(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 388 541**

(51) Int. CI.:
G06F 19/20 (2011.01)
C12Q 1/68 (2006.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(96) Número de solicitud europea: **06734379 .8**
(96) Fecha de presentación: **03.02.2006**
(97) Número de publicación de la solicitud: **1859376**
(97) Fecha de publicación de la solicitud: **28.11.2007**

(54) Título: **Método de selección de sondas optimizadas**

(30) Prioridad:
04.02.2005 US 650265 P

(45) Fecha de publicación de la mención BOPI:
16.10.2012

(45) Fecha de la publicación del folleto de la patente:
16.10.2012

(73) Titular/es:
Roche NimbleGen, Inc.
1 Science Court
Madison WI 53711 , US

(72) Inventor/es:
RICHMOND, Todd;
NORTON, Jason;
NUWAYSIR, Emile, F.;
GREEN, Roland y
NUWAYSIR, Kate

(74) Agente/Representante:
de Elzaburu Márquez, Alberto

DESCRIPCIÓN

Método de selección de sondas optimizadas.

Antecedentes de la invención

La llegada de la tecnología de micromatrices de ADN hace posible la construcción de una matriz de cientos de miles de secuencias de ADN en un área muy pequeña, del tamaño de un portaobjetos de microscopio.

Véase, por ejemplo, la patente de EE.UU. Nº 6.375.903 y la patente de EE.UU N° 5.143.854. La descripción de la patente de EE.UU. Nº 6.375.903, permite la construcción del llamado sintetizador de matrices sin máscara (MAS) instrumental en el que se utiliza la luz para dirigir la síntesis de las secuencias de ADN, la dirección de la luz que se controla utilizando un dispositivo digital (DMD).Utilizando un instrumental MAS, la selección de secuencias de ADN que se van a construir en la micromatriz está bajo el control del software de manera que se pueden construir por encargo matrices personalizadas individualmente. En general, la tecnología de síntesis de micromatrices de ADN basada en MAS permite la síntesis en paralelo de más de 786.000 oligonucleótidos únicos en un área muy pequeña de un portaobjetos de microscopio estándar. Las micromatrices se sintetizan generalmente mediante el uso de luz para dirigir que oligonucleótidos se sintetizan en lugares específicos en una matriz, estos lugares se denominan posiciones. Típicamente, se sintetiza una secuencia de nucleótidos en cada posición de la matriz, es decir, hay múltiples sondas en cada posición, pero todas estas sondas tienen la misma secuencia de nucleótidos. Para determinadas aplicaciones, pueden estar presentes oligonucleótidos de secuencias diferentes dentro de una posición de la matriz, y se pueden controlar la proporción y la dirección (5'-3 ', o 3'-5') de estos oligonucleótidos.

Con la disponibilidad de la totalidad del genoma de cientos de organismos, para los que generalmente se ha depositado una secuencia de referencia en una base de datos pública, las micromatrices se han utilizado para realizar análisis de secuencia del ADN aislado de tales organismos. Los métodos de micromatrices que por ejemplo, permiten cuantificar los cambios en el número de copias de ADN son útiles para la determinación de las aberraciones cromosómicas en eucariotas superiores que a menudo están vinculadas a estados de enfermedad. Los cambios en el número de copias suelen ser el resultado de amplificación o deleciones de fragmentos de los cromosomas. Si bien se pueden detectar fácilmente amplificaciones y deleciones o translocaciones grandes por métodos tradicionales de cariotipado, la amplificación o la supresión de fragmentos de ADN más pequeños dentro de un cromosoma puede ser difícil o imposible de detectar por estos métodos. En consecuencia, resulta cada vez más importante para el análisis genético utilizar las sondas de oligonucleótidos más precisas.

Recientemente, varios grupos de investigación han desarrollado métodos para optimizar las sondas. Por ejemplo, para evitar la hibridación cruzada de secuencias muy similares en una micromatriz, los investigadores han desarrollado un método para determinar el número y la longitud óptima de las sondas específicas de genes para estudios de perfiles de transcripción de precisión. El estudio examinó longitudes de sondas de 25 a 1000 nt. Se encontró que las sondas largas producían una intensidad de señal mejor que las sondas cortas. Sin embargo, la intensidad de la señal de las sondas cortas se podría mejorar añadiendo espaciadores o utilizando una mayor concentración de sonda para localización. (Véase Chou et al., Optimization of probe length and the number of probes per gene for optimal microarray analysis of gene expression. *Nucleic Acids Res*.2004 Jul 08; 32 (12): e99). Se cree que el uso de métodos alternativos de optimización de sondas en la identificación de modificaciones genéticas sería una contribución deseable a la técnica.

La patente de EE.UU 2002/13301 está relacionada con métodos para la selección de sondas de ácidos nucléicos. La patente internacional WO 01/05935 está relacionada con el diseño de sondas iterativas y perfiles de expresión detallados con matrices de síntesis in-situ flexibles. La patente internacional WO02/42485 está relacionada con métodos y productos de programas de ordenador para seleccionar sondas de ácidos nucléicos. La patente americana EE.UU 2004/0101846 se refiere a métodos para la identificación de secuencias adecuadas para uso en matrices de ácidos nucleicos.

