a alrededor de 1000 residuos nucleotídicos, preferiblemente de alrededor de 250 a alrededor de 800 residuos nucleotídicos y más preferiblemente de alrededor de 400 a alrededor de 600 residuos nucleotídicos (por ejemplo, mediante la nebulización de DNA genómico como se indica en la solicitud de patente europea PE 0 552 290.

Los parámetros de reducción de complejidad genética pueden escogerse de forma casi arbitraria, dependiendo de la necesidad del usuario de seleccionar secuencias, y vienen definidos por las secuencias de las sondas oligonucleotídicas. En algunas realizaciones, dichas sondas definen una pluralidad de exones, intrones o secuencias regulatorias de una pluralidad de locus genéticos.

En algunas realizaciones, dichas sondas definen la secuencia completa de al menos un único locus genético, y dicho locus posee un tamaño de al menos 100 kb y preferiblemente al menos 1 Mb o un tamaño como el especificado anteriormente. En algunas realizaciones, dichas sondas definen lugares que se conoce contienen SNP. En algunas realizaciones, las sondas definen un chip de embaldosado (del inglés *tiling array*). Dicho chip de embaldosado, en el contexto de la presente invención, se contempla que está diseñado para capturar la secuencia completa de al menos un cromosoma completo de manera uniforme.

En algunas realizaciones, la población de sondas comprende al menos una segunda sonda para cada secuencia diana que se ha de enriquecer, que se caracteriza por que dicha segunda sonda posee una secuencia que es complementaria de dicha primera secuencia. El soporte sólido de acuerdo con la presente invención es un microchip de ácidos nucleicos o una población de cuentas. Las cuentas comprenden, por ejemplo, cuentas de vidrio, metal, cerámica o cuentas poliméricas. Si el soporte sólido es un microchip, es posible sintetizar las sondas de captura de oligonucleótidos *in situ* directamente sobre dicho soporte sólido. Por ejemplo, las sondas pueden sintetizarse sobre el microchip utilizando un sintetizador de chips sin máscara (US 6.375.903). Las longitudes de las sondas de oligonucleótidos pueden variar, dependen del diseño experimental y sólo están limitadas por las posibilidades de síntesis de tales sondas. Preferiblemente, la longitude promedio de la población de sondas es de alrededor de 20 a alrededor de 100 nucleótidos, preferiblemente de alrededor de 40 a alrededor de 85 nucleótidos, en particular de

alrededor de 45 a alrededor de 75 nucleótidos. Si el soporte sólido es una población de cuentas, las sondas de captura se sintetizan inicialmente sobre un microchip utilizando un sintetizador de chips sin máscara, luego se liberan o se escinden de acuerdo con la metodología estándar conocida, opcionalmente se amplifican y luego se inmovilizan sobre dicha población de cuentas de acuerdo con los métodos conocidos en la materia. En algunas realizaciones, las cuentas están empaquetadas en una columna, de forma que la muestra se aplica y se hace pasar a través de la columna para reducir su complejidad genética. En algunas realizaciones, la hibridación tiene lugar en una suspensión acuosa que comprende las cuentas con las múltiples moléculas de oligonucleótido inmovilizadas.

En una realización, cada sonda oligonucleotídica contiene un grupo químico o enlazante, por ejemplo una porción que permite su inmovilización sobre un soporte sólido (por ejemplo, un grupo inmovilizable). Por ejemplo, la biotina se utiliza para la inmovilización sobre soportes sólidos recubiertos de estreptavidina. En otra realización, tal porción es un hapteno como la digoxigenina, que se utiliza para la inmovilización sobre un soporte sólido recubierto con un anticuerpo que reconoce el hapteno (por ejemplo un anticuerpo de unión a digoxigenina).

En algunas realizaciones, las sondas de ácido nucleico para las moléculas de ácido nucleico diana se sintetizan sobre un soporte sólido, se libera del soporte sólido como un conjunto de sondas y se amplifican mediante, por ejemplo, una PCR. En algunas realizaciones, un conjunto amplificado de sondas liberadas se inmoviliza de forma covalente o no covalente sobre el soporte, como cuentas de vidrio, metal, cerámica o cuentas poliméricas, u otro soporte sólido. En algunas realizaciones, las sondas se diseñan para una liberación sencilla del soporte sólido al proporcionar, por ejemplo, en el extremo o cerca del extremo de la sonda próximo al soporte, una secuencia de ácido nucleico lábil en medio ácido o alcalino que libera las sondas bajo condiciones de pH reducido o elevado, respectivamente, o mediante la incorporación en el extremo de la sonda de una diana de escisión de una endonucleasa de restricción, u otras dianas de escisión enzimática. Se conocen en la material varias químicas de enlazante escindible. En algunas realizaciones, el soporte sólido se proporciona en una columna con un punto de entrada y uno de salida de fluido. En algunas realizaciones, se incorpora un nucleótido biotinilado en la secuencia de la sonda y el soporte se recubre con estreptavidina para la captura de la sonda biotinilada.

La presente invención comprende la captura de secuencias de ácido nucleico diana que se encuentran en moléculas de ácido nucleico diana. Las moléculas de ácido nucleico diana incluyen ácidos nucleicos de cualquier origen, en forma purificada, sustancialmente purificada o no purificada. En algunas realizaciones, no es necesario que el material de origen de ácido nucleico comprenda un complemento completo de moléculas de ácido nucleico genómico de un organismo. En algunas realizaciones, la muestra de ácido nucleico es biológica. En algunas realizaciones, las muestras biológicas de ácido nucleico se obtienen a partir de animales y comprenden ácidos nucleicos aislados de fluidos, sólidos, tejidos, etc. En algunas realizaciones, las muestras biológicas de ácido nucleico también pueden proceder de animales no humanos, lo que incluye pero no se limita a, vertebrados como los roedores, primates no humanos, ovinos, bovinos, rumiantes, lagomorfos, porcinos, caprinos, equinos, caninos, felinos, aves, etc. En algunas realizaciones, los ácidos nucleicos biológicos también puede obtenerse a partir de plantas, procariotas (por ejemplo, bacterias) y virus (por ejemplo, DNA o RNA). Sin embargo, se contempla que la presente invención no está limitada por el origen de la muestra de ácidos nucleicos, y cualquier ácido nucleico de cualquier reino biológico puede encontrar utilidad en los métodos aquí descritos. En las realizaciones preferibles, las muestras de ácido nucleico son de humano, o se derivan de humanos, por ejemplo de pacientes individuales, muestras de tejido o cultivos celulares. Como se utiliza aquí, el término "moléculas de ácido nucleico diana" se refiere a las moléculas de una región genómica diana a estudiar. Las sondas preseleccionadas determinan el rango de las moléculas de ácido nucleico diana. Un experto con la ayuda de esta descripción apreciará el rango completo de posibles dianas y dianas asociadas.

