(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 675 618**

(51) Int. CI.:

C40B 20/04 (2006.01)
C12Q 1/68 (2008.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(86) Fecha de presentación y número de la solicitud internacional: **25.11.2013 PCT/US2013/071656**

(87) Fecha y número de publicación internacional: **30.05.2014 WO14082032**

(96) Fecha de presentación y número de la solicitud europea: **25.11.2013 E 13857053 (6)**

(97) Fecha y número de publicación de la concesión europea: **04.04.2018 EP 2922989**

(54) Título: **Métodos para la secuenciación estandarizada de ácidos nucleicos y usos de los mismos**

(30) Prioridad:

26.11.2012 US 201261729853 P
27.11.2012 US 201261730463 P
14.03.2013 US 201361784394 P

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
11.07.2018

(73) Titular/es:

THE UNIVERSITY OF TOLEDO (100.0%)
2801 Bancroft St., Ms 218
Toledo, OH 43606, US

(72) Inventor/es:

WILLEY, JAMES, C.;
BLOMQUIST, THOMAS y
CRAWFORD, ERIN

(74) Agente/Representante:

LINAGE GONZÁLEZ, Rafael

DESCRIPCIÓN

Métodos para la secuenciación estandarizada de ácidos nucleicos y usos de los mismos

Declaración respecto a la investigación patrocinada por el gobierno federal

La invención se realizó con apoyo del gobierno de EE.UU. con los números de subvención CA138397 y HL108016 concedidos por los Institutos Nacionales de Salud. El gobierno de EE.UU. tiene ciertos derechos en la invención.

Campo de la invención

La presente invención se refiere métodos para la secuenciación estandarizada de ácidos nucleicos y usos de los mismos.

Antecedentes

La identificación de la información genética se está convirtiendo en una pieza clave de información para el diagnóstico y tratamiento de muchas enfermedades. Para hacer que dicha herramienta de diagnóstico esté fácilmente disponible, se desea que esta identificación sea lo más eficiente y económica posible. Para aspectos de diagnóstico, médicos, regulatorios y éticos, esta identificación debe ser lo más exacta posible para descartar mediciones falsas.

Además del deseo de adquirir información de material genético humano, existe un gran interés en la adquisición de información genética de, por ejemplo, mitocondrias, agentes patógenos y organismos que causan enfermedades.

Un método para la adquisición de información es el método de secuenciación de Sanger de análisis del genoma. Están disponibles otros métodos que proporcionan un rendimiento mejorado en comparación con el método de secuenciación de Sanger. Dichos métodos incluyen una tecnología de secuenciación corta en paralelo de alta densidad, secuenciación de nueva generación (es decir, NextGen o "NGS"), que intentan proporcionar una visión del ARN en muestras biológicas más completa y exacta que el método de secuenciación de Sanger.

La secuenciación de nueva generación (NGS) es útil en una multitud de aplicaciones clínicas en virtud de su análisis automatizado y altamente paralelizado de plantillas de ácidos nucleicos. Sin embargo, el límite de las cuestiones clínicas que la NGS puede abordar viene determinado en gran medida por: i) la fuente de la plantilla de ácido nucleico en la secuencia ascendente (por ejemplo, tejido humano, muestra microbiana, etc.), y ii) si la variación biológica clínicamente relevante en la plantilla de ácido nucleico es mayor que la variación técnica (que a menudo se introduce mediante dichas variantes como el flujo de trabajo para la preparación de muestras, la secuenciación y/o el análisis de datos).

El flujo de trabajo para la preparación de bibliotecas de NGS varía ampliamente, pero en términos generales se pueden agrupar en dos enfoques: 1) digestión o fragmentación de la muestra de ácido nucleico con unión posterior a una secuencia adaptadora universal, o 2) PCR con cebadores específicos de la diana que incorporan una secuencia adaptadora universal en sus extremos 5'. En ambos enfoques, si una plantilla de ácido nucleico es ARN, se usa una etapa de transcripción inversa para crear la plantilla de ADN requerida para la secuenciación.

Una preocupación con la NGS es que estos métodos de secuenciación cuantitativos tienen una alta variación intralaboratorio e interlaboratorio. Por lo tanto, este problema reduce el valor de cualquier resultado y ha impedido el uso de estos métodos de secuenciación en el diagnóstico molecular.

Por ejemplo, a menudo se introducen inadvertidamente sesgos no sistemáticos (es decir, no reproducibles) (es decir, errores) durante la preparación de la biblioteca de secuenciación. Estos sesgos no sistémicos son un obstáculo importante para la implementación de la NGS como una medida de rutina fiable y eficiente de la abundancia de ácidos nucleicos (cuantificación) en el entorno clínico.

La fuente más probable de sesgo no sistemático (que impide de este modo la comparación interlaboratorio y, por tanto, el uso clínico de rutina, de datos cuantitativos de NGS) se deriva de los problemas que surgen de la fragmentación del ácido nucleico, la unión del adaptador y la PCR.

Además, aunque no es explícitamente necesario, la FDA ha publicado indicaciones y recomendaciones para la industria acerca de que los dispositivos de diagnóstico in vitro (IVD) basados en PCR deberían contener controles internos de amplificación (IAC) para controlar las sustancias que interfieren y para verificar que un resultado negativo para una muestra no es causado por inhibidores.

Además, con el fin de evitar el error de muestreo estocástico y de garantizar mediciones fiables, es necesario secuenciar (es decir, leer) un número suficiente de copias del analito que se está midiendo. Un problema es que el intervalo de representación de los transcritos después de la preparación de la biblioteca a menudo permanece muy

alto, típicamente un millón de veces o más, lo que impone un alto coste. Esto se debe a que los transcritos de cada gen se deben secuenciar al menos 10 veces (para garantizar 10 "lecturas"). Para garantizar 10 lecturas de los genes menos representados, es necesario leer un gen representado a un nivel un millón de veces más alto al menos 10 millones de veces.

Por lo tanto, un método de NGS que reduce la variación entre experimentos y entre laboratorios en la medición del número de copias de ácido nucleico en las muestras será de gran utilidad tanto para aplicaciones clínicas como de investigación.

Sumario de la invención

En el presente documento se describe un método para proporcionar reproducibilidad en la medición del número de copias de ácido nucleico en las muestras, que comprende medir una relación proporcional de al menos un evento de secuenciación de la diana natural de al menos un ácido nucleico en una muestra con respecto al control interno de amplificación (IAC) competitivo respectivo para ese ácido nucleico.

En el presente documento también se describe un método en el que el al menos un evento comprende: una observación, un recuento y/o una lectura entre la diana natural y su IAC respectivo.

En el presente documento también se describe un método para controlar el error no sistemático en la preparación de bibliotecas de NGS basada en PCR, que comprende compartir sitios de cebado idénticos con una plantilla de ácido nucleico natural de interés a fin de imitar la cinética de la diana natural en la reacción de PCR y, por lo tanto, de controlar la variación en la eficiencia de la PCR específica de la diana.

En el presente documento también se describe el uso de un método de IAC competitivo para contemplar la convergencia de la representación del analito diana en una muestra, al tiempo que conserva información cuantitativa de la representación original tanto de la diana de baja abundancia como de la diana de alta abundancia, lo que permite la medición cuantitativa de la representación original con un bajo número de lecturas de secuenciación.

Además, en el presente documento se describe un método para determinar una cantidad de un primer ácido nucleico, que comprende:

proporcionar una serie de mezclas estandarizadas diluidas en serie que comprenden una plantilla competitiva para dicho primer ácido nucleico y una plantilla competitiva para un segundo ácido nucleico presente en varias muestras que comprenden dicho primer ácido nucleico, en el que dichas plantillas competitivas están en concentraciones conocidas entre sí;

combinar una de dichas muestras que comprende dicho primer ácido nucleico con una primera de dichas mezclas estandarizadas diluidas en serie;

coamplificar dicho primer ácido nucleico y dicha plantilla competitiva para dicho primer ácido nucleico para producir el producto amplificado del mismo;

obtener una primera relación, comparando dicha primera relación dicho producto amplificado de dicho primer ácido nucleico con dicho producto amplificado de dicha plantilla competitiva para dicho primer ácido nucleico;

determinar si dicha primera relación está dentro de aproximadamente 1:10 a aproximadamente 10:1;

si no es así, repetir dichas etapas de combinación, coamplificación, obtención y determinación con una segunda de dichas mezclas estandarizadas diluidas en serie;

coamplificar dicho segundo ácido nucleico y dicha plantilla competitiva para dicho segundo ácido nucleico para producir el producto amplificado del mismo;

obtener una segunda relación, comparando dicha segunda relación dicho producto amplificado de dicho segundo ácido nucleico con dicho producto amplificado de dicha plantilla competitiva para dicho segundo ácido nucleico; y

comparar dicha primera relación y dicha segunda relación.

El método puede incluir comparar dicho proyecto amplificado de dicho primer ácido nucleico con dicho proyecto amplificado de dicha plantilla competitiva para dicho primer nucleico, determinando si dicha primera relación está dentro de aproximadamente 1:100 a aproximadamente 100:1 o de 1:1000 a aproximadamente 1000:1 o de 1:10 000 a aproximadamente 10 000:1,

si no es así, repetir dichas etapas de combinación, coamplificación, obtención y determinación con una segunda de dichas mezclas estandarizadas diluidas en serie;

coamplificar dicho segundo ácido nucleico y dicha plantilla competitiva para dicho segundo ácido nucleico para producir el producto amplificado del mismo; obteniendo una segunda relación;

en el que dicha segunda relación compara dicho producto amplificado de dicho segundo ácido nucleico con dicho proyecto amplificado de dicha plantilla competitiva para dicho segundo ácido nucleico; y

comparar dicha primera relación y dicha segunda relación.