Mei *et al* (2003) *Proc. Natl. Acad. Sci EEUU* 100(20):11237-11242 se refiere a una selección de sondas para matrices de oligonucleótidos de alta densidad.

Breve compendio de la invención

La presente invención se resume como un método para optimizar las sondas de hibridación de oligonucleótidos para uso en la investigación básica y clínica. La premisa que subyace a esta estrategia de optimización es que las sondas que presentan intensidades de señal correspondientes a diluciones seriadas de una muestra genómica, y que presentan una intensidad de señal fuerte y constante están bien adaptadas para su uso como sondas optimizadas para una variedad de técnicas de hibridación. En particular, la invención proporciona un método para optimizar sondas de oligonucleótidos para su uso en técnicas de hibridación de micromatrices.

Por lo tanto, la presente invención proporciona un método para la optimización de sondas de oligonucleótidos para su utilización en ensayos basados en hibridación, donde dicho método incluye las siguientes etapas:

a) proporcionar una pluralidad de sondas de oligonucleótidos en una matriz de hibridación,

b) proporcionar diluciones seriadas de una muestra genómica, en donde la muestra genómica está marcada

c) hibridar la muestra genómica marcada y diluida en serie con las sondas en la matriz, de tal manera que se produzca una intensidad de señal para cada una de las sondas, en donde la etapa de hibridación se lleva a cabo una vez por lo menos

d) generar computacionalmente los datos de la regresión ponderada a partir de la intensidad de la señal producida por cada una de las sondas

e) identificar las sondas optimizadas a partir de una pluralidad de sondas de oligonucleótidos en una matriz de hibridación para uso en el ensayo utilizando un algoritmo de selección de sondas; en donde las sondas muestran intensidades que corresponden a las diluciones seriadas de la muestra genómica, son reproducibles y están fuertemente relacionadas con las sondas no optimizadas.

Un aspecto de la invención prevé que las sondas de oligonucleótidos sean ADN o ARN.

En otro aspecto, la invención proporciona un método para la optimización de las sondas para cualquier ensayo basado en hibridación seleccionado del grupo que consiste en micromatrices, ensayos basados en esferas, ensayos de genotipado y ensayos de ARNi.

Un aspecto adicional de la invención consiste en utilizar el método de la invención en la optimización de las sondas utilizadas en los campos de la genómica, farmacogenómica, descubrimiento de fármacos, caracterización de alimentos, genotipado, diagnóstico, monitorización de la expresión génica, perfilado de la diversidad genética, ARNi, secuenciación del genoma completo y descubrimiento de polimorfismos, o cualquier otra aplicación que implique la detección de una alteración genética que implique una amplificación o deleción en un cromosoma.

Otras ventajas y características de la presente invención se harán evidentes a partir de la siguiente descripción.

Breve descripción de las diversas vistas de los dibujos

FIG. 1 es un diagrama de Venn que muestra una comparación de los grupos de sondas originales y óptimas que indica que la selección inicial del rango *in silico* se puede mejorar con los datos empíricos de hibridación.

FIG. 2 es una representación gráfica de una pendiente frente a la representación de la intensidad de la señal que muestra que las sondas más luminosas no son siempre las mejores para medir el cambio en la concentración de ADN.

FIG. 3 es una representación gráfica de un r^2 frente a la representación de la intensidad de la señal que muestra que las sondas más luminosas no son siempre las más reproducibles.

FIG. 4 es una representación gráfica de una pendiente frente a un valor r^2 de la línea de regresión ponderada para las sondas individuales después de la serie de diluciones.

Descripción detallada de la invención

La presente invención se refiere a un método para optimizar las sondas de oligonucleótidos para diversas investigaciones básicas y aplicaciones médicas. La premisa que subyace a esta estrategia de optimización es que las sondas que muestran intensidades de señal correspondientes a diluciones seriadas de una muestra genómica, y que muestran una intensidad de señal fuerte y reproducible son las más adecuadas para su uso como sondas optimizadas en técnicas de hibridación diversas. En particular, la invención proporciona un método para optimizar las sondas de oligonucleótidos para su uso en técnicas de hibridación en micromatrices. El método incluye proporcionar varias sondas de oligonucleótidos a una matriz de hibridación; suministrar diluciones seriadas de una muestra genómica, donde la muestra genómica está marcada; hibridar la muestra genómica marcada y diluida en serie con las sondas en la matriz, de tal manera que produzca una intensidad de señal para cada una de las sondas, en donde la etapa de hibridación se lleva a cabo una vez por lo menos; generar computacionalmente los datos de la regresión ponderada a partir de la intensidad de la señal producida por cada una de las sondas, identificar las sondas optimizadas utilizando un algoritmo de selección de sondas; en donde las sondas muestran intensidades de señal que corresponden a las diluciones seriadas de la muestra genómica, son reproducibles y están fuertemente relacionadas con las sondas no optimizadas.

En una realización, la presente invención proporciona un método para identificar y seleccionar las sondas optimizadas para su uso en experimentos de hibridación. En la práctica de la invención, la muestra genómica (marcada y diluida en serie) se hibridó con sondas de oligonucleótidos en una micromatriz de tal manera que se genera una intensidad de señal. Por "muestra genómica," se entiende cualquier fuente, incluyendo plantas, animales, tales como mamíferos, embriones humanos, recién nacidos, adultos, genomas recombinantes, células madre, líneas celulares de tumores sólidos humanos y muestras de tejidos.