Las moléculas de ácido nucleico de la presente descripción son normalmente ácidos desoxiribonucleicos o ácidos ribonucleicos, e incluyen los productos sintetizados *in vitro* por conversión de una molécula de ácido nucleico (por ejemplo DNA, RNA y cDNA) en otra, así como las moléculas sintéticas que contienen análogos de nucleótidos. En las realizaciones preferibles, las moléculas de ácido nucleico son moléculas de DNA, preferiblemente moléculas de DNA genómico. En algunas realizaciones, las moléculas de ácido nucleico están fragmentadas. En algunas realizaciones, las moléculas de ácido nucleico están desnaturalizadas. En algunas realizaciones, las moléculas de DNA desnaturalizadas, preferiblemente moléculas derivadas del genoma, son más cortas que las moléculas de ácido nucleico genómicas de aparición en la naturaleza, lo que comprende, por ejemplo, moléculas fragmentadas de ácido nucleico.

Una secuencia o región diana de la presente descripción comprende uno o más bloques continuos de varias megabases (Mb), o varias regiones menores contiguas o no contiguas, como todos los exones de uno o más cromosomas, o lugares que se conoce contienen SNP. Por ejemplo, el soporte sólido puede soportar un chip de embaldosado diseñado para capturar uno o más cromosomas completos, partes de uno o más cromosomas, todos los exones, todos los exones de uno o más cromosomas, exones seleccionados, intrones y exones de uno o más genes, regiones reguladoras génicas y otros similares. En algunas realizaciones, para aumentar la probabilidad de que las dianas deseadas no únicas o difíciles de capturar se enriquezcan, las sondas pueden ir dirigidas a

secuencias asociadas con la secuencia diana deseada (por ejemplo, en el mismo fragmento pero en un punto separado), en la que fragmentos genómicos en cuestión que contienen tanto la diana deseada como las secuencias asociadas serán capturadas y enriquecidas. Las secuencias asociadas pueden estar adyacentes o separadas de las secuencias diana, pero un experto apreciará que cuanto más cerca estén las dos porciones entre ellas, más probable será que los fragmentos genómicos contengan ambas porciones. En algunas realizaciones, para reducir más el impacto limitado de la hibridación cruzada con moléculas no relacionadas con la diana, y potenciando así la integridad del enriquecimiento, se realizan rondas secuenciales de captura utilizando conjuntos de sondas de captura distintos pero relacionados dirigidos contra la región diana. Las sondas relacionadas son las sondas correspondientes a las regiones de gran proximidad entre ellas en el genoma y que hibridan con el mismo fragmento de DNA genómico.

En algunas realizaciones, los métodos de enriquecimiento uniforme de la presente invención comprenden fragmentos de moléculas de ácido nucleico, por ejemplo fragmentos de DNA genómico, en un rango de tamaño compatible con la tecnología post-enriquecimiento uniforme en la que se utilizarán los fragmentos enriquecidos. En algunas realizaciones, el tamaño de los fragmentos comprende de aproximadamente 100 nucleótidos a aproximadamente 1000 residuos nucleotídicos o pares de bases, de aproximadamente 250 a aproximadamente 800 residuos nucleotídicos, de aproximadamente 400 a aproximadamente 600 residuos nucleotídicos, y más preferiblemente, de aproximadamente 500 residuos nucleotídicos o pares de bases.

Un experto puede producir moléculas de ácido nucleico fragmentadas de tamaño aleatorio o no aleatorio a partir de moléculas de mayor tamaño mediante, por ejemplo, la fragmentación o escisión química, física o enzimática utilizando protocolos bien conocidos. La fragmentación química puede realizarse utilizando, por ejemplo, metales ferrosos (por ejemplo, Fe-EDTA). Los métodos físicos incluyen, por ejemplo, la sonicación, fuerza hidrodinámica o la nebulización (por ejemplo, como en la solicitud de patente europea PE 0 552 290) y otras fuerzas de cizalla. Los protocolos enzimáticos pueden utilizar, por ejemplo, nucleasas como la nucleasa micrococal (Mnasa) y la exonucleasa (como ExoI o Ba131) o endonucleasas de restricción. La presente invención no está limitada por el método utilizado para producir las moléculas de ácido nucleico fragmentadas, como el DNA genómico fragmentado, de hecho se contempla el uso de cualquier de método fragmentación para proporcionar las moléculas de ácido nucleico fragmentadas para poner en práctica la presente invención.

En algunas realizaciones, la presente invención proporciona métodos para reducir la complejidad genómica y determinar múltiples secuencias al incorporar el paso de ligar moléculas adaptadoras en uno o ambos extremos de las moléculas de ácido nucleico fragmentadas. En las realizaciones preferibles, los adaptadores están ligados a ambos extremos de las moléculas de ácido nucleico fragmentadas. En algunas realizaciones, las moléculas adaptadoras de la presente invención comprenden oligonucleótidos de doble cadena y de extremo romo. En algunas realizaciones, cuando se ligan los adaptadores a las moléculas de ácido nucleico fragmentadas proporcionan puntos de amplificación de dichas moléculas de ácido nucleico, con al menos un cebador, y dicho cebador comprende una secuencia que corresponde o que hibrida específicamente bajo condiciones de hibridación con la secuencia de dichas moléculas adaptadoras. En algunas realizaciones, los enlazantes oscilaron de aproximadamente 12 a aproximadamente 100 pares de bases, de aproximadamente 18 a aproximadamente 80 pares de bases, preferiblemente de aproximadamente 20 a aproximadamente 24 pares de bases.