Además, los productos de la serie de reacciones de coamplificación se pueden combinar y amplificar en una segunda ronda utilizando pares de cebadores que reconocen cada producto DN y DC de la primera ronda de amplificación y que también tienen un cebador con código de barras específico del gen y un cebador universal en el extremo 5' para facilitar la secuenciación.

Breve descripción de los dibujos

La patente o el archivo de solicitud pueden contener uno o más dibujos ejecutados en color y/o una o más fotografías. La Oficina de Patentes proporcionará copias de esta patente o publicación de solicitud de patente con dibujo(s) a color y/o fotografía(s) previa solicitud y pago de la tarifa necesaria.

Figura 1: ADNg A549 valorado con relación a la mezcla ICA. Gráfico que muestra la valoración de una mezcla de controles internos de amplificación (IAC) competitivos aproximadamente equimolares en relación con una cantidad fija de entrada de ADN genómico (ADNg) de 100 000 copias en cada PCR múltiple. En el eje X se representa la cantidad inicialmente estimada de cada IAC en la mezcla equimolar. El eje Y representa la frecuencia de eventos de secuenciación (lecturas) observados para cada plantilla natural dividida por la suma de las frecuencias de lectura para la plantilla natural y su respectivo IAC competitivo.

Figuras 2A-2F: Gráficos que muestran la valoración de una mezcla de controles internos de amplificación (IAC) competitivos aproximadamente equimolares en relación con: figura 2A) una cantidad de entrada fija de 100 000 copias de ADN genómico (ADNg) en PCR múltiple, o de 11 ng de material de ADNc de ARN transcrito inversamente de muestras SEQC: figura 2B) A-RT1, figura 2C) A-RT2, figura 2D) B, figura 2E) C y figura 2F) D, en cada PCR múltiple. En el eje X se representa la cantidad inicialmente estimada de cada IAC en la mezcla equimolar. El eje Y representa la frecuencia de eventos de secuenciación (lecturas) observados dividida por la suma de las frecuencias de lectura para la plantilla natural y su respectivo IAC competitivo.

Figura 3: Gráficos que muestran el análisis de la exactitud del ensayo. Los valores medidos para la muestra SEQC C-RT1 se compararon con los valores esperados (% de diferencia) en base a los valores de las muestras SEQC A y B. La diferencia porcentual entre la señal predicha C' y la señal del ensayo real C se usó como una indicación de la exactitud relativa (ER) del ensayo. Una puntuación de ER ΔC para un gen diana se definió como (C-C'/C'), respectivamente. Se presentan el valor medio de ER (línea), los cuartiles de la mediana (recuadro), la desviación estándar (bigotes) y los valores atípicos. El valor medio de ER fue muy cercano al valor estimado. Solo ciertos ensayos tienen una ER con una diferencia superior al 25 % con respecto a la media.

Figura 4A: Gráfico que muestra las dianas génicas (n = 88) medidas entre las muestras SEQC ART1, B, C en las que se evaluó la expresión diferencial (ED) entre genes; es decir, si su ED experimentaba un cambio de entre 1,5 y 3,0 veces, entre 2 y 3 veces, entre 3 y 5 veces, entre 5 y 10 veces o más de 10 veces. El control para el cambio falso positivo o negativo se evaluó comparando SEQC A-RT1 frente a SEQC A-RT2.

Figura 4B: La tabla 1 presenta las estadísticas resumidas de la expresión diferencial en la muestra C en comparación con las esperadas en base a las muestras A y B.

Figura 5: Gráficos que muestran la reproducibilidad del ensayo entre transcripciones inversas. Se midieron dos transcripciones inversas de la muestra SEQC A (RT1 frente a RT2) para la expresión de 119 dianas génicas que superaron con éxito los criterios de rendimiento de la figura 1.

Figura 6: Gráfico que muestra los mismos datos que en la figura 3 y la figura 4. Mediciones esperadas en la muestra C (eje x) frente a mediciones observadas (eje y). La transcripción inversa de la muestra SEQC C se midió para la expresión de 119 dianas génicas que superaron con éxito los criterios de rendimiento de la figura 1. De las 119 dianas génicas, 88 tenían $R^2 > 0,95$ para el ajuste de la curva de la ecuación de la pendiente para determinar el punto de equivalencia y la concentración de cada diana (eje y) (figura 2).

Figura 7: Gráfico que muestra la convergencia y el aumento de la uniformidad de 97 dianas génicas de la figura 5. En el eje X se representan los datos de la figura 5 en proporción a la plantilla de mayor abundancia. En el eje Y se representa la proporción real de lecturas de secuenciación en las que está la diana génica en proporción a la plantilla de secuencia más alta.

Figura 8: Gráfico que muestra la curva ROC para detectar un cambio de > 3 veces para la secuenciación de ARN (RNA-Seq) utilizando la plataforma Illumina de Bullard et al. BMC Bioinformatics 2010, 11:94.

Figura 9: Proporciona una ilustración esquemática de una mezcla maestra de PCR con una mezcla de controles internos de amplificación (IAC).

Figuras 10-11: Gráficos que muestran la valoración de una mezcla de controles internos de amplificación contra ADNg y ADNc de SEQC.

Figura 12: Gráfico que muestra la misma secuenciación de réplicas de preparación de biblioteca (intrasitio), en el que el eje X = 1,8 millones de lecturas de secuenciación y el eje Y = 3,0 millones de lecturas de secuenciación.

Figura 13: Gráfico que muestra la preparación de una biblioteca separada secuenciada (intrasitio), en el que el eje X = 2,6 millones de lecturas de secuenciación y el eje Y = 4,8 millones de lecturas de secuenciación.

Figuras 14A-14B: Gráficos que muestran la predicción de la medición de las muestras C y D en base a las mediciones de las muestras A y B (intrasitio); en los que el eje X = 15,2 millones de lecturas de secuenciación y el eje Y = 4.9 millones de lecturas de secuenciación.

Figura 15: Gráfico que muestra la comparación entre laboratorios de las mediciones (intersitio), es decir, preparaciones de bibliotecas separadas secuenciadas en diferentes sitios (intersitio), en el que el eje X = 2,6 millones de lecturas de secuenciación y el eje Y = 0,4 millones de lecturas de secuenciación.

Figuras 16A-16B: Gráficos que muestran la curva ROC para detectar con exactitud los cambios basados en la figura 13 (resultado 4), que muestran la curva ROC para la expresión diferencial basada en la figura 14 (resultado 4).

Figura 17: Gráfico que muestra la preparación de una biblioteca basada en PCR en la que las concentraciones de dianas naturales convergen, reduciendo la profundidad de lectura requerida.

Figuras 18A-18B: Flujo de trabajo y análisis de datos de la secuenciación de ARN estandarizada (STARSEQ)

Figuras 19A-19C: STARSEQ reduce el sobremuestreo sin compresión de señales.

Figuras 20A-20B: STARSEQ reduce las lecturas de secuenciación requeridas hasta 10 000 veces.

Figuras 21A-21E: Rendimiento de STARSEQ con materiales de referencia ERCC.

Figuras 22A-22F: Rendimiento de STARSEQ con dianas de ADNc endógeno.

Figuras 23A-23B: Comparación de plataformas de STARSEQ con respecto a qPCR TaqMan y a RNA-Seq Illumina.

Figura 24: Representación de las diferencias entre las mediciones de TaqMan y de STARSEQ.

Figura 25: Representación de las diferencias entre las mediciones de RNA-Seq Illumina y de STARSEQ.

Figura 26: Rendimiento del ensayo.

Figura 27: "Negativo verdadero" en STARSEQ frente a TaqMan y RNA-Seq.

Figura 28: Desviación estándar de las mediciones de ERCC.

Descripción detallada

En el presente documento se describen métodos para evaluar ácidos nucleicos, y aplicaciones y métodos comerciales que emplean dichas composiciones y métodos. Algunos aspectos de la presente divulgación se refieren a mejoras con respecto a Willey y Willey et al., las patentes de EE.UU. n.º 5.043.390; 5.639.606; 5.876.978 y 7.527.930.

Métodos para evaluar un ácido nucleico

En el presente documento se describen métodos para evaluar cantidades de un ácido nucleico en una muestra. En algunos modos de realización, el método permite medir pequeñas cantidades de un ácido nucleico, por ejemplo, en el que el ácido nucleico se expresa en bajas cantidades en una muestra, en el que pequeñas cantidades del ácido nucleico permanecen intactas y/o en el que se proporcionan pequeñas cantidades de una muestra.

"Muestra biológica", tal como se usa en el presente documento, se puede referir a material recogido para su análisis,

por ejemplo, un hisopo de cultivo, una pizca de tejido, una extracción de biopsia, un vial de un líquido corporal, por ejemplo, saliva, sangre y/u orina, etc., que se obtiene con fines de investigación, diagnóstico o de otra índole de cualquier entidad biológica.

Muestra también se puede referir a cantidades típicamente recogidas en biopsias, por ejemplo, biopsias endoscópicas (usando cepillo y/o fórceps), biopsias aspirativas con aguja (que incluyen biopsias aspirativas con aguja fina), así como cantidades proporcionadas en poblaciones de células clasificadas (p. ej., poblaciones de células clasificadas por flujo) y/o materiales microdiseccionados (por ejemplo, tejidos microdisecados capturados por láser). Por ejemplo, las biopsias de lesiones sospechosas de cáncer en el pulmón, mama, próstata, tiroides y páncreas comúnmente se realizan mediante aspiración con aguja fina (FNA), la médula ósea también se obtiene mediante biopsia y los tejidos del cerebro, del embrión en desarrollo y de modelos animales se pueden obtener mediante muestras microdiseccionadas capturadas con láser.