Se entiende por marcaje de la muestra genómica, el marcaje terminal con biotina. Como alternativa, los expertos en la técnica considerarán que podrían ser igualmente adecuados otros métodos de marcaje para práctica de la invención. Asimismo, si bien la presente invención no se limita a un conjunto particular de condiciones de hibridación, en una realización preferida de la invención, las hibridaciones se realizaron en un tampón MES (pH 6,6), bajo condiciones estrictas a 45ºC durante 16-18 horas.

Se ajustó una regresión lineal ponderada a las diluciones seriadas para cada sonda de oligonucleótidos y se calculó la pendiente y el coeficiente de correlación. Se utiliza una transformación de log2 de la intensidad de señal y una regresión lineal ponderada para minimizar el efecto de los valores atípicos en los datos. Se calculan las ponderaciones de la regresión lineal ajustando la recta, calculando los residuos, y después utilizando las ponderaciones de un cálculo del promedio robusto de tukey (véase Hubbel, et al, Robust estimators or expression analysis, *Bioinformatics* 2002 Dec; 18 (12) :1585-92) sobre los residuos que se ajustan a una regresión ponderada. Al mismo tiempo, se calculó el promedio robusto de tukey de las intensidades de señal de la sonda a la dilución 4X para representar la intensidad de la señal total de la sonda. La tabla 1, que se muestra a continuación proporciona una descripción de la información necesaria para identificar las sondas optimizadas.

Tabla 1

NOMBRE DE LA COLUMNA	DESCRIPCIÓN
SEQ_ID	Identificador único para secuencia ID
SONDA_ID	Identificador único para sonda ID
POSICION	Posición de la sonda en la secuencia parental
PENDIENTE_PONDERADA	Pendiente ponderada de la serie de diluciones
R_CUADRADO_PONDERADA	Coeficiente de correlación ponderada
INTENSIDAD	Promedio robusto de tukey de la intensidad de la señal de la dilución 4x

Los datos de la regresión ponderada se cargan en una base de datos relacional MySQL (disponible a través de MySQL AB) y las sondas se seleccionan mediante una modificación del algoritmo de selección por rango disponible a través de NimbleGen Systems, Inc., Madison, WI. El algoritmo de selección por rango se describe en el diseño experimental que se ejemplifica más adelante. La modificación es el cambio en las puntuaciones. Por ejemplo, en lugar de por criterios de unicidad, los datos utilizados se obtienen a partir de la hibridación. A continuación se describen las ponderaciones para los diferentes grupos de datos. En una realización típica, el método se lleva a cabo mediante una consulta a la base de datos y la recogida de toda la información de las sondas para una SEQ ID dada. A continuación, se realiza un primer pase a través de las sondas para calcular una puntuación inicial. El primer pase es el cálculo de la puntuación inicial para cada sonda y la selección real de la primera sonda. Esta puntuación inicial se calcula utilizando las siguientes ponderaciones:

* Pendiente ponderada * 100

*Coeficiente de correlación ponderada * 100

*Intensidad * 6

* -3 * Log2 (distancia desde el extremo 3 ') = peso posicional

El objetivo es que cada uno de los principales componentes contribuyan con aproximadamente 1/3 de la puntuación final, y que la ponderación posicional desempeñe un papel menor ya que el proceso inicial de selección de la sonda debería haber espaciado adecuadamente las sondas. Una sonda que empareje perfectamente con las series de dilución de la muestra genómica tendrá una pendiente de 2 (4 veces la dilución en el espacio de log 2). Por lo tanto, la máxima contribución de la pendiente es de 200. Sin embargo, muy pocas sondas sobrepasan un valor de 1. Esto no es infrecuente – la intensidad de señal no se ajusta perfectamente a la concentración de ADN. Una sonda que no responde tendrá una pendiente de 0. Las sondas pueden tener pendientes negativas - éstas añadirán un valor negativo a la puntuación, contrario por lo tanto a la selección de estas sondas. De esta manera, la contribución mínima de la pendiente es de -200. Sin embargo, un intervalo efectivo basado en la experimentación iría desde -100 a aproximadamente 100.

Del mismo modo, el coeficiente de correlación (r^2) puede variar de 0 (ninguna correlación) a 1 (correlación perfecta). Por lo tanto, el intervalo de contribución del coeficiente de correlación es aproximadamente de 0 a 100. Para una intensidad de señal, los datos de una imagen TIFF de 16-bit pueden variar de 1 a 65536. En un espacio de log2, el intervalo va de 0 a 16, de manera que multiplicando por 6 da un intervalo de 0 a 96.

Para la ponderación posicional, la máxima longitud de secuencia para un trascrito, por ejemplo, en un genoma bacteriano es poco probable que sea mucho más larga de 8196 pb (2^{19}), de modo que la penalización máxima por la distancia desde el extremo 3 ' es -3 * log2 (8196 pb) = -39. Una sonda en el extremo 3 'no tendría penalización de manera que el intervalo de la contribuciones para la posición estará de entre 0 y -39. La mayoría de los transcritos bacterianos, en promedio, van a ser inferiores a un millar de pares de bases, por lo que el intervalo más realista estaría entre 0 y -30 aproximadamente.