Cuando se ligan cebadores de extremo romo a ácidos nucleicos fragmentados, se contempla que los ácidos nucleicos fragmentados son en sí mismos de extremo romo. El relleno de los extremos de las moléculas de ácido nucleico para crear moléculas de extremo romo, previamente a la ligación con otras moléculas, como las moléculas adaptadoras, es bien conocido en la materia, por ejemplo utilizando métodos que comprenden el uso de dNTP y polimerasas de DNA como la polimerasa de DNA T4 o la Klenow. Los extremos 5' pulidos de las moléculas de ácido nucleico fragmentadas se fosforilan entonces utilizando, por ejemplo, la quinasa de polinucleótidos T4, que añade grupos fosfato a los extremos 5', lo que permite la subsiguiente ligación de las moléculas adaptadoras. La ligación de las moléculas adaptadoras se realiza de acuerdo con cualquier método conocido en la materia, por ejemplo, realizando una reacción de ligación que comprende la ligasa de DNA T4.

En algunas realizaciones, la ligación de adaptadores a las moléculas de ácido nucleico fragmentadas se realiza previamente a la hibridación con las sondas de oligonucleótido, mientras en otras realizaciones se realiza tras la hibridación con las sondas de oligonucleótido. En las realizaciones en las que la ligación se realiza a continuación, es preferible que los ácidos nucleicos enriquecidos que se liberan del soporte sólido en forma de cadena sencilla se rehibriden y seguidamente se realice una reacción de extensión del cebador y una reacción de relleno de acuerdo con los métodos estándar conocidos en la materia.

En algunas realizaciones, la ligación de moléculas adaptadoras permite un paso de subsiguiente amplificación de las moléculas capturadas. En algunas realizaciones, las moléculas adaptadoras comprenden una secuencia, que resulta en una población de fragmentos con secuencias terminales idénticas en ambos extremos del fragmento. Como tal, sería suficiente utilizar un único cebador en un potencial paso de amplificación realizado a continuación. En algunas

realizaciones, las moléculas adaptadoras comprenden dos secuencias diferentes, por ejemplo una secuencia A y una secuencia B. Así, puede resultar una población de moléculas enriquecidas compuestas por tres secuencias diferentes en los extremos de los ácidos nucleicos fragmentados: (i) fragmentos con un adaptador (A) en un extremo y otro adaptador (B) en el otro extremo, (ii) fragmentos con adaptadores A en ambos extremos, y (iii) fragmentos con adaptadores B en ambos extremos. La generación de moléculas enriquecidas de acuerdo con el tipo (i) es ventajosa si se realiza una amplificación y secuenciación, por ejemplo, utilizando el instrumento 454 Life Sciences Corporation GS20 y GSFLX (manual de prep. de bibliotecas GS20, Dic. 2006, número de publicación de la patente PCT WO 2004/070007).

En algunas realizaciones, si uno de tales adaptadores, por ejemplo el adaptador B, comprende una modificación con biotina, las moléculas (i) y (iii) pueden entonces capturarse sobre partículas magnéticas recubiertas de estreptavidina (SA) para su posterior aislamiento, y los productos de (ii) se eliminan con un lavado. En el caso de que el DNA enriquecido e inmovilizado en SA sea de cadena sencilla tras la elución del chip de captura/ soporte sólido, es ventajoso pasar el DNA a una doble cadena. En este caso, los cebadores complementarios al adaptador A pueden añadirse a los productos que se han depositado sobre SA lavados. Como las porciones que son B-B (iii anteriormente) no poseen A o su complementaria disponible, sólo los productos con adaptadores A-B y capturados por la SA pasan a ser de doble cadena tras la extensión del cebador a partir de un cebador complementario de A. A continuación, las moléculas de DNA de doble cadena unidas a dichas partículas magnéticas se desnaturalizan de forma termal o química (por ejemplo con NaOH) de modo que la cadena que se sintetiza de nuevo se libera a la solución. Debido al fuerte enlace entre biotina y estreptavidina, las moléculas con dos adaptadores B no se liberarán a la solución. La única cadena disponible para su liberación es la cadena sintetizada por extensión del cebador complementaria de A a complementaria de B. Dicha solución que comprende las moléculas diana de cadena sencilla con un adaptador A en un extremo y un adaptador B en el otro extremo, puede por ejemplo, unirse a continuación a otro tipo de cuenta que comprende una secuencia de captura que sea suficientemente complementaria de las secuencias adaptadoras A o B para su posterior procesado.

En algunas realizaciones, la presente invención no está limitada a un conjunto particular de condiciones de hibridación. Sin embargo, preferiblemente se utilizarán condiciones de hibridación astringentes, como conocerán los expertos en la materia y como se describen aquí. En algunas realizaciones, la presente invención proporciona un lavado de la reacción de hibridación, eliminando así las moléculas de ácido nucleico no unidas y las unidas de forma no específica. En algunas realizaciones, la presente invención proporciona lavados de astringencia diferencial, por ejemplo, un tampón de lavado I que comprende 0,2x SSC, SDS al 0,2% (v/v) y DTT 0,1 mM, un tampón de lavado II que comprende 0,2x SSC y DTT 0,1 mM y un tampón de lavado III que comprende 0,5x SSC y DTT 0,1 mM. La presente invención no se ve limitada por la composición de los tampones de hibridación y/o de lavado, de hecho puede utilizarse cualquier composición en la práctica de los métodos de la presente invención. En algunas realizaciones, las secuencias diana de hibridación se eluyen del soporte sólido utilizando, por ejemplo agua o una solución con pocos solutos similar conocida para los expertos en la materia.

En algunas realizaciones, la presente invención proporciona un enriquecimiento uniforme de las secuencias de ácido nucleico diana para su posterior uso en los métodos de secuenciación basados en chips con secuencias diana, aleatoria (del inglés *shotgun*), capilar u otros métodos conocidos en la materia. En general, las estrategias para la secuenciación aleatoria de fragmentos generados al azar son rentables y pueden integrarse fácilmente en un proyecto, pero la invención mejora la eficiencia de la estrategia aleatoria al presentar fragmentos de ácido nucleico enriquecidos de forma uniforme de una o más regiones genómicas de interés para la secuenciación. Así, la presente invención proporciona la capacidad de centrar las estrategias de secuenciación en regiones genómicas específicas, como cromosomas o exones individuales (por ejemplo, mediante la selección consciente no uniforme mediante una distribución de las sondas no uniformes), por ejemplo, con el propósito de una secuenciación con finalidad clínica.