"Entidad biológica", tal como se usa en el presente documento, se puede referir a cualquier entidad capaz de albergar un ácido nucleico, incluyendo cualquier especie, por ejemplo, un virus, una célula, un tejido, un cultivo in vitro, una planta, un animal, un sujeto que participa en un ensayo clínico y/o un sujeto al que se le diagnostica o trata una enfermedad o afección.

"Muestra", tal como se usa en el presente documento, se puede referir al material de muestra biológica usado para un ensayo, reacción, ejecución, estudio y/o experimento dado. Por ejemplo, una muestra puede comprender una alícuota del material de muestra biológica recogido, hasta e incluyendo toda la muestra biológica. Tal como se usa en el presente documento, los términos ensayo, reacción, ejecución, estudio y/o experimento se pueden usar indistintamente.

La muestra biológica recogida puede comprender menos de aproximadamente 100 000 células, menos de aproximadamente 10 000 células, menos de aproximadamente 5000 células, menos de aproximadamente 1000 células, menos de aproximadamente 500 células, menos de aproximadamente 100 células, menos de aproximadamente 50 células o menos de aproximadamente 10 células.

La valoración, evaluación y/o medición de un ácido nucleico se pueden referir a proporcionar una medida de la cantidad de un ácido nucleico en una muestra biológica y/o muestra, por ejemplo, para determinar el nivel de expresión de un gen. En algunos modos de realización, proporcionar una medida de una cantidad se refiere a detectar una presencia o ausencia del ácido nucleico de interés. En algunos modos de realización, proporcionar una medida de una cantidad se puede referir a cuantificar una cantidad de un ácido nucleico, por ejemplo, proporcionar una medida de la concentración o nivel de la cantidad del ácido nucleico presente. En algunos modos de realización, proporcionar una medida de la cantidad de ácido nucleico se refiere a enumerar la cantidad del ácido nucleico, por ejemplo, indicar un número de moléculas del ácido nucleico presente en una muestra. El "ácido nucleico de interés" se puede denominar ácido nucleico "diana" y/o un "gen de interés", por ejemplo, un gen que se está evaluando se puede denominar gen diana. El número de moléculas de un ácido nucleico también se puede denominar número de copias del ácido nucleico encontradas en una muestra y/o muestra biológica.

Como se usa en el presente documento, "ácido nucleico" se puede referir a una forma polimérica de nucleótidos y/o moléculas de tipo nucleótido de cualquier longitud. En ciertos modos de realización, el ácido nucleico puede servir como plantilla para la síntesis de un ácido nucleico complementario, por ejemplo, mediante incorporación complementaria de bases de unidades de nucleótidos. Por ejemplo, un ácido nucleico puede comprender ADN de origen natural, por ejemplo, ADN genómico; ARN, por ejemplo, ARNm, y/o puede comprender una molécula sintética, que incluye pero no se limita a ADNc y moléculas recombinantes generadas de cualquier manera. Por ejemplo, el ácido nucleico se puede generar a partir de síntesis química, transcripción inversa, replicación de ADN o una combinación de estos métodos de generación. El enlace entre las subunidades se puede proporcionar mediante fosfatos, fosfonatos, fosforamidatos, fosforotioatos o similares, o mediante grupos no fosfato, tales como, pero sin limitación, enlaces de tipo péptido utilizados en ácidos peptidonucleicos (PNA). Los grupos de enlace pueden ser quirales o aquirales. Los polinucleótidos pueden tener cualquier estructura tridimensional, que abarca moléculas monocatenarias, bicatenarias y de triple hélice que pueden ser, por ejemplo, ADN, ARN o moléculas híbridas de ADN/ARN.

Una molécula de tipo nucleótido se puede referir a un resto estructural que puede actuar sustancialmente como un nucleótido, por ejemplo, exhibiendo complementariedad de bases con una o más de las bases que se producen en el ADN o ARN y/o siendo capaz de incorporar bases complementarias. Los términos "polinucleótido", "molécula de polinucleótido", "molécula de ácido nucleico", "secuencia de polinucleótido" y "secuencia de ácido nucleico" se pueden usar indistintamente con "ácido nucleico" en el presente documento. En algunos modos de realización específicos, el ácido nucleico que se ha a medir puede comprender una secuencia correspondiente a un gen específico.

En algunos modos de realización, la muestra biológica recogida comprende ARN que se va a medir, por ejemplo, ARNm expresado en un cultivo de tejido. En algunos modos de realización, la muestra biológica recogida comprende ADN que se ha a medir, por ejemplo, ADNc transcrito inversamente a partir de transcritos. En algunos

modos de realización, el ácido nucleico que se ha a medir se proporciona en una mezcla heterogénea con otras moléculas de ácido nucleico.

El término "plantilla natural", tal como se usa en el presente documento, se puede referir al ácido nucleico obtenido directa o indirectamente de una muestra biológica que puede servir como una plantilla para la amplificación. Por ejemplo, se puede referir a moléculas de ADNc, que corresponden a un gen cuya expresión se ha a medir, en el que el ADNc se amplifica y se cuantifica.

El término "cebador" se refiere en general a un ácido nucleico capaz de actuar como un punto de iniciación de la síntesis a lo largo de una hebra complementaria cuando las condiciones son adecuadas para la síntesis de un producto de extensión del cebador.

Descripción general del método

La preparación de una biblioteca de secuenciación implica la combinación de algunas de las siguientes etapas o de todas ellas: 1) fragmentación de ácido nucleico; 2) clonación in vivo, que sirve para unir secuencias adaptadoras de ácidos nucleicos flanqueantes; 3) unión de adaptadores in vitro; 4) adición de adaptadores basada en PCR; y, 5) tecnología de tipo sonda de inversión unimolecular con o sin relleno de polimerasa, y unión de la sonda para capturar la secuencia mediante circularización, con el adaptador contenido dentro de la secuencia de la sonda.

La definición de "adaptador de ácido nucleico" es que el "adaptador de ácido nucleico" puede servir como cualquiera o todos de los siguientes: a) el sitio de reconocimiento del cebador de secuenciación, b) la secuencia de código de barras de nucleótidos para descircunvolucionar la muestra que fue preparada para secuenciación durante el análisis, y c) el sitio de ácido nucleico universal que permite la amplificación de múltiples plantillas, o la adición adicional de secuencias de cola de fusión a través de la amplificación.

La biblioteca de secuenciación preparada a partir de una o más de las etapas 1-5 anteriores se analiza entonces en un instrumento de secuenciación y se secuencia una muestra representativa de la biblioteca. Se cuenta el número de veces que se observa cada diana única de ácido nucleico y se evalúa la proporción relativa entre recuento de cada diana única de ácido nucleico. Esta proporción relativa, sin embargo, no representa la verdadera proporcionalidad de la abundancia entre cada diana única de ácido nucleico en la muestra original.

Esta pérdida de representación original es un artefacto técnico (por ejemplo, error, sesgo) de las etapas 1-5. Además, este error no es sistemático, es decir, no tiene la misma cantidad de sesgo, entre al menos los siguientes errores: i) etapas de preparación de la biblioteca (1-5 anteriores); ii) réplicas de la biblioteca de secuenciación; iii) diferentes tiempos de réplicas; iv) diferentes técnicos preparan la biblioteca; y/o, v) preparar la biblioteca en un laboratorio diferente.

Dado que este error no sistemático en cuanto a la proporción de ácido nucleico, en efecto, se dirige a los errores (i-v), cualquier comparación de los resultados entre las preparaciones de la biblioteca para la misma muestra es propensa a error, lo que limita la aplicación de la secuenciación como una herramienta de para medir copias de ácido nucleico de forma económica y fiable.

Un modo de realización descrito en el presente documento es un método que utiliza una mezcla de un número conocido (es decir, abundancia, concentración y/o cantidad) de moléculas de ácido nucleico patrón interno correspondientes a dianas únicas de ácido nucleico (también definidas como "diana natural" o DN) que se han de mezclar en una muestra de ácido nucleico antes de la preparación de la biblioteca para secuenciación, o antes de la secuenciación (si no se requiere la preparación de la biblioteca).

Cada diana de ácido nucleico es similar a su respectivo patrón interno, con la excepción de uno o más cambios en la secuencia de ácido nucleico. Estas diferencias entre la diana natural y el patrón interno son identificables con la secuenciación, y pueden incluir deleciones, adiciones o alteración del orden o la composición de los nucleótidos utilizados.

Mediante la introducción de patrones internos en una muestra de dianas de ácido nucleico antes de la preparación de la biblioteca, el error no sistemático introducido en las etapas 1-5 (así como el sesgo específico del instrumento secuenciador) es experimentado por la diana natural y por el patrón interno de manera similar.

Al final de la secuenciación, la proporción de eventos de secuenciación (es decir, observaciones, recuentos, lecturas) entre la diana natural y su respectivo patrón interno se evalúa, junto con el número original de entrada de moléculas de ácido nucleico patrón interno en la muestra antes de la preparación de la biblioteca, para determinar de manera cuantificable la cantidad original de cada diana natural en la muestra original antes de la preparación de la biblioteca y la secuenciación.