Después del cálculo puntuación inicial, la sonda con la puntuación más alta se selecciona como la primera en el rango de sondas. Los pases posteriores se pueden llevar a cabo, volviendo a calcular la puntuación añadiendo una bonificación a las sondas que están más alejadas de las sondas previamente seleccionadas, en lugar de penalizar la distancia desde el extremo 3 '. Si la ubicación de la primera sonda seleccionada está en el extremo 3 'en el transcrito más largo, entonces la bonificación máxima sería la misma que la penalización inicial, por lo que la bonificación de posición estará de nuevo en el intervalo de 0 a aproximadamente 30. A medida que se seleccionan más sondas, sin embargo, la bonificación máxima debe necesariamente disminuir puesto que los intervalos entre las sondas seleccionadas disminuyen. Por lo tanto, si todos los otros valores de los datos se mantienen iguales, las puntuaciones disminuirán con cada selección de la sonda sucesiva.

Se incluyen los siguientes ejemplos como ilustraciones adicionales no limitativas de realizaciones particulares de la invención.

Ejemplo

En una realización preferida de la invención, el objetivo fue comenzar con un conjunto inicial de veintiún sondas de 24-mer para cada uno de los 2682 genes de *Lactococcus lactis subsp. cremoris SK11*, y después llevara a cabo el nuevo método de optimización de sondas descrito en la presente memoria para seleccionar las 5 mejores sondas para cada uno de los genes colocados en un único pocillo de 13000 posiciones de formato 12plex de NimbleScreen (disponible a través de NimbleGen Systems, Inc.).

Diseño Experimental

Sondas de Lactococcus lactis subsp. cremoris SK11

En el diseño del experimento, se utilizó la selección de rango estándar de NimbleGen para seleccionar veintiuna sondas de 24-mer a partir de cada una de las 2682 secuencias de *Lactococcus lactis subsp. cremoris SK11* para utilizarlas como conjunto inicial de sondas. La selección de rango estándar es un proceso interactivo, basado en puntuaciones que se utiliza para seleccionar sondas de hibridación basándose en 4 parámetros. La puntuación inicial se calcula utilizando los cuatro parámetros siguientes.

Unicidad ponderada * 100. Por el término "unicidad ponderada" se entiende una medida booleana (0 o 1) de si el oligonucleótido de 24-mer está a 3 desemparejamientos ponderados de cada uno de los otros oligonucleótidos de 24-mer en el genoma diana. Esta medida es independiente de un emparejamiento exacto de 24-mer con otro oligonucleótido.

(24-mer de frecuencia-1) * se penalizan los oligonucleótidos -10.24-mer que tienen más de una coincidencia exacta con el genoma diana.

Puntuación de la composición de pares de bases * 50. La puntuación de la composición de pares de bases es una medida Booleana de si el oligonucleótido 24-mer cumple una serie de reglas basadas en la composición de pares de bases del oligonucleótido, series de bases homopoliméricas y una puntuación de autocomplementariedad.

Ponderación posicional igual a -10 * El log2 de la distancia de la sonda desde el extremo 3 'de la secuencia o del transcrito. Después de seleccionar la primera sonda para cada secuencia, la ponderación posicional se modifica con una bonificación que depende de la distancia a la sonda seleccionada más cercana, que como resultado fuerza incluso el espaciado a lo largo de la secuencia diana.

Sondas de *E. coli*

Las sondas de *E. coli* K12 se seleccionaron como controles de normalización de intensidad. Se seleccionaron un total de 6044 sondas de E. coli de la siguiente manera. Las sondas se colocaron con un intervalo de 10 pares de bases por todo el genoma de *E. coli* K12. Este grupo de sondas se hizo pasar por proceso estándar de selección de sondas de NimbleGen System, Inc. (descrito anteriormente) para recoger información de la sonda, utilizando tanto hebras directas como inversas del genoma de *L. cremoris* como única diana. En lugar del proceso de selección de rangos normal, sin embargo, se aplicó un filtro sencillo para incluir sólo las sondas que podrían ser sintetizadas en 72 ciclos o menos; que no aparecían como una coincidencia exacta en el genoma de *L. cremoris*, y que estaban a 3 desemparejamientos ponderados de distancia de cualquier 24-mer de *L. cremoris* .

Sondas de GC aleatorias

Se colocaron en la matriz sondas aleatorias (1900) de contenido en GC definido (6-14%) como controles de normalización de intensidad final baja. El porcentaje de GC se calculó sobre la base del contenido promedio de GC (+ / -2 desviaciones estándar) de las sondas de *L. cremoris* en la matriz.

Diseño

La optimización del diseño se llevó a cabo en la plataforma estándar de expresión de NimbleGen, 385.000 posiciones con densidad de formato de posición 01:02, sin desemparejamientos. Para compensar los posibles problemas de uniformidad en la matriz, las sondas se dispusieron en tiras verticales sobre la matriz, para un total de 6 grupos de replicación. Cada grupo de sondas/ control se coloco aleatoriamente en recipientes solapados identificados (ECOLI_BLOCK1, LCRE_BLOCK1, RANDOM_BLOCK1, etc.)