Como conocerá un experto, la secuenciación mediante la síntesis se entiende como un método de secuenciación que controla la generación de productos colaterales tras la incorporación de un desoxinucleósido trifosfato específico durante la reacción de secuenciación (Rhonaghi, M., et al., Science 281 (1998) 363-65). Por ejemplo, una de las realizaciones más prominentes de secuenciación mediante una reacción de síntesis es el método de secuenciación del pirofosfato. En la pirosecuenciación, la generación de pirofosfato durante la incorporación de nucleótidos se monitoriza mediante una cascada enzimática que resulta en la generación de una señal quimioluminiscente. El sistema 454 Genome Sequencer System (Roche Applied Science, nº de catálogo 04760085001) se basa en la tecnología de secuenciación de pirofosfato. Para la secuenciación en un instrumento 454 GS20 o 454 FLX, el tamaño promedio de los fragmentos de DNA genómico está preferiblemente en el rango de 200 o 600 pb, respectivamente. La secuenciación mediante reacciones de síntesis también puede comprender una reacción de secuenciación de tipo colorante terminador. En este caso, los bloques de construcción dNTP incorporados comprenden una señal detectable, como una señal fluorescente, que evita la posterior extensión de la cadena de DNA naciente. La señal se elimina y se detecta tras la incorporación del bloque de construcción de dNTP al híbrido entre molde / extensión del cebador, por ejemplo, utilizando una polimerasa de DNA que realiza actividad exonucleasa de 3'-5' o de corrección de galeradas.

En algunas realizaciones, las secuencias diana enriquecidas de manera uniforme se eluyen del microchip y se secuencian. En algunas realizaciones, la secuenciación se realiza utilizando un secuenciador 454 Life Sciences Corporation. En algunas realizaciones, la presente invención proporciona una amplificación de la secuencia diana tras la elución mediante una PCR por emulsión (emPCR) siguiendo los protocolos proporcionados por el fabricante. Las cuentas que comprenden los ácidos nucleicos diana amplificados de forma clónica mediante la emPCR se transfieren a una placa picotitulada de acuerdo con los protocolos proporcionados por el fabricante y se someten a una reacción de secuenciación de pirofosfato para la determinación de su secuencia.

En algunas realizaciones, el análisis de los datos se realiza sobre las secuencias diana unidas previamente, o en lugar de, su elución. El análisis de los datos se realiza, por ejemplo, para determinar la redistribución de la sonda necesaria y para verificar la redistribución de la sonda una vez se ha completado. El análisis de los datos se realiza utilizando cualquier escáner de chips, por ejemplo un escáner fluorescente Axon GenePix 4000B. Una vez se han capturado los datos mediante el escáner, se utilizan programas de bioinformática para analizar los datos capturados. Los programas de bioinformática útiles en el análisis de datos a partir de formatos de microchips fluorescentes incluyen, pero no se limitan a SignalMap™ (NimbleGen) y NimbleScan™ (NimbleGen), aunque pueden utilizarse del mismo modo cualquier escáner y programa de bioinformática capaz de capturar y analizar los datos generados por los métodos de la presente invención. Los datos obtenidos pueden leerse, por ejemplo, en cualquier pantalla de computador u otro dispositivo capaz de mostrar los datos como se muestra en la Figura 1.

Puede proporcionarse un equipo que comprende reactivos y/o otros componentes (por ejemplo, tampones, instrucciones, superficies sólidas, contenedores, programas, etc.) suficientes y necesarios para realizar un enriquecimiento uniforme (o enriquecimiento no uniforme) de moléculas de ácido nucleico diana. Pueden proporcionarse al usuario los equipos en uno o más contenedores (que además comprendan uno o más tubos, paquetes, etc) que pueden requerir un almacenaje diferente, por ejemplo un almacenaje diferente de los componentes/ reactivos del equipo debido a requisitos de luz, temperatura, etc. particulares para cada componente/ reactivo del equipo. En algunas realizaciones, un equipo comprende una o más moléculas adaptadoras de doble cadena, y asimismo los adaptadores comprenden una o más secuencias. En algunas realizaciones, un equipo comprende uno o más soportes sólidos, en los que dichos soportes sólidos pueden ser un microchip o una pluralidad de cuentas como las aquí descritas. En algunas realizaciones, el equipo comprende al menos uno o más compuestos y reactivos para realizar reacciones enzimáticas, por ejemplo una o más de entre una polimerasa de DNA, una quinasa de polinucleótidos T4, una ligasa de DNA T4, una solución de hibridación de chips, una solución de lavado de chips y similares. En algunas realizaciones, se proporcionan una o más soluciones de lavado en un equipo, y dichas soluciones de lavado comprenden SSC, DTT y opcionalmente SDS. En algunas realizaciones, un equipo comprende uno o más tampones de lavado, cuyos ejemplos incluyen, pero no se limitan al tampón de lavado I (0,2 x SSC, SDS al 0,2% (v/v), DTT 0,1 mM) y/o tampón de lavado II (0,2 x SSC, DTT 0,1 nM) y/o tampón de lavado III (0,5 x SSC, DTT 0,1 mM). En algunas realizaciones, un equipo comprende una solución de elución de chips, en la que dicha solución de elución comprende agua purificada y/o una solución que contiene tampón TRIS y/o EDTA. En algunas realizaciones, un equipo comprende una segunda molécula adaptadora, en la que una cadena de oligonucleótido de dicha primera o segunda molécula adaptadora comprende una modificación que permite la inmovilización en un soporte sólido. Por ejemplo, tal modificación puede ser un marcaje de biotina que puede utilizarse para la inmovilización sobre un soporte sólido recubierto de estreptavidina. Alternativamente, dicha modificación puede ser un hapteno como la digoxigenina, que puede utilizarse para la inmovilización sobre un soporte sólido recubierto con un anticuerpo que reconoce un hapteno.

Los siguientes ejemplos se proporcionan para demostrar e ilustrar en más detalle ciertas realizaciones y aspectos de la presente invención preferibles y no deben tomarse como limitantes del alcance de la misma.

Si se proporciona un rango de valores, debe entenderse que cada valor intermedio, hasta un décimo de la unidad del límite inferior a no ser que el contexto claramente lo indique de otro modo, entre el límite superior e inferior de ese rango también se incluye de forma específica. Cada rango menor entre cualquier valor indicado o valor intermedio en un rango indicado y cualquier otro valor indicado o intermedio en ese rango indicado está incluido en la invención. Los límites superiores e inferiores de estos rangos menores pueden estar incluidos o excluidos de forma independiente en el rango, y cada rango en el que alguno, ninguno o ambos límites están incluidos en los rangos menores, también se encuentra incluido en la invención, sujeto a cualquier límite excluido específicamente en el rango indicado. Si el rango indicado incluye uno o ambos límites, los rangos que excluyen alguno o ambos de estos límites incluidos también se incluyen en la invención.