Dado que la inclusión del patrón interno controla por tanto el error y los cambios relativos en proporción entre dianas naturales durante las etapas 1-5 y la posterior secuenciación, el método descrito en el presente documento también

permite amplificar preferentemente (es decir, enriquecer) las dianas naturales de baja abundancia con respecto a las dianas naturales de mayor abundancia durante la preparación de la biblioteca. Esta amplificación o enriquecimiento preferente se puede aprovechar de modo que, al final de la preparación de la biblioteca de secuenciación, la proporción relativa entre cada diana natural única convergirá hacia una abundancia equimolar (es decir, uniforme) en la biblioteca. Esto da como resultado una cobertura más uniforme de la profundidad de secuenciación entre dianas naturales. Y, dado que el patrón interno también experimenta la amplificación o enriquecimiento preferente, este método permite determinar de forma cuantificable la cantidad original de cada diana natural en la muestra original antes de la preparación de la biblioteca.

En un ejemplo no limitante, por cada reducción de 10 veces en la profundidad de la proporción entre dianas naturales se consigue una reducción de aproximadamente 10 veces en el coste de la secuenciación directa, ya que se requieren 10 veces menos lecturas de secuenciación.

La adición de una mezcla de patrones de ácido nucleico antes de la preparación de la biblioteca de secuenciación (o antes de la secuenciación si no se requiere la preparación de la biblioteca) proporciona así una cuantificación exacta de las dianas naturales en el punto final con la secuenciación.

El uso de una mezcla estandarizada de patrones internos de ácido nucleico permite una comparación directa de los resultados entre laboratorios para el diagnóstico molecular de ácido nucleico y otros resultados de secuenciación cuantitativos.

Además, la adición adicional de patrones internos permite la convergencia de la abundancia de dianas naturales, reduciendo así los costes de la secuenciación directa por la normalización de la proporción de las abundancias de las dianas naturales entre sí.

La inclusión de una mezcla de ácido nucleico patrón interno de cantidad conocida (es decir, abundancia, concentración y/o número) durante la preparación de la biblioteca proporciona ciertas ventajas. Dado que se podría desconocer cuál de las etapas 1-5 o la secuenciación podría introducir el error, el presente método reduce este sesgo, permitiendo así la comparación de resultados entre bibliotecas y entre laboratorios y, al mismo tiempo, proporcionando la capacidad de reducir el coste de la secuenciación directa a través de la convergencia de concentración de las dianas de ácido nucleico.

En ciertos modos de realización, el método descrito en el presente documento incluye un número conocido de moléculas de patrón interno para cada gen que se ha a medir en la muestra de ácido nucleico antes de la secuenciación, o antes de la preparación de la biblioteca para la secuenciación.

La preparación de una mezcla estandarizada de patrones internos puede ser utilizada por múltiples laboratorios, aumentando así la fiabilidad de la medición de cada gen diana y aumentando la reproducibilidad de la medición entre experimentos y entre laboratorios. La medición del número de copias para cada ácido nucleico con respecto a un número conocido de copias de sus respectivas moléculas patrón interno dentro de una mezcla estandarizada de patrones internos, y el uso de misma mezcla estandarizada de patrones internos en diversos experimentos y laboratorios, aumenta así la fiabilidad y el control de calidad al controlar la variación introducida por la preparación de la biblioteca de secuenciación.

El método descrito en el presente documento utiliza una transcripción inversa específica de genes y/o una PCR para la preparación de la biblioteca para la cuantificación mediante secuenciación. La optimización de la PCR permite la multiplexación de hasta 100, 300, 500, 1000 o más genes para producir suficiente producto de PCR de cada gen diana para la cuantificación mediante secuenciación. La optimización de la PCR puede generar una convergencia de la representación inicial de los transcritos entre genes de 10 veces, 100 veces, 1000 veces, 10 000 veces o más, mientras se mantiene la capacidad de cuantificar la representación inicial relativa de los transcritos a través de la medición de cada gen con respecto a su patrón interno respectivo. Por lo tanto, la inclusión de un número conocido de copias de patrones internos en la muestra antes de la preparación de la biblioteca (o antes de la secuenciación si no se requiere la preparación de la biblioteca) controla los cambios posteriores en la representación de los transcritos. Ahora es posible optimizar la convergencia entre genes sin perder la información relativa a la representación inicial. Por ejemplo, puede haber una convergencia de más de 1000 veces, lo que resulta en una reducción del requisito de "lectura" de 10 000 000 a 10 000.

Además, en algunos de los métodos descritos en el presente documento en los que cada chip para un típico secuenciador de nueva generación permite realizar 10 millones de lecturas, este resultado permite aumentar el número de muestras analizadas/chip de 1 a 1000. Actualmente, dado que un chip para un secuenciador típico cuesta 1000 $, el coste del chip/muestra se reduce de aproximadamente 1000 $ a aproximadamente 1,00 $.

Además, los transcritos raros se puede medir con significación estadística. Por ejemplo, se puede determinar el número de copias de un ácido nucleico correspondiente a un transcrito génico, por ejemplo, el número de copias/célula, en el que el gen se expresa en un bajo número de copias. La enumeración de menos de aproximadamente 1000 moléculas puede permitir la medición de menos de aproximadamente 10 copias/célula de al

menos 100 transcritos génicos diferentes en una pequeña muestra biológica. Los métodos son capaces de medir y/o enumerar menos de aproximadamente 10 copias/célula de al menos 100 transcritos génicos diferentes en una pequeña muestra biológica.

En todavía algunos de los métodos descritos en el presente documento se pueden obtener más mediciones de una muestra biológica y/o muestra dada, por ejemplo, del tamaño normalmente utilizado para medir que escasas copias de un ácido nucleico correspondiente a un gen. Por ejemplo, la práctica de algunos métodos permite medir y/o enumerar menos de aproximadamente 100, menos de aproximadamente 50, menos de aproximadamente 20, menos de aproximadamente 10, menos de aproximadamente 8 o menos de aproximadamente 5 copias/célula de al menos aproximadamente 20, al menos aproximadamente 50, al menos aproximadamente 80, al menos aproximadamente 100, al menos aproximadamente 120, al menos aproximadamente 150 o al menos aproximadamente 200 ácidos nucleicos diferentes en una muestra, por ejemplo, correspondientes a diferentes transcritos génicos.

El material expresado puede ser endógeno a la entidad biológica, por ejemplo, los transcritos de un gen expresado de forma natural en un tipo de célula dado, o el material expresado que se ha a medir puede ser de naturaleza exógena. Por ejemplo, los métodos se pueden usar para cuantificar genes transfectados después de una terapia génica y/o un gen indicador en ensayos de transfección transitoria, por ejemplo, para determinar la eficacia de la transfección.

Ejemplos

El método de preparación de la biblioteca de NGS basada en PCR incorpora controles internos de amplificación (IAC) competitivos. Este método controla la mayoría de los errores no sistemáticos introducidos durante la preparación de la biblioteca de NGS y permite la comparación entre laboratorios de los datos de NGS cuantitativa (qNGS).

El IAC competitivo controla por lo tanto el error no sistemático en la preparación de la biblioteca de NGS basada en PCR al compartir sitios de cebado idénticos con una plantilla de ácido nucleico natural de interés a fin de imitar la cinética de la diana natural en la reacción de PCR y, por lo tanto, al controlar la variación en la eficiencia de la PCR específica de la diana.

En los métodos descritos en el presente documento, dado que el IAC competitivo experimenta la misma cinética que una plantilla de ácido nucleico natural, la relación proporcional de las lecturas de secuenciación de la diana natural con respecto a su IAC competitivo respectivo no cambia durante la preparación de la biblioteca de NGS.

Por otra parte, si se conoce la concentración de IAC competitivos colocado en la preparación de la muestra, ahora es posible calcular con exactitud la abundancia original de moléculas de ácido nucleico natural que estaba presente al comienzo de la preparación de la biblioteca de NGS.

Como un ejemplo, cuando varios laboratorios utilizan la misma mezcla de IAC competitivos en múltiples estudios diferentes, cada uno de los múltiples laboratorios ha mostrado que sus resultados son concordantes.

Por lo tanto, el uso de IAC competitivos en la preparación de la biblioteca de NGS basada en PCR permite análisis multiplexados altamente rentables de múltiples dianas de ácido nucleico en múltiples muestras con un alto grado de exactitud y reproducibilidad.

Un beneficio adicional de incorporar IAC competitivos es que los protocolos que dan como resultado la normalización (es decir, la convergencia) de cada diana natural hacia la concentración equimolar (es decir, uniforme), tales como PCR múltiple, se pueden implementar utilizando dicho método.

Se debe entender que con la normalización de las concentraciones de las plantillas, ahora es posible que la abundancia entre las plantillas de ácidos nucleicos naturales pueda variar en más de un millón de veces. En el pasado, la plantilla de ácido nucleico natural más altamente representada se sobremuestrearía innecesariamente y se secuenciaría diez millones de veces para secuenciar la plantilla de ácido nucleico menos representada (por ejemplo, al menos diez veces para detectar con exactitud un cambio de 2 veces (Potencia = 80 %); Tasa de error tipo 1 = 0,05)). Sin embargo, el uso del método de IAC competitivos descrito en el presente documento proporciona normalización en la representación de los analitos diana, pero aún retiene información cuantitativa de la representación original de las dianas de alta y baja abundancia con un número bajo de lecturas de secuenciación. La reducción en el sobremuestreo de dianas de ácidos nucleicos sobrerrepresentadas da como resultado un reducido coste y error de muestreo estocástico asociado con la secuenciación profunda.

Ejemplo 1

PCR múltiple con IAC competitivos para la preparación de la biblioteca de NGS y la medición posterior de la abundancia de ácidos nucleicos

Se obtuvieron grupos de valoración de ARN como material de referencia utilizado en el proyecto de control de calidad de secuenciación (SEQC) patrocinado por la FDA (en el que ya se ha medido la abundancia de ácidos nucleicos mediante múltiples plataformas de qPCR, Microarray y NGS en una variedad de condiciones).