Marcaje

El ADN genómico se amplificó utilizando un kit de REPLI-g (Qiagen, Inc.), fue extraído con fenol y precipitado con etanol. Se prepararon tres muestras utilizando 2,5 µg (microgramos) de ADN genómico control amplificado de *E. coli* y tres cantidades diferentes de muestra de ADN genómico amplificado de *L. cremoris* (0,625 µg, 2,5 µg y 10,0 µg). El control y la muestra de ADN se combinaron y se sometieron a digestión con ADNasaI de manera que el tamaño del fragmento final oscilara entre 50-200 pb. El fragmento de ADN se marco en su extremo terminal con Biotina-N6-ddATP usando una transferasa terminal como preparación para la hibridación.

Optimización de las hibridaciones y barrido

Para las hibridaciones se utilizaron 2.5 µg de ADN genómico de *E. coli* y 3 concentraciones diferentes de ADN genómico de *L. cremoris*: 0.625 µg, 2,5 µg y 10,0 µg, proporcionando diluciones 0,25X, 1X y 4X. Se llevaron a cabo dos hibridaciones para cada dilución en condiciones de expresión de la hibridación, y se realizaron tres barridos estándar para cada matriz a ajustes de voltaje variables de los tubos fotomultiplicadores (PMT), dando un total de 18 imágenes (3 diluciones x 2 replicaciones x 3 ajustes de PMT). Los voltajes del PMT abarcaron un intervalo de 100 V, con pasos de 50, asegurando que una serie de barridos capturaría el intervalo completo de datos sin saturar las posiciónes. PMT1 fue el ajuste medio, PMT2 fue el extremo más alto, y PMT3 fue el extremo más bajo.

Normalización de los datos

Cada imagen fue extraída utilizando el software 2.0 NimbleScan de NimbleGen y se guardó como un archivo PAIR de NimbleGen. Se realizó la extracción por recipientes. Después de la extracción, los datos en el archivo de PAIR se combinaron y reordenaron, de modo que cada bloque de replicación que contenía el ECOLI, RANDOM y LCRE se colocaba en una columna separada en un único archivo de datos. Cada conjunto de datos de los tres diferentes PMT fue tratado por separado. Se produjo un archivo para cada PMT que contenía las columnas que se describen en la Tabla 2.

Tabla 2

NOMBRE DE LA COLUMNA	DESCRIPCION
GEN_EXP_OPCION	ECOLI,LCRE o ALEATORIO
SEQ_ID	Ecoli K12, RANDOM. Designación GC, o SEQ_ID *de L. cremoris*
SONDA_ID	SONDA_IDs individuales
POSICIONES	Posición en el genoma para *E. coli* o posición en la secuencia para *L. cremoris*
CHIP_ID_PMT1_BLOCK1	1ª columna de datos para el experimento
CHIP_ID_PMT1_BLOCK2	2ª columna de datos para el experimento
Etc.	Etc.

La normalización se realizó utilizando R, un lenguaje y entorno para procesado de datos estadísticos y gráficos, y el paquete vsn (normalización estabilizadora de la varianza) del proyecto BioConductor (http://bioconductor.org). El "paquete VSN" funciona calculando las "variaciones muestra a muestra mediante desplazamiento y escalado, y transforma las intensidades a una escala donde la varianza es aproximadamente independiente de la intensidad del promedio. La transformación estabilizadora de la varianza es equivalente al logaritmo natural en el intervalo de alta intensidad, y a una transformación lineal en el intervalo de baja intensidad. En un intervalo intermedio, la función arsinh interpola sin problemas entre los dos ". Para la normalización de datos, se utilizaron las sondas "ECOLI" y "RANDOM" para generar parámetros de normalización que posteriormente se aplicaron al conjunto de datos.

Después de la normalización, las sondas de *L. cremoris* se pasaron a archivos de texto para su posterior procesamiento. El archivo de datos normalizado tiene el mismo formato de columnas que el archivo de datos en bruto, pero contiene solamente la información de las sondas de *L. cremoris.*

Cálculo de la recta de mejor ajuste a la serie de diluciones

La idea que subyace a la estrategia de optimización es que las sondas con las intensidades de señal que mejor siguen la serie de diluciones, y que constantemente tienen la máxima luminosidad van a ser seleccionadas como sondas optimizadas. Para encontrar esas sondas se ajusta una recta de regresión ponderada a la serie de diluciones para cada sonda y se calcula la pendiente y el coeficiente de correlación. Se utilizan una transformación de log2 de la intensidad de señal y una regresión lineal ponderada para minimizar el efecto de los valores extremos en los datos. Las ponderaciones de la regresión lineal se calculan ajustando la recta, calculando los residuos, y a continuación, utilizando las ponderaciones a partir del cálculo del promedio robusto tukey de los residuos para adaptarse a una regresión ponderada. Al mismo tiempo, se calcula el promedio robusto tukey de las intensidades de señal de la sonda a la dilución 4x para representar la intensidad de la señal total de la sonda. Toda esta información se pasa a un archivo de texto. La tabla 3, que se muestra a continuación proporciona una descripción de las columnas de datos en este archivo.