Ejemplo 1. Diseño del chip de captura inicial

Se diseñaron cinco microchips de captura de secuencias que tenían como diana regiones anidadas de extensión decreciente (5 Mpb, 2 Mpb, 1 Mpb, 500 Kpb y 200 Kpb), cada una de ellas centrada aproximadamente en la coordenada chr17:38490539. Se creó una base de datos común de secuencias de sondas con una longitud mediana

de 75 pb y capaces de producir síntesis en no más de 188 ciclos (NimbleGen, Madison WI). Cada diseño de captura se componía de no más de 385.000 sondas seleccionadas de esta base de datos con un espaciado de coordenadas de las sondas lo más cercano posible dentro del respectivo intervalo genómico. Como la capacidad del chip superó el número de sondas únicas en el intervalo diana en los diseños de 200 Kpb y 500 Kpb, cada sonda se replicó ocho y cuatro veces, respectivamente, en esos chips.

Ejemplo 2. Preparación de la muestra y captura del microchip

Se compró DNA genómico purificado (línea celular de linfoma de Burkitt, ATCC #NA04671) del Coriell Institute for Medical Research (Coriell Cell Repositories, Camden NJ) y se amplificó utilizando un equipo de amplificación del genoma completo de Qiagen (Hilden, Alemania). Tras la amplificación, se sonicaron 20 ug de DNA, dando lugar a fragmentos de un tamaño promedio de 500 pb. Los fragmentos se trataron con el fragmento Klenow de la polimerasa de DNA I (New England Biolabs, Beverly MA) que genera extremos romos y luego se fosforiló en 5' con la quinasa de polinucleótidos (New England Biolabs) siguiendo los protocolos establecidos. Se hibridaron y ligaron a los extremos del DNA genómico fragmentado los enlazantes oligonucleótidos sintéticos 5'-Pi-GAGGATCCAGAATTCTCGAGTT-3' (Id. de Sec. Nº1) y 5'-CTCGAGAATTCTGGATCCTC- 3' (Id. de Sec. Nº2). Los fragmentos de DNA genómico adaptados con enlazantes se hibridaron en microchips de captura en presencia de tampón de hibridación 1 x de NimbleGen (NimbleGen) durante aproximadamente 65 horas a 42 °C con mezclado activo utilizando una estación de hibridación MAUI (NimbleGen), siguiendo los protocolos del fabricante. Tras la hibridación, los chips se lavaron 3 veces, con lavados de 5 minutos cada uno, con el tampón de lavado astringente (NimbleGen), seguido de un enjuague con los tampones de lavado 1, 2 y 3 (NimbleGen), siguiendo los protocolos del fabricante de la guía del usuario de chips NimbleChip™ para el análisis de CGH. Los fragmentos de DNA capturados se eluyeron inmediatamente con 2 x 250 ml de agua a 95 °C. Las muestras se secaron, resuspendieron y se amplificaron mediante una reacción en cadena de la polimerasa mediada por ligación (LM-PCR) utilizando cebadores complementarios a los enlazantes ligados.

Ejemplo 3-Secuenciación y procesamiento de datos de secuencias

Los enlazantes compatibles con la secuenciación 454 (454, Branford CT) se ligaron a los, fragmentos capturados de DNA eluidos. Los fragmentos resultantes se amplificaron en cuentas usando PCR por emulsión (emPCR) y se secuenció utilizando el instrumento de secuenciación 454, siguiendo los protocolos del fabricante. Como cada fragmento secuenciado contenía el enlazante de 20 pb para la LM-PCR, la mayoría de las lecturas de secuenciación 454 comprende esta secuencia enlazante.

Se aplicaron filtrados de calidad estándar y funciones de llamada de bases del instrumento 454 para proporcionar lecturas de secuencia y las correspondientes puntuaciones de calidad. Las secuencias adaptadoras y el cebador de secuenciación se eliminaron de las lecturas de secuencia. Antes de las lecturas de mapeado del ensamblaje hgl8 del genoma humano, se enmascararon las porciones repetitivas de cada lectura con probabilidad de mapear localizaciones no exclusivas (por ejemplo, alinear con alta identidad con localizaciones múltiples y dispares en el genoma) utilizando WindowMasker (Morgulis, A., et al ., Bioinformatics 22 (2) (2006) 134-41). Las lecturas se mapearon en el genoma usando NCBI Megablast (Zhang, Z., et al., J Comput Biol. 7 (1/2) (2000) 203-14). Después de descartar coincidencias con el genoma de menos de 95% de identidad, las lecturas restantes se clasificaron como mapeo único si, para cada lectura, se produjo o bien 1) una sola coincidencia con el genoma, o 2) solo se pudo identificar claramente una mejor coincidencia. En este último caso, solo se seleccionó un emparejamiento entre varios, si este único emparejamiento tenía tanto la mayor longitud y la más fuerte homología. De lo contrario, la lectura se etiquetó como mapeado no único. Todos los análisis posteriores se limitaron a lecturas de mapeado únicas.

Ejemplo 4- análisis de captura de datos

La densidad de sondas de captura de un intervalo genómico determinado se calculó promediando sobre cada base en el intervalo el número de sondas de captura que se solapan con esa base. Del mismo modo, la profundidad de lectura bruta en un intervalo genómico determinado se calculó promediando sobre cada base en el intervalo el número de secuencias de lectura de mapeado únicas que se solapan con esa base. Las profundidades de lectura entre distintos experimentos de captura / secuenciación se normalizó dividiendo por el número total de bases secuenciadas mapeadas únicamente en las regiones diana. El intervalo central de 200kb cubierto por los cinco experimentos de captura inicial se segmentó en ventanas de 100 pb que no se solapaban y se calcularon las profundidades de lectura y densidad de sonda de captura en cada ventana (Figura 1a). Dentro de cada ventana, se utilizó la regresión lineal para ajustar las profundidades de lectura para capturar la densidad de la sonda, con la intersección limitada a (0,0), la pendiente resultante (o "respuesta de captura") en cada ventana cuantificó la afinidad de captura local en esa región (Figura 1b).