Las bibliotecas de NGS se prepararon a partir del material de referencia sometido a transcripción inversa utilizando PCR múltiple en presencia de cebadores e IAC competitivos para 150 dianas génicas.

La preparación de la biblioteca de NGS se evaluó usando PCR múltiple con IAC competitivos para la reproducibilidad de la medición de la abundancia de ácidos nucleicos en sitios de ensayo individuales, entre laboratorios, y en diferentes plataformas de medición de ácidos nucleicos.

Los costes y ventajas de la PCR múltiple con IAC competitivos para la preparación de la biblioteca de NGS se compararon con un protocolo comúnmente empleado de preparación de bibliotecas de NGS basada en Illumina y en qPCR TaqMan para medir con exactitud la abundancia de ácidos nucleicos en un entorno clínico.

Métodos y resultados

Se diseñaron cebadores directos e inversos que corresponden a 101 regiones de pares de bases (es decir, amplicón) para cada uno de los 150 genes transcritos de manera única en el genoma humano. Cada cebador se diseñó con una temperatura de fusión uniforme de 68 °C. Cada cebador también contenía una secuencia de cola universal que se puede usar para PCR de múltiples plantillas, tal como se usa en la adición de secuencias de código de barras y secuencias de adaptador de secuenciación después de la PCR múltiple inicial. Estos cebadores fueron sintetizados por Integrated DNA Technologies (IDT) y se combinaron en una relación equimolar, y se diluyeron a una concentración de trabajo final de 50 nM de cada cebador. Una mezcla correspondiente de 150 controles internos de amplificación (IAC) competitivos, cada uno con 101 bases de longitud, fue sintetizada por Integrated DNA Technologies (IDT). Cada uno de los IAC competitivos contenía sitios de cebado específicos de la diana idénticos a sus respectivas dianas de plantilla de ácido nucleico natural. Dentro de estos sitios idénticos de cebado directo e inverso se encontraron seis cambios de nucleótidos en la porción interna de la secuencia, que permitían diferenciar un IAC competitivo de su diana natural correspondiente durante el análisis de datos posterior a la secuenciación.

IDT combinó cada IAC competitivo en una mezcla a una concentración aproximadamente equimolar con respecto al resto. Debido a que la mezcla de IAC competitivos puede no haberse obtenido en una relación 1:1 exacta, la abundancia absoluta de copias de cada uno de los IAC competitivos y su proporción en relación con el resto para cada uno de los 150 IAC competitivos se determinaron mediante valoración con respecto a una cantidad conocida de material de referencia de ADN genómico (ADNg). El material de referencia de ADN genómico puede servir como un reactivo de normalización, ya que entre cada una de las secuencias genómicas únicas existe una proporción uno a uno con respecto al resto en todo el genoma. Por lo tanto, las diferencias percibidas en la concentración de IAC competitivos cuando se valora frente a ADNg indican en realidad una diferencia sistemática en la proporción que existe entre los IAC competitivos en la mezcla. Esta diferencia sistemática viene determinada por la valoración frente a una cantidad fija de ADNg y siempre se aplica a futuros cálculos y mediciones obtenidas usando ese lote o mezcla particular de IAC (figura 1).

La figura 1 muestra la valoración de una mezcla de controles internos de amplificación (IAC) con respecto a una cantidad fija de entrada de ADNg de 100 000 copias en cada PCR múltiple. En el eje Y se representa la frecuencia o proporción de lecturas naturales observadas dividida por la suma de las lecturas naturales y sus respectivas lecturas de IAC competitivos. En el eje X se muestra la cantidad inicialmente estimada de cada diana en una mezcla aproximadamente equimolar de IAC competitivos. Se introdujeron 10 diluciones que van desde 10 000 000 de copias de cada IAC (log10 de concentración = 0) hasta 1000 (log10 de concentración = -2) copias en cada una de las 10 reacciones para generar la curva mostrada. De los 150 conjuntos de cebadores diseñados, IAC competitivos y las respectivas dianas naturales, 119 se valoraron con una bondad de ajuste ($R^2 > 0,95$). Más de un 95 % de los IAC competitivos estaba dentro de un intervalo de 10 veces el punto de equivalencia esperado, (Natural)/(Natural + IAC) = 0,5, cuando se diluyó a 100 000 copias (10 000 000 copias iniciales de IAC diluidas a 100 000, o Log10 de concentración de -2, es decir, 100 veces con respecto a 10 000 000). La nueva concentración sirvió como la concentración real para cada uno de los 119 ensayos realizados en la mezcla de IAC competitivos y sirvió como referencia de exactitud absoluta (es decir, exactitud verdadera) en µg.

Por lo tanto, después de probar los 150 ensayos contra la mezcla valorada de IAC con una cantidad fija de ADNg, se determinó que 119 de los 150 ensayos tenían suficientes características de rendimiento (representación de Hill, $R^2 >$ 0,95). Estas correcciones se aplicaron posteriormente a todas las mediciones futuras realizadas con esta mezcla de IAC.

La Fase III del proyecto MAQC, también conocido como el proyecto de Control de Calidad de Secuenciación (SEQC), generó cuatro grupos de dos tipos de muestras de ARN: ARN de referencia humano universal (UHRR) de Stratagene y un ARN de referencia de cerebro humano (HBRR) de Ambion. Los cuatro grupos incluyeron las dos muestras de ARN de referencia, así como dos mezclas de las muestras originales: Muestra A, 100 % de UHRR;

Muestra B, 100 % de HBRR; Muestra C, 75 % de UHRR y 25 % de HBRR; y Muestra D, 25 % de UHRR y 75 % de HBRR. Esta combinación de fuentes de ARN biológicamente diferentes y diferencias de valoración conocidas proporcionó un método para evaluar la exactitud de una plataforma basada en los genes expresados diferencialmente detectados. Se usaron alícuotas de diez (10) µg de estos grupos de ARN para las muestras A, B, C y D.

Cada uno de los materiales de referencia de los grupos de valoración de ARN (muestras A, B, C, D) se sometieron a transcripción inversa, como se describe en Canales et al., 2006, con la excepción de que la transcriptasa inversa Superscript III de Invitrogen se utilizó en lugar de la transcriptasa inversa MMLV y de que se colocó 1 µg de ARN en cada reacción de transcripción inversa. Además, la muestra A se transcribió de forma inversa dos veces a partir de dos preparaciones separadas de mezcla maestra de transcripción inversa, con el fin de determinar la varianza introducida por la transcripción inversa en la preparación de la biblioteca de secuenciación.

Se añadió un (1) µl de cada uno de estos 5 grupos de valoración de ARN transcritos de forma inversa de ADNc (muestras A-RT1, A-RT2, B, C y D) en 1 de 12 reacciones de PCR múltiple que contenían una mezcla de diluciones en serie de una mezcla de controles internos de amplificación (IAC) competitivos que representa 150 dianas. Estas 12 diluciones en serie de mezclas de IAC competitivos van desde 107 copias cargadas hasta 103. Se consumió un total de 12 µl de cada muestra durante la PCR múltiple, lo que corresponde a ~133 ng de ARN en total para cada muestra.

La concentración a la que el material natural estaba a la misma concentración que el IAC competitivo (es decir, el punto de equivalencia) para cada diana génica se determinó en cada material de referencia transcrito de forma inversa (muestras SEQC A-RT1, A-RT2, B, C y D) y se determinó usando la ecuación de Hill (figura 2).

El gráfico de la figura 2 muestra la valoración de una mezcla de controles internos de amplificación (IAC) con respecto a una cantidad fija de 100 000 copias de ADNg, u 11 ng de material de ADNc de ARN transcrito de forma inversa (muestras SEQC A-RT1, A-RT2, B, C y D), introducidos en cada PCR múltiple. En el eje Y se representa la frecuencia o proporción de lecturas naturales observadas dividida por la suma de las lecturas naturales y sus respectivas lecturas de IAC competitivos. En el eje X se muestra la cantidad inicialmente estimada de cada diana en una mezcla aproximadamente equimolar de IAC competitivos. Se introdujeron diluciones que van desde 10 000 000 de copias de cada IAC (log10 de concentración = 0) hasta 1000 (log10 de concentración = -2) copias en cada una de las 10 reacciones para generar la curva anterior.

Dado que las muestras C y D representan una valoración cruzada conocida entre muestras A y B, se evaluó la exactitud de la plataforma de genes expresados diferencialmente (figura 3). Los valores medidos para la muestra SEQC C-RT1 se compararon con los valores esperados (% de diferencia) en base a los valores de las muestras SEQC A y B. La diferencia porcentual entre la señal predicha C' y la señal del ensayo real C se usó como una indicación de la exactitud relativa (ER) del ensayo. Una puntuación de ER ΔC para un gen diana se definió como (C-C'/C'), respectivamente. La distribución de la diferencia porcentual de la puntuación de ER esperada para cada gen se presenta en un diagrama de cajas para qNGS estandarizada (n = 88) y RT-PCR estandarizada (n = 201). Los componentes del diagrama de cajas son: línea horizontal, mediana; caja, rango intercuartil; bigotes, rango intercuartil de 1,5×; cuadrados negros, valores atípicos.