Tabla 3

NOMBRE DE LA COLUMNA	DESCRIPCION
SEQ_ID	Identificador único para SEQ_ID
SONDA_ID	Identificador único para SONDA_ID
POSITION	Posición de la sonda en la secuencia parental
PENDIENTE_PONDERADA	Pendiente ponderada de la serie de diluciones
R_CUADRADO_PONDERADA	Coeficiente de correlación ponderada
INTENSIDAD	Promedio robusto de tukey de la intensidad de la señal de la dilución 4x

Bases para la optimización de la sonda

Una idea equivocada frecuente en los estudios empíricos de optimización de sondas es que las sondas con intensidad de la señal más luminosa son las sondas mejores. Los siguientes gráficos muestran que a menudo este no es el caso. La figura 2 muestra que, en promedio, las sondas que se encuentran en el intervalo medio de intensidades de señal varían en paralelo con la serie de diluciones de ADN mejor que las sondas más luminosas. La figura 3 muestra que las sondas comprendidas entre la mitad y el extremo mas alto de intensidad también tienden comportarse de manera más consecuente, de acuerdo con el coeficiente de correlación de la línea de regresión. La figura 4 muestra un diagrama de dispersión de la pendiente en comparación con el valor r^2 de la línea de regresión ponderada para las sondas individuales después de la serie de diluciones. Esto demuestra que hay un gran número de sondas que cumplen los criterios de rendimiento constante y capacidad de variar en paralelo con la serie de diluciones de ADN. La cola de la izquierda es el resultado de las pendientes negativas -que indica que hay sondas que se vuelven más tenues a medida que aumenta la concentración de ADN. Algunas de las sondas con pendientes negativas tienen valores r^2 que son bastante altos. Esto puede indicar que hay algún tipo de hibridación competitiva en juego, ya que cuando la cantidad de ADN de *L. cremoris* aumenta, la intensidad de la señal de la sonda disminuye.

Selección de las sondas óptimas

Los datos de la regresión ponderada se cargan en una base de datos MySQL y se seleccionan las sondas óptimas mediante una modificación del algoritmo de selección de rango de NimbleGen. Las ponderaciones de los diferentes datos se basaron en los gráficos anteriores, y en los resultados de los experimentos de optimización anteriores. Se compararon los datos de los tres ajustes PMT y se encontró que eran esencialmente los mismos. Se seleccionó PMT2, el conjunto de datos más luminoso, para la optimización, ya que las intensidades de señal de la sonda tenían el mayor intervalo. El procedimiento de selección de las sondas detallado se llevo a cabo tal como se ha descrito anteriormente.

También se prevé que el proceso de selección de sondas óptimas que se describe en este ejemplo (es decir, introducir los datos de regresión en una base de datos, consultar la base de datos y seleccionar las sondas optimizadas) se podría también llevar a cabo utilizando un "archivo de texto delimitado por tabuladores" en lugar de almacenar los datos en una base de datos y posteriormente recuperar los datos. En consecuencia, la base de datos que se describe en la presente memoria está destinada a ser sólo una herramienta de conveniencia y no un medio para limitar el método de la invención.

Resultados

Superposición del grupo de sondas

La figura 1 es un diagrama de Venn del grupo(s) de las sondas originales seleccionadas por rango y del grupo optimizado final. La intersección de las 5 sondas originales superiores y del conjunto de las sondas óptimas es aproximadamente igual a la que cabría esperar al azar. Esto indica que la selección de rango *in silico* inicial puede ser mejorada por los datos de hibridación empíricos.

Gráficos que muestran las sondas óptimas

La figura 2 muestra que, en promedio, las sondas que se encuentran en el intervalo de intensidades de señal media varían en paralelo a la serie de diluciones de ADN mejor que las sondas más luminosas. La figura 3 muestra que las sondas comprendidas entre la mitad y el extremo más alto de intensidad también tienden comportarse de manera más consecuente, de acuerdo con el coeficiente de correlación de la línea de regresión. La figura 4 muestra un diagrama de dispersión de la pendiente en comparación con el valor r2 de la línea de regresión ponderada para las sondas individuales después de la serie de diluciones. Esto demuestra que hay un gran número de sondas que cumplen los criterios de rendimiento constante y de capacidad variar en paralelo con la serie de diluciones de ADN. Por lo tanto, los resultados comprobados en el experimento cumplen con nuestras expectativas, mostrando la mayoría de las sondas óptimas las pendientes más altas, los valores r^2 mas grandes e intensidad de señal relativa intermedia. Al examinar las gráficas, en ocasiones hay sondas seleccionadas que no cumplen los criterios anteriores. En general, estas sondas pertenecen a genes en los que todas las sondas eran no óptimas. Esto indica que las sondas finales seleccionadas fueron las mejores de un lote malo, y puede que no sea posible seleccionar sondas buenas de este subgrupo muy pequeño de genes. Estas sondas/ genes deben ser vistas con desconfianza en otras hibridaciones de ARN posteriores.