Ejemplo 5- redistribución de sondas

Las sondas de captura en los cinco diseños iniciales de captura se colocaron aproximadamente de forma uniforme dentro del intervalo de diana correspondiente (figura 1c). Asimismo, se preparó un diseño de control con una distribución uniforme de la sonda de captura a través de la diana central de 200 Kpb (Figura 1d). Se preparó un diseño optimizado (Figura 1e) moviendo las sondas de las regiones de alta respuesta de captura que requieren menos sondas a las regiones que necesitan más sondas para cubrir la profundidad de lectura uniforme deseada.

Un modelo de regresión lineal con ajuste empírico, a partir de la primera serie de experimentos de captura se utilizó para predecir la profundidad de lectura resultante de una densidad determinada de sondas de captura en cada región diana, permitiendo de esta manera la mejor distribución de las sondas diana para producir de manera óptima una profundidad de lectura uniforme tras la captura y secuenciación.

La Tabla 1 muestra ejemplos de las estadísticas de cobertura de secuencia para los chips control y de captura redistribuidos. Después de corregir los datos de variación experimental tras la captura de secuencia (por ejemplo, mediana de cobertura dividida por el número de lecturas de dianas), el chip redistribuido demuestra una mejora de aproximadamente un 20% en la uniformidad de captura sobre el chip de control (última columna).

Tabla 1

	Control	Redistribuido
Lecturas totales	342796	383658
pb totales	7,15 E+07	7,52 E+07
Hibridaciones totales	465564	532545
Por debajo del umbral del 95%	27,14% (126333)	33,22% (176928)
Sin hibridación	24,93% (85453)	38,82% (148950)
Mapado de forma única	73,60% (252284)	59,60% (228658)
pb mapadas de forma única	78,13% (55873369)	65,39% (49201425)
Bases diana cubiertas	146295	151732
% bases diana cubiertas	95,00%	98,50%
Nº lecturas en la región diana	16,11% (40649)	32,76% (74905)
Promedio de la cobertura	57,3	102,8
Mediana de la cobertura	42	93
Uniformidad de la captura	100,00%	120,16%

Listado de secuencias

<110> Roche Diagnostics GmbH F. Hoffmann-La Roche AG

<120> Métodos y sistemas para el enriquecimiento uniforme de regiones genómicas

<130> 25831 WO

<150> US 61/032594

<151> 2008-02-29

<160> 2

<170> PatentIn version 3.2

<210> 1

<211> 22

<212> DNA

<213> Artificial

<220>

<223> enlazante de oligonucleótidos

<400> 1

gaggatccag aattctcgag tt 22

<210> 2

<211> 20

<212> DNA

<213> Artificial

<220>

<223> enlazante de oligonucleótidos

<400> 2

ctcgagaatt ctggatcctc 20

REIVINDICACIONES

1. Un método para el enriquecimiento uniforme de una población de moléculas de ácido nucleico en una muestra, que comprende:

a) proporcionar una muestra de moléculas de ácido nucleico que comprende una pluralidad de secuencias de ácido nucleico diana,

b) hibridar la muestra a un soporte que comprende sondas de ácido nucleico inmovilizadas bajo condiciones que soportan la hibridación entre las sondas de ácido nucleico inmovilizadas y la pluralidad de secuencias de ácido nucleico diana, en el que dichas sondas de ácido nucleico inmovilizadas son complementarias de dicha pluralidad de secuencias de ácido nucleico diana, en el que la densidad de dichas sondas de ácido nucleico inmovilizadas para producir de forma óptima una profundidad de lectura uniforme se predice utilizando un modelo de regresión lineal con ajustes empírico, que ajusta la profundidad de lectura a la densidad de las sondas de ácido nucleico inmovilizadas, y en el que las sondas de ácido nucleico inmovilizadas proporcionan una hibridación uniforme entre dicha pluralidad de secuencias de ácido nucleico diana, y

c) separar las secuencias de ácido nucleico no hibridadas de las secuencias de ácido nucleico diana hibridadas, enriqueciendo así de forma uniforme una población de moléculas de ácido nucleico en una muestra.

2. El método de acuerdo con la reivindicación 1, en el que dicha separación comprende el lavado de dicho soporte.

3. El método de acuerdo con las reivindicaciones 1-2, que además comprende la fragmentación de dicha muestra de moléculas de ácido nucleico previamente a dicha hibridación.

4. El método de acuerdo con la reivindicación 3, que además comprende la ligación de una molécula adaptadora en uno o ambos extremos de una pluralidad de moléculas de ácido nucleico fragmentadas previamente a dicha hibridación.

5. El método de acuerdo con las reivindicaciones 1-4, que además comprende la desnaturalización de dicha muestra de moléculas de ácido nucleico previamente a dicha hibridación.

6. El método de acuerdo con las reivindicaciones 1-5, que además comprende la elución del soporte de una pluralidad de secuencias de ácido nucleico diana hibridadas.

7. El método de acuerdo con la reivindicación 6, que además comprende la secuenciación de las secuencias de ácido nucleico diana eluidas.

8. El método de acuerdo con las reivindicaciones 1-7, en el que dicho soporte es un portaobjetos para microchips.

9. El método de acuerdo con las reivindicaciones 1-7, en el que dicho soporte es una cuenta.

10. El método de acuerdo con las reivindicaciones 1-9, en el que dicha población de moléculas de ácido nucleico es una población de moléculas de DNA genómico.

11. El método de acuerdo con las reivindicaciones 1-9, en el que dicha población de moléculas de ácido nucleico es una población de moléculas de DNA genómico amplificadas.