Las dianas génicas (n = 88) medidas entre las muestras SEQC A-RT1, B, C se evaluaron para la expresión diferencial (ED) intergénica (figura 4A); la ED experimentaba un cambio de entre 1,5 y 3,0 veces, entre 2 y 3 veces, entre 3 y 5 veces, entre 5 y 10 veces o más de 10 veces El control para el cambio falso positivo o negativo se evaluó comparando SEQC A-RT1 frente a SEQC A-RT2. Las estadísticas resumidas que muestran la expresión diferencial en la muestra C en comparación con las esperadas basadas en las muestras A y B se muestran en la figura 4B - tabla 1.

El ensayo de reproducibilidad entre transcripciones inversas se muestra en la figura 5. Se midieron dos transcripciones inversas de la muestra SEQC A (RT1 frente a RT2) para la expresión de 119 dianas génicas que superaron con éxito los criterios de rendimiento de la figura 1. De las 119 dianas génicas, 97 tenían $R^2 > 0,95$ para el ajuste de la curva de la ecuación de la pendiente para determinar el punto de equivalencia y la concentración de cada diana (figura 2).

Los mismos datos que en la figura 3 y la figura 4 están representados en la figura 6. Mediciones esperadas en la muestra C (eje x) vs. mediciones observadas (eje y). La transcripción inversa de la muestra SEQC C se midió para la expresión de 119 dianas génicas que superaron con éxito los criterios de rendimiento de la figura 1. De las 119 dianas génicas, 88 tenían $R^2 > 0,95$ para el ajuste de la curva de la ecuación de la pendiente para determinar el punto de equivalencia y la concentración de cada diana (eje y) (figura 2).

La convergencia y la mayor uniformidad de 97 dianas génicas de la figura 5 se muestran en la figura 7. En el eje X se representan los datos de la figura 5 en proporción a la plantilla de mayor abundancia. En el eje Y se representa la proporción real de lecturas de secuenciación en las que está la diana génica en proporción a la plantilla de

secuencia más alta. Tenga en cuenta que la medición y la exactitud no están comprimidas (figuras 4-6); sin embargo, un 75 % de las dianas génicas están dentro de una abundancia de lecturas de secuenciación de 10 veces entre sí. Es decir, la profundidad de secuenciación disminuyó desde aproximadamente 1000 veces hasta 10 veces. Esto representa una disminución de 100 veces en el coste de la secuenciación directa.

La curva ROC para detectar un cambio de > 3 veces para la RNA-Seq utilizando la plataforma Illumina de Bullard et al. BMC Bioinformatics 2010, 11:94 se muestra en la figura 8. En comparación con la figura 4B, esta curva ROC representa una exactitud de ~75 % de RNA-Seq para detectar un cambio > 3 veces. Mientras tanto, la qNGS estandarizada descrita en el presente documento tiene una exactitud superior a un 97 % (figura 4B). Se debe observar que el método de qNGS estandarizada utilizó una profundidad de secuenciación de 10 veces para detectar con exactitud un cambio de 3 veces en una diferencia proporcional de 1000 veces entre dianas naturales. Por el contrario, la secuenciación de ARN tradicional requeriría 100 veces más lecturas para llegar a una exactitud similar. En un ejemplo se utilizaron 5 millones de lecturas de secuenciación para cuantificar con exactitud 97 genes usando el método de qNGS estandarizada. En comparación, la secuenciación de ARN tradicional habría requerido más de 500 millones de lecturas para una cuantificación exacta.

Ejemplo 2

La secuenciación cuantitativa después de la preparación de la biblioteca basada en PCR con mezclas de patrones internos ha mejorado el rendimiento analítico y reducido el coste

Los sesgos no sistemáticos introducidos durante la preparación de bibliotecas de secuenciación de nueva generación (NGS) como la fuente principal de variación técnica han impedido la aplicación de la NGS para medir la abundancia de ácidos nucleicos en el entorno clínico.

Los costes de los diagnósticos clínicos actuales por qPCR están vinculados al coste de la química (normalmente fluorescente) que utilizan, y linealmente relacionados con el número de dianas de ácido nucleico que están interrogando. Además, cada diana de ensayo requiere un recipiente de reacción separado y múltiples controles, que pueden llegar a ser prohibitivamente costosos. Estos costes acumulativos evitan la aparición de diagnósticos clínicos más complejos basados en la medición de múltiples dianas de ácidos nucleicos. Las alternativas más rentables para la medición de la abundancia multiplexada de dianas de ácidos nucleicos no son tan flexibles para lograr nuevas dianas de ensayo en línea de forma rentable sin alterar los paneles de genes existentes, o no son susceptibles de estandarización y reproducibilidad entre sitios de datos cuantitativos. Mientras que la NGS es susceptible de análisis cuantitativos rentables altamente multiplexados de múltiples muestras de pacientes y dianas de ácidos nucleicos, existe la necesidad de una forma eficiente de permitir la comparación de resultados cuantitativos de NGS entre sitios, y de evitar la necesidad de secuenciación profunda para medir con exactitud la abundancia de ácidos nucleicos.

En este ejemplo se especifica un protocolo de preparación de una biblioteca de NGS basada en PCR que incorpora mezclas de controles internos de amplificación (IAC) competitivos (es decir, patrones internos) controlado para la mayoría de los sesgos introducidos durante la preparación de la biblioteca de NGS, que permite a los laboratorios clínicos ofrecer paneles de diagnóstico moderadamente complejos rentables a partir de datos NGS cuantitativa.

Se obtuvieron grupos de valoración de ARN como material de referencia utilizado en el proyecto de control de calidad de secuenciación (SEQC) patrocinado por la FDA (muestras A, B, C y D). Dado que las muestras de ARN C y D del proyecto SEQC representan una valoración cruzada conocida entre las muestras A y B del proyecto SEQC, es posible comparar el valor de expresión SEQC con los valores de expresión medidos y esperados para determinar la exactitud del método. Utilizando PCR múltiple con cebadores y IAC competitivos para 150 dianas génicas, se prepararon bibliotecas de NGS a partir de: 1) ADNg para evaluar el rendimiento analítico general, y 2) ADNc de material de referencia SEQC de transcripción inversa para determinar la exactitud.

Resultados

El uso de ADNg mezclado con mezclas de IAC competitivos valorados en serie como entrada, se observó un intervalo dinámico lineal de más de 10^6 órdenes de magnitud, con un R^2 promedio = 0,995 (0,993 - 0,997; IC del 95 %). El coeficiente de correlación de los valores esperados vs. observados para la muestra C fue R^2 = 0,96, y la muestra D fue R^2 = 0,94, con una exactitud determinada por la curva ROC para detectar un cambio de 3 veces de un 97 % (95 - 99 %; IC del 95 %). El coeficiente de correlación entre sitios de las mediciones basadas en tan solo 400 000 lecturas de secuenciación fue R^2 = 0,92 en un intervalo dinámico lineal de ~10^5 órdenes de abundancia entre dianas naturales.

El método descrito en el presente documento supera fuentes clave de sesgo no sistemático introducido durante la preparación de la biblioteca de NGS. Esto permite obtener resultados de NGS cuantitativos reproducibles entre laboratorios y entre plataformas, y un camino claro hacia la aprobación regulatoria de aplicaciones de diagnóstico clínico.

El método descrito en el presente documento (una NGS con controles internos de amplificación (IAC)) proporciona reproducibilidad intrasitio e intersitio de datos de secuenciación cuantitativa de nueva generación (qNGS). El método descrito en el presente documento también reduce la necesidad de secuenciación profunda y, por lo tanto, el coste de la secuenciación directa, al converger el número de lecturas requerido para secuenciar adecuadamente las dianas de ácidos nucleicos tanto raras como de alta abundancia.

La figura 9 proporciona una ilustración esquemática de una mezcla maestra de PCR con una mezcla de controles internos de amplificación (IAC). El IAC sirve como una referencia intersitio entre bibliotecas. El IAC es estable durante un período de tiempo prolongado (por ejemplo, durante años). La mezcla de IAC controla el sesgo de PCR y está presente en una concentración conocida. La mezcla de cebadores específicos de la diana incluye cientos de dianas por cada reacción. Los cebadores específicos de la diana contienen una cola universal.

Las figuras 10A-10B son gráficos que muestran la valoración de una mezcla de controles internos de amplificación contra ADNg y ADNc de SEQC. La representación está en el formato de una curva dosis-respuesta para la inhibición de un sistema enzimático. La enzima es la Taq polimerasa. El inhibidor es la concentración de los controles internos de amplificación (IAC) competitivos. La dosis-respuesta se mide como la proporción de lecturas de secuenciación observada para la diana de ADN genómico (ADNg) natural o la diana de ADN complementario (ADNc) natural frente a la suma de las lecturas de secuenciación natural y de IAC. La gráfica de ADNg representa 119 de las 150 dianas génicas diseñadas (~80 % de la tasa de éxito del diseño del ensayo). El coeficiente de correlación promedio para ajustar una ecuación de Hill de la pendiente fijada con tres parámetros para cada uno de los 119 ensayos fue R^2 = 0,995 (0,993 - 0,997, IC del 95 %).

La CI50 (concentración inhibitoria del 50 %) promedio fue de $10^{4,98}$ con el número de copias de ADNg de entrada siendo 10^5. Por lo tanto, la valoración de una mezcla de controles internos de amplificación proporciona una exactitud verdadera, no relativa, de la medición de copias de una mezcla compleja de ácidos nucleicos.

En las figuras 10-11 (*resultado 1*), la gráfica de ADNc representa 110 de 119 ensayos de dianas génicas de trabajo. Nueve (9) ensayos tenían una profundidad de lectura insuficiente de al menos 1 lectura de secuenciación tanto para la diana natural como para el control interno de amplificación. La CI50 (concentración inhibidora del 50 %) promedio se determinó para cada diana de ácido nucleico en las muestras SEQC A, B, C y D en una variedad de condiciones y se usó en ejemplos posteriores. La comparación de los resultados se realizó para:

La figura 12 (*resultado 2*) que muestra la misma secuenciación de réplicas de preparación de biblioteca (intrasitio), en la que el eje X = 1,8 millones de lecturas de secuenciación y el eje Y = 3,0 millones de lecturas de secuenciación.