Se entiende que algunas adaptaciones de la invención descritas en esta memoria serían optimizaciones rutinarias para los expertos en la técnica, y pueden ser implementadas sin apartarse del espíritu de la invención, o del alcance de las reivindicaciones adjuntas.

REIVINDICACIONES

1. Un método para optimizar sondas de oligonucleótidos para uso en un ensayo de hibridación, donde dicho método comprende las siguientes etapas:

a) proporcionar una pluralidad de sondas de oligonucleótidos en una matriz de hibridación;

b) proporcionar diluciones seriadas de una muestra genómica, en donde la muestra genómica está marcada;

c) hibridar la muestra genómica marcada y diluida en serie con las sondas en la matriz, de tal manera que se produce una intensidad de señal para cada una de las sondas, en donde la etapa de hibridación se lleva a cabo una vez por lo menos;

d) generar computacionalmente los datos de la regresión ponderada a partir de la intensidad de la señal producida por cada una de las sondas; e

e) identificar las sondas optimizadas a partir de una pluralidad de sondas de oligonucleótidos en una matriz de hibridación para uso en el ensayo utilizando un algoritmo de selección de sondas; donde las sondas que muestran intensidades que corresponden a las diluciones seriadas de la muestra genómica, son reproducibles, y están fuertemente relacionadas con sondas no optimizadas

2. El método de la reivindicación 1, en el que las sondas de oligonucleótidos son ADN o ARN.

3.El método de la reivindicación 2, en el que los ensayos basados en hibridación se seleccionan del grupo que consiste en ensayos basados en micromatrices, ensayos basados en esferas, ensayos de genotipado, y ensayos de RNAi.

4. El método de la reivindicación 3, en el que los ensayos basados en hibridación se realizan en los campos de la genómica, farmacogenómica, descubrimiento de fármacos, caracterización de alimentos, genotipado, diagnóstico, monitorización de la expresión génica, perfilado de diversidad genética, secuenciación de genoma completo y descubrimiento de polimorfismos, o cualquier otra aplicación que implique la detección de alteraciones genéticas que implican una amplificación o deleción en un cromosoma.

5. El método de la reivindicación 1, en el que el algoritmo de selección de sondas para identificar las sondas optimizadas es un algoritmo de selección de rango.

6. El método de la reivindicación 5, en el que el algoritmo de selección de rango para identificar las sondas optimizadas comprende calcular una puntuación para cada sonda utilizando el criterio que comprende la ponderación de la pendiente, el coeficiente de correlación ponderado, la intensidad y la ponderación posicional.

7. El método de la reivindicación 1, en el que el ensayo basado en hibridación es un ensayo basado en micromatrices.

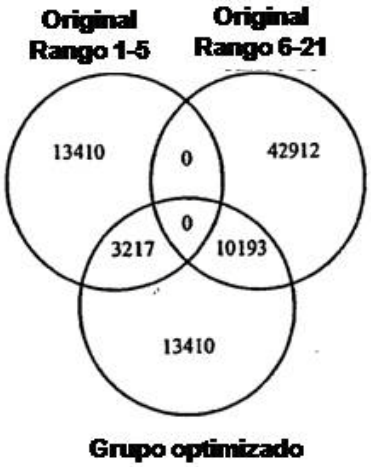
Original
Rango 1-5
Original
Rango 6-21
13410
0
42912
0
3217
10193
13410
Grupo optimizado

FIG 1

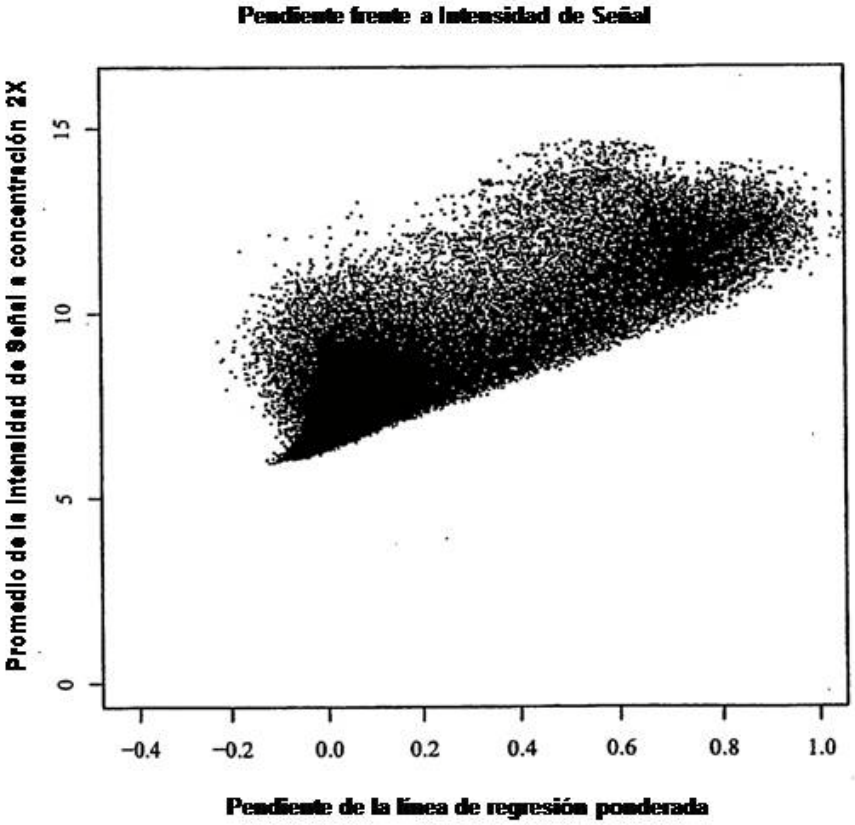
Pendiente frente a Intensidad de Señal
Promedio de la Intensidad de Señal a concentración 2X
0
5
10
15
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
Pendiente de la línea de regresión ponderada