Fig. 1

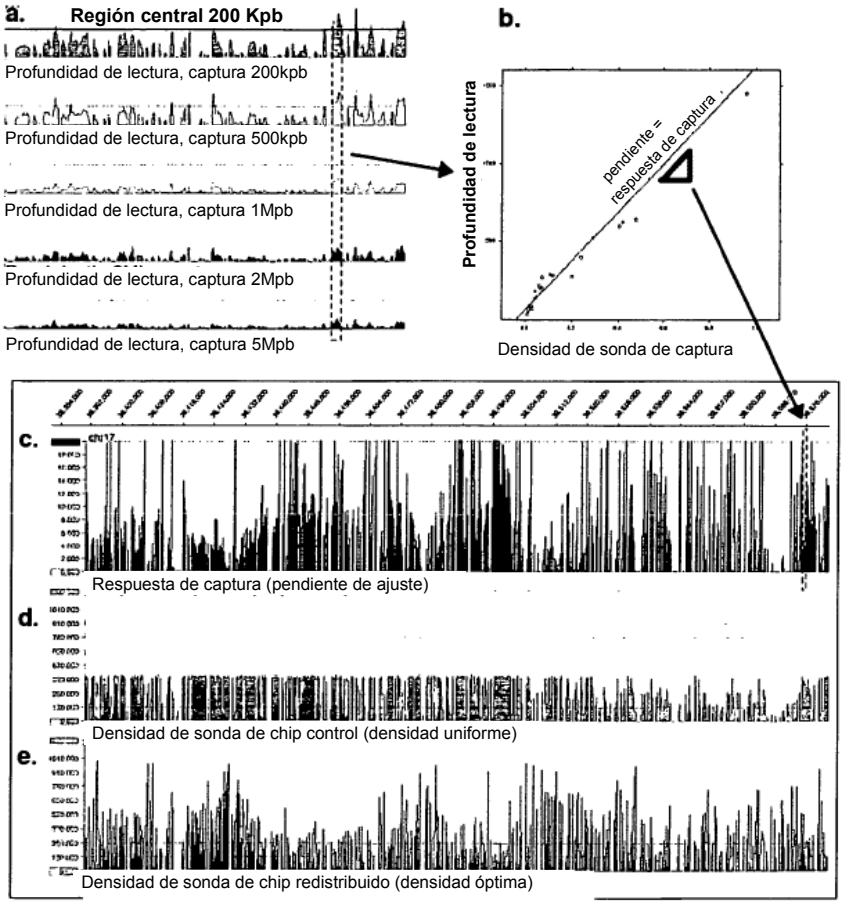

a.
Región central 200 Kpb
Profundidad de lectura, captura 200kpb
Profundidad de lectura, captura 500kpb
Profundidad de lectura, captura 1Mpb
Profundidad de lectura, captura 2Mpb
Profundidad de lectura, captura 5Mpb
b.
Profundidad de lectura
pendiente = respuesta de captura
Densidad de sonda de captura
c.
Respuesta de captura (pendiente de ajuste)
d.
Densidad de sonda de chip control (densidad uniforme)
e.
Densidad de sonda de chip redistribuido (densidad óptima)

Fig. 2

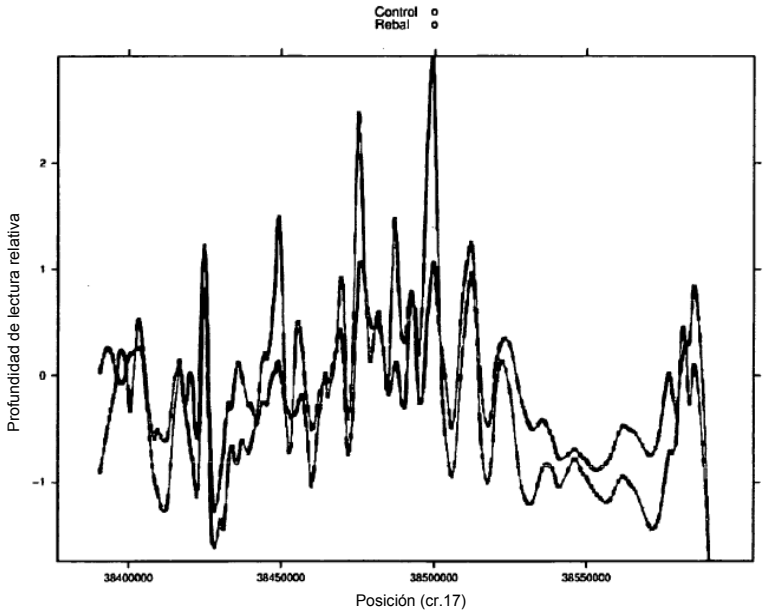
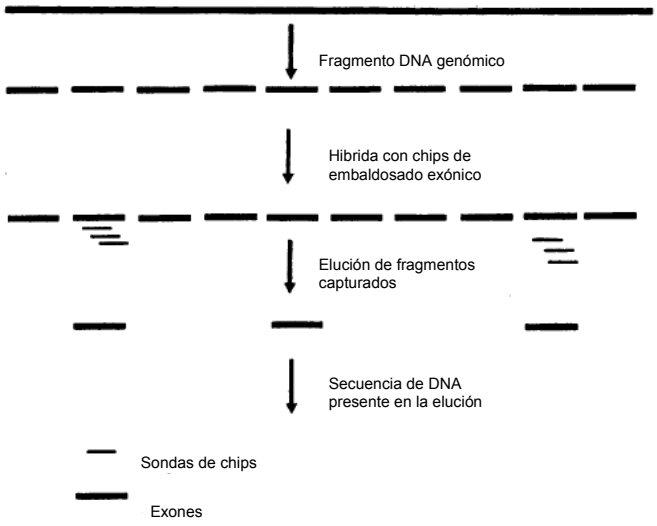

Control
Rebal
Profundidad de lectura relativa
2
1
0
-1
38400000
38450000
38500000
38550000
Posición (cr.17)

Fig. 3

a)


Fragmento DNA genómico
Hibrida con chips de embaldosado exónico
Elución de fragmentos capturados
Secuencia de DNA presente en la elución
Sondas de chips
Exones

b)

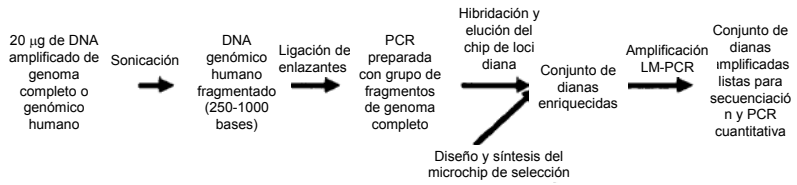

20 μg de DNA amplificado de genoma completo o genómico humano
Sonicación
DNA genómico humano fragmentado (250-1000 bases)
Ligación de enlazantes
PCR preparada con grupo de fragmentos de genoma completo
Hibridación y elución del chip de loci diana
Diseño y síntesis del microchip de selección
Conjunto de dianas enriquecidas
Amplificación LM-PCR
Conjunto de dianas amplificadas listas para secuenciación y PCR cuantitativa