La figura 13 (*resultado 3*) que muestra la preparación de la biblioteca separada secuenciada (intrasitio), en la que el eje X = 2,6 millones de lecturas de secuenciación y el eje Y = 4,8 millones de lecturas de secuenciación.

Las figuras 14A-14B (*resultado 4*) que muestra la medición predictiva de las muestras C y D en base a las mediciones de las muestras A y B (intrasitio); en la que el eje X = 15,2 millones de lecturas de secuenciación y el eje Y = 4,9 millones de lecturas de secuenciación.

La figura 15 (*resultado 5*) que muestra la comparación entre laboratorios de mediciones (intersitio), es decir, preparaciones de bibliotecas separadas secuenciadas en diferentes sitios (intersitio) en la que el eje X = 2,6 millones de lecturas de secuenciación y el eje Y = 0,4 millones de lecturas de secuenciación.

Las figuras 16A-16B (*resultado 6*) que muestra la curva ROC para detectar con exactitud los cambios basados en la figura 13 (resultado 4), que muestran la curva ROC para la expresión diferencial basada en la figura 14 (resultado 4).

La figura 17 (*resultado 7*) que muestra la preparación de una biblioteca basada en PCR en la que las concentraciones de dianas naturales convergen, reduciendo la profundidad de lectura requerida. La convergencia de las concentraciones de amplicón de plantilla natural durante la preparación de la biblioteca basada en PCR reduce el número de lecturas de secuenciación para secuenciar adecuadamente todas las dianas. Los controles internos de amplificación proporcionan el punto de referencia necesario al comienzo de la preparación de la biblioteca de secuenciación basada en PCR para medir con exactitud cada diana de ácido nucleico a pesar de la convergencia de la concentración de las plantillas (véanse las figuras 12-16 - Resultados 2-6). En este ejemplo, la profundidad de secuencia directa se reduce en 1000 veces, y todas las dianas están dentro de 100 veces entre sí.

También se ha de entenderse que está dentro del alcance contemplado de la presente divulgación que los métodos descritos en el presente documento incluyen el uso de paneles clínicos de complejidad moderada basados en la preparación de una biblioteca de NGS basada en PCR con controles internos de amplificación. Ejemplos no limitantes incluyen paneles para: prueba de riesgo de cáncer de pulmón (15 genes); prueba de diagnóstico del cáncer de pulmón (4 genes); prueba de quimiorresistencia al cáncer de pulmón (20 genes) y prueba de transcrito de fusión BCR-ABL (2 genes).

Ejemplo 3

Secuenciación de ARN estandarizada (STARSEQ)

Se evaluó la secuenciación de ARN estandarizada (STARSEQ) usando dos materiales de referencia separados: 1) ADN genómico (ADNg) derivado de la sangre de un individuo fenotípicamente normal (muestra desidentificada 723) del Centro Médico de la Universidad de Toledo (UTMC) de acuerdo con un protocolo aprobado por la junta de revisión institucional del UTMC, y 2) cuatro muestras de ARN de referencia (A, B, C y D) proporcionadas por el proyecto de control de calidad de secuenciación (SEQC) patrocinado por la FDA (anteriormente consorcio MAQC). La muestra A consiste en ARN de referencia humano universal obtenido de Stratagene. La muestra B consiste en ARN de referencia del cerebro humano obtenido de Ambion. Para el proyecto SEQC, las muestras A y B se combinaron luego con las mezclas 1 y 2 de enriquecimiento de ARN de control del consorcio ERCC (External RNA Controls Consortium) de Ambion, respectivamente, para lograr una concentración final de un 2 % en las muestras A y B en base a la concentración total de ARN

Cada mezcla de enriquecimiento de controles de ARN ERCC contiene los mismos controles que abarcan un intervalo dinámico mayor que 10^6, pero en diferentes formulaciones. Dentro de cada mezcla de formulación hay 4 subgrupos que exhiben diferencias de abundancia conocidas entre la mezcla 1 y 2; diferencia de 0,5x, 0,67x, 1,0x y 4,0x veces. Las muestras A y B se combinaron luego en mezclas proporcionales 3:1 y 1:3 para crear las muestras C y D, respectivamente. El material de "referencia" de ADNg representa una muestra en la que la mayoría de las dianas endógenas se encuentran en una proporción muy cercana de 1:1 entre sí. Considerando que las muestras A-D representan una mezcla compleja de dianas de ARN sintéticas (controles ERCC) y endógenas en proporciones conocidas que se pueden usar como indicadores reales para evaluar las características de rendimiento analítico de un método en un intervalo dinámico de abundancias de más de 10^6 veces.

Transcripción inversa de materiales de referencia de ARN

Diez microgramos de cada uno de los materiales de ARN de referencia de las muestras A-D a una concentración de 1 µg/µl se obtuvieron del proyecto SEQC patrocinado por la FDA (fda.gov/ScienceResearch /BioinformaticsTools /Microarray Quality Control Project). Para cada muestra se transcribieron de forma inversa dos alícuotas de 2 µg de ARN. Cada reacción de transcripción inversa tuvo lugar en un volumen de 90 µl utilizando el protocolo del fabricante para la transcripción inversa de Superscript III (Life Technologies) y el cebado con oligo(dT). Después de la transcripción inversa, los dos productos de ADNc de 90 µl para cada muestra se combinaron en un solo volumen de 180 µl (transcripción inversa 1; RT1). Para la muestra A, se transcribió de forma inversa un conjunto adicional de dos alícuotas de 2 µg de ARN usando una mezcla maestra separada (transcripción inversa 2; RT2).

Selección de dianas de ensayo STARSEQ

El consorcio MAQC (MicroArray Quality Control) seleccionó previamente una lista de 1297 genes para evaluar el rendimiento de plataformas de qPCR múltiples y micromatrices. De esta lista, se seleccionaron 150 dianas endógenas para desarrollar ensayos STARSEQ. Estos 150 ensayos se eligieron, en parte, porque las dianas génicas que representan se expresan en un rango dinámico mayor que 10^6. Estos reactivos se usaron para medir la proporción absoluta y relativa de cada diana génica en ADNg y muestras de ARN de referencia de transcripción inversa A-D. Además, 28 de las 92 dianas del consorcio ERCC (External RNA Control Consortium) también se seleccionaron para desarrollar ensayos STARSEQ.

Diseño y síntesis de cebadores STARSEQ

Los cebadores de PCR directa e inversa se diseñaron para las regiones correspondientes de amplicón de 101 pb para cada uno de 150 genes transcritos de forma única en el genoma humano y 28 dianas ERCC. Cada conjunto de cebadores directos e inversos se diseñó con una temperatura de fusión uniforme de 68 ºC utilizando el software Primer3 (Untergasser et al, NAR, 2012). Con el fin de minimizar el cebado fuera de la diana, se verificó la especificidad del par de cebadores usando GenomeTester 1.3 para identificar cualquier amplicón adicional de menos de 1000 pb de tamaño. Cada cebador también contiene una secuencia de cola universal no presente en el genoma humano, que se puede usar para la adición por PCR multiplantilla de códigos de barras y adaptadores de secuenciación específicos de la plataforma. Las colas universales directas son idénticas en cuanto a los adaptadores de secuencia utilizados para la extensión de cebador en matriz (APEX-2), mientras que la secuencia inversa de cola es la misma que la directa, con la excepción de las últimas cuatro bases 3', que permiten la direccionalidad durante la secuenciación. Los cebadores específicos de la diana con colas universales para las 150 dianas endógenas y las 28 dianas ERCC fueron sintetizados por Integrated DNA Technologies (IDT) y Life technologies, respectivamente. Se creó un grupo de cebadores para dianas endógenas o de ERCC combinando cebadores sintetizados en una relación equimolar y diluyendo a una concentración de trabajo final de 50 nM para cada cebador en tampón Tris-EDTA diluido.

Diseño y síntesis de la mezcla de patrones internos competitivos de STARSEQ

Cada patrón interno (PI) competitivo de 101 pb se diseñó para retener sitios de cebado específicos de diana idénticos a su diana de ácido nucleico natural respectivo (figuras 18A-18B). Dentro de estos sitios de cebado idénticos hay seis cambios de nucleótidos, para poder diferenciar un SI competitivo de su diana natural correspondiente durante el análisis de datos posterior a la secuenciación. Los 150 PI competitivos correspondientes a las dianas endógenas fueron sintetizados por Integrated DNA Technologies (IDT), y los 28 PI competitivos correspondientes a las dianas de ERCC fueron sintetizados por Life Technologies.

Para las 150 plantillas de PI competitivos correspondientes a las dianas endógenas, la concentración se midió por densidad óptica en IDT, y posteriormente se combinó en una relación molar estequiométrica 1:1 basada en estas mediciones. La concentración de cada PI se determinó empíricamente por valoración cruzada de la mezcla en relación con una entrada fija de ADNg de 100 000 copias (ID 723). En el ADNg de un individuo fenotípicamente sano, ahora se cree que la mayoría de los loci estarían en una proporción 1:1 entre sí, proporcionando un material de referencia razonable y rentable para determinar la concentración real para cada plantilla de PI competitivo.