FIG 2

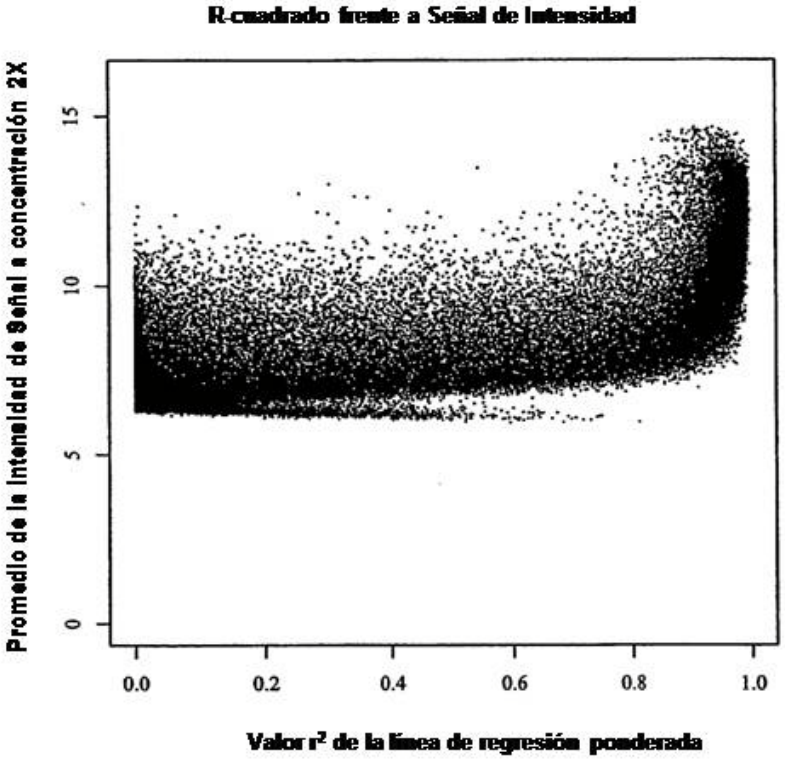
R-cuadrado frente a Señal de Intensidad
Promedio de la Intensidad de Señal a concentración 2X
0
5
10
15
0.0
0.2
0.4
0.6
0.8
1.0
Valor r² de la línea de regresión ponderada

FIG 3

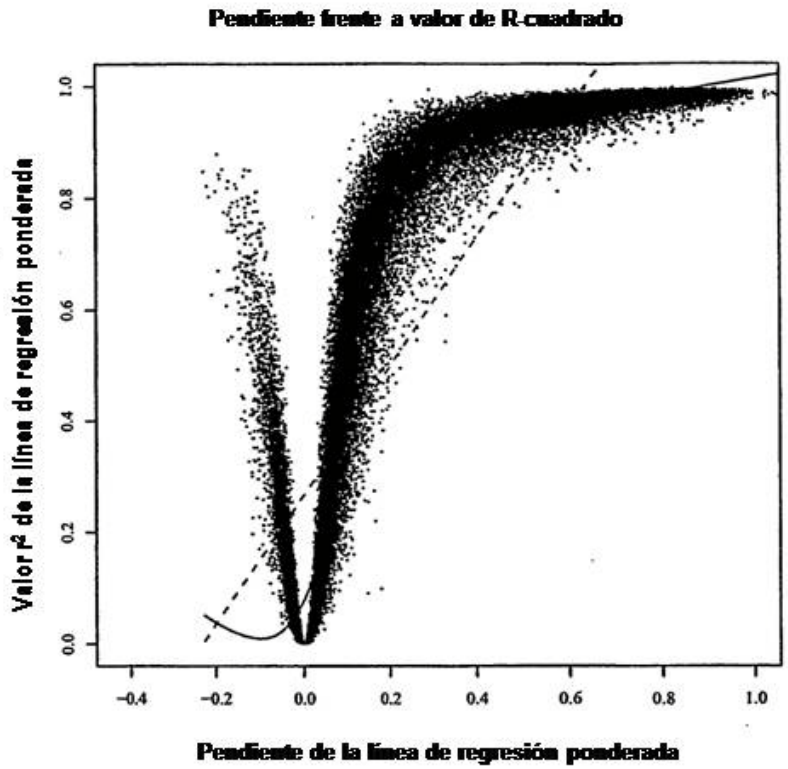
Pendiente frente a valor de R-cuadrado
Valor r² de la línea de regresión ponderada
0.0
0.2
0.4
0.6
0.8
1.0
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
Pendiente de la línea de regresión ponderada

FIG 4