Fig. 4

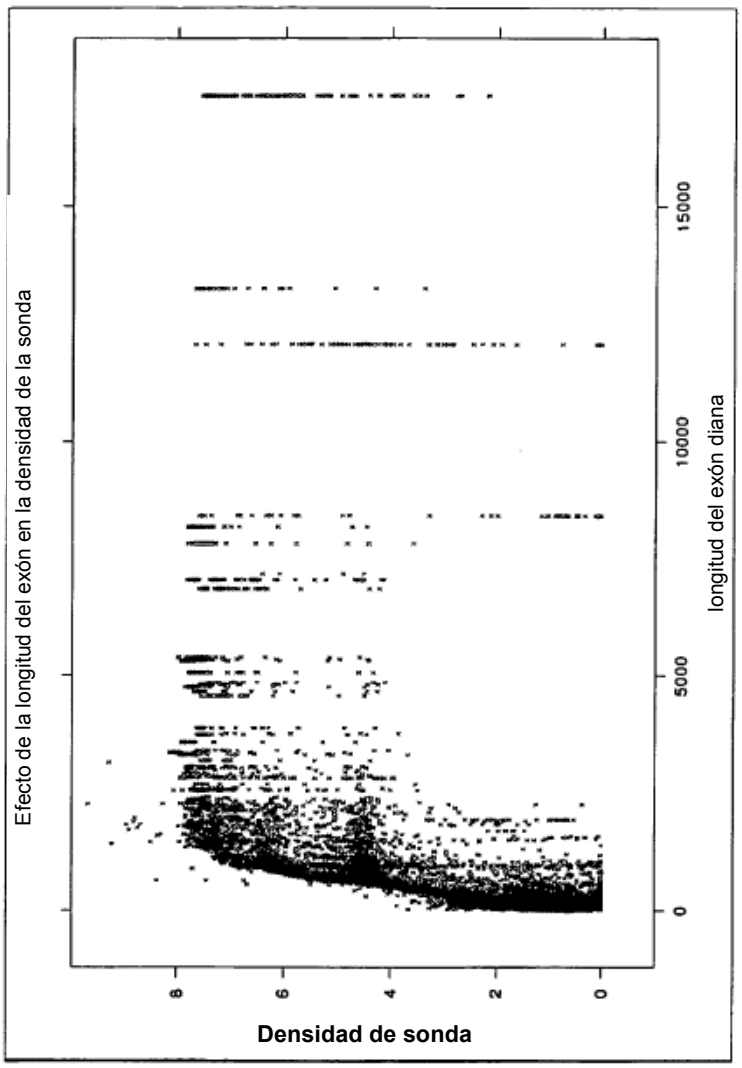

Efecto de la longitud del exón en la densidad de la sonda
longitud del exón diana
0
5000
10000
15000
8
6
4
2
0
Densidad de sonda

Fig. 5

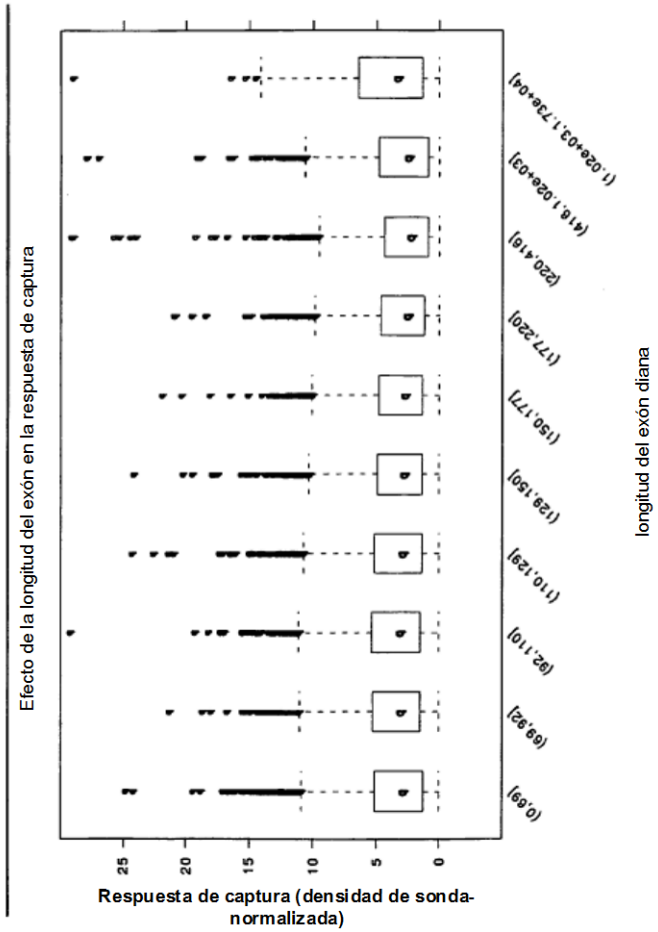
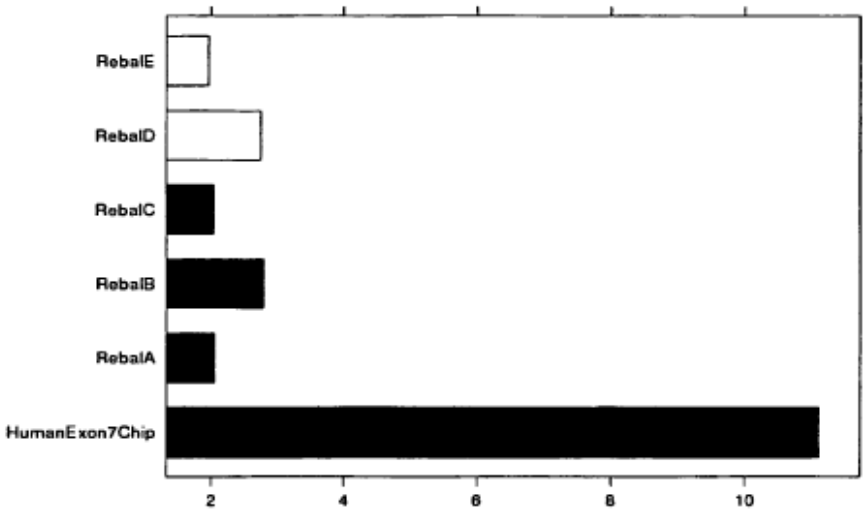

Efecto de la longitud del exón en la respuesta de captura
Respuesta de captura (densidad de sonda-normalizada)
25
20
15
10
5
0
(0,69]
(69,92]
(92,110]
(110,129]
(129,150]
(150,177]
(177,220]
(220,416]
(416,1.02e+03]
(1.02e+03,1.73e+04]
longitud del exón diana

Fig. 6



Puntuación de cobertura no uniforme (DE de la cobertura media por locus)