Para las 28 plantillas de PI competitivos correspondientes a las dianas de ERCC, no existe dicho material de referencia para la normalización. Por lo tanto, cada patrón se amplificó por separado con cebadores directos e inversos (sin secuencias universales), se purificó en columna (kit de purificación QIAquick PCR), se visualizó y cuantificó para un solo pico a 101 bases en un Bioanalizador Agilent 2100 usando chips de ADN con reactivos del kit DNA 1000 de acuerdo con el protocolo del fabricante (Agilent Technologies Deutschland GmbH, Waldbronn, Alemania). A continuación, se combinaron patrones cuantificados en una relación molar estequiométrica 1:1 para crear una mezcla concentrada de patrones internos (PI). Tanto las mezclas de dianas endógenas como de dianas de ERCC de PI competitivos se diluyeron luego en serie a concentraciones de trabajo y se usaron en todos los experimentos posteriores como una mezcla de referencia para cuantificar copias absolutas de cada transcrito en las muestras A-D (figuras 18A-18B).

PCR competitiva múltiple con cebadores específicos de diana con cola universal

Para cada reacción en cadena de la polimerasa (PCR) competitiva múltiple se preparó un volumen de reacción de 10 µl que contenía: 1 µl de plantillas naturales, 1 µl de mezcla de PI competitivos a concentraciones variables de entrada, 1 µl de mezcla de cebadores correspondiente, 1 µl de dNTP 2 mM, 1 µl de tampón de reacción 10x de Idaho Technology con $MgCl_2$ 30 mM, 0,1 µl de Taq polimerasa GoTaq Hot Start de Promega (5 u/µl) y 4,9 µl de agua libre de RNAsa (figura 18A). El ADN genómico se añadió a 10 reacciones de PCR múltiple separadas que contenían una mezcla diluida en serie de mezcla de PI competitivos que representaba 150 dianas endógenas. Estas 10 diluciones representan una serie de diluciones de 3 veces de una mezcla de PI que varían en abundancia desde 2 x 10^6 a 10^3 copias cargadas. El ADNc de las muestras A-D para RT1 se añadió a 5 reacciones de PCR múltiple separadas que contenían una mezcla diluida en serie de una mezcla de PI competitivos que representa 28 dianas de ERCC. Estas 5 diluciones representan una serie de diluciones de una mezcla de PI, con 10^6, 10^5, 10^4, 10^3 y 300 copias cargadas. Los ARN de transcripción inversa para las muestras A (RT1 y RT2), B, C y D se añadieron a 12 reacciones de PCR múltiple separadas que contenían una mezcla diluida en serie de una mezcla de PI competitivos que representaba 150 dianas endógenas. Estas 12 diluciones representan una serie de diluciones de 3 veces de una mezcla de PI que varían en abundancia desde 6 x 10^7 a 3,4 x 10^2 copias cargadas. Se consumió un total de 17 µl de cada muestra de ADNc durante la PCR competitiva múltiple, lo que corresponde a ~377 ng de ARN para cada muestra.

Flujo de trabajo y análisis de datos de la secuenciación de ARN estandarizada (STARSEQ)

La figura 18A (DN = diana natural (por ejemplo, ADNc, ADNg, etc.); PI = Patrón interno) muestra una molécula de ADN monocatenario o bicatenario que a) es homóloga a una diana natural específica en las secuencias del cebador y por lo tanto compite por la amplificación con la diana natural, pero b) contiene una o más sustituciones de bases internas en los sitios del cebador y por lo tanto es distinguible de la diana natural. La plantilla de PI para cada gen está en una relación fija con relación al PI para los otros genes en una mezcla de patrones internos.

La figura 18B muestra que la relación proporcional entre las dianas naturales en la muestra original se conserva durante la amplificación y la secuenciación porque a) la competencia entre cada DN y su PI respectivo conserva la concentración original para cada DN, y b) los PI están en una relación fija relativa entre sí La determinación de la abundancia de la diana natural en la muestra original se obtiene multiplicando la relación de recuentos de secuenciación para DN y PI (DN:PI) por la concentración de patrón interno (PI) cargado en la preparación de la biblioteca de amplicones (es decir, determinación del punto de equivalencia). No se muestran las dianas naturales para los que no se pudieron medir los valores en al menos tres puntos de dilución. Figura 18B - panel superior: muestra la linealidad de la valoración cruzada de la mezcla de patrones internos competitivos con una cantidad constante de ADN genómico (ADNg) para 123 dianas. Las líneas punteadas representan un intervalo de predicción del 95 % para los valores de relación DN:PI. Figura 18B - panel central: muestra la linealidad de la valoración cruzada de la mezcla de patrones internos competitivos con una cantidad constante de 26 dianas naturales de ERCC de las muestras A, B, C y D. Cada diana de ERCC está a una concentración diferente que abarca un intervalo dinámico mayor que 10^6 en abundancia. Figura 18B - panel inferior: muestra la linealidad de la valoración cruzada de la mezcla de patrones internos competitivos con una cantidad constante de dianas naturales de ADNc endógeno de

las muestras A, B, C y D (mismas dianas que las evaluadas en ADNg, panel superior).

PCR Touchdown con PCR competitiva múltiple

El aumento del nivel de PCR múltiple requiere una disminución proporcional en la concentración de los cebadores usados. La disminución de la concentración de cebador tiene dos efectos predominantes en la PCR múltiple: 1) reduce la formación de productos de dímeros de cebadores, y 2) estabiliza la formación de amplicones en etapas tempranas impidiendo que los dNTP se conviertan en un reactivo limitado (método de menos cebador). Este último efecto es importante, ya que permite que todas las plantillas de diana alcancen la fase de meseta y, en presencia de PI competitivo, reduce drásticamente la sobremuestreo / secuenciación de dianas de alta abundancia sin compresión de señales (figuras 19A-19C, figuras 20A-20B).

STARSEQ reduce el sobremuestreo sin compresión de señales

La figura 19A representa dos dianas naturales (DN) en una muestra hipotética de ADNc. Una diana natural es de alta abundancia, 10^8 copias (DN "Abundante"), mientras que la otra es de baja abundancia, 10^2 copias (DN "Rara"), lo que representa una diferencia de un millón de veces en la abundancia entre las dianas. Esta muestra hipotética de ADNc se combina con una mezcla de patrones internos (PI) con una relación fija de concentraciones de 10^5 copias.

La figura 19B representa la preparación de la biblioteca por PCR competitiva múltiple para la figura 19A. Los diagramas de amplificación de PCR para DN "Abundante" y "Rara" se separan por motivos de claridad, pero ocurren en la misma reacción. Durante la PCR competitiva múltiple, cada diana natural compite por igual con su patrón interno competitivo respectivo para dNTP, polimerasa y una concentración limitante de cebadores. Debido a que la concentración inicial de cada par de cebadores de diana es la misma, cada reacción competitiva se estabilizará alrededor de la misma concentración del punto final (~10^9 copias).

En la figura 19C, la competencia igual entre cada DN y PI respectivo conserva la relación proporcional entre las dianas naturales en la muestra original, lo que permite la medición de la abundancia de la diana natural sin compresión de señales. Sin embargo, un intervalo de plantillas de 10^6 se reduce a 10^3 después de la preparación de la biblioteca por PCR competitiva múltiple, dando como resultado una reducción de 1000 veces en el sobremuestreo de la diana de alta abundancia.

La mezcla de una muestra de dianas naturales en múltiples relaciones con la mezcla de PI (figura 18A) da como resultado un mayor grado de uniformidad en la concentración de plantillas que el que se puede obtener con la adición de un solo patrón interno (figura 19A).

STARSEQ reduce las lecturas de secuenciación requeridas hasta 10 000 veces

La figura 20A muestra los datos reales de secuenciación proporcional para las dianas de ADNc de ERCC (n = 104) y endógenas (n = 400). El eje X representa la abundancia proporcional de cada diana en una preparación de biblioteca normalizada para la diana de menor abundancia (establecida en 10°). El eje Y está en unidades de lecturas de secuenciación proporcional (cobertura) requeridas para secuenciar la diana de menor abundancia al menos una vez.

La figura 20B es una tabla resumen de la figura 20A, en la que el número de lecturas de secuenciación representa la suma de todas las lecturas de secuenciación para observar todas las dianas al menos una vez. El número requerido de lecturas de secuenciación de ARN tradicional se calcula en base a una supuesta relación 1:1 entre las copias de diana presentes en la biblioteca y la secuencia de cobertura requerida. La reducción en las lecturas de secuenciación requeridas conseguida por STARSEQ es el cociente entre las lecturas de secuenciación de ARN tradicional y las lecturas de secuenciación de STARSEQ.

Sin embargo, hay un límite al que se pueden diluir los cebadores y aun así amplificar con éxito las dianas de interés. Este límite se puede reducir a través de varios enfoques: 1) aumentar la temperatura de fusión del cebador, y 2) aumentar el tiempo durante el cual se produce la hibridación para permitir la eventual unión del cebador. Ambas soluciones pueden exacerbar el cebado fuera de la diana. Este obstáculo aparente que se muestra ahora en el presente documento se puede remediar mediante el uso de un protocolo de PCR touchdown modificado. En este protocolo se incorporan altas temperaturas de hibridación durante los ciclos iniciales de PCR para aumentar la rigurosidad de la unión del cebador, lo que reduce el cebado fuera de la diana. En ciclos subsiguientes, la temperatura de hibridación se reduce gradualmente dando como resultado un rendimiento aumentado una vez que se ha formado suficiente producto específico durante ciclos de alta rigurosidad anteriores. Usando este marco, se desarrolló el siguiente protocolo: Cada mezcla de reacción competitiva múltiple se realizó en un termociclador de aire RapidCycler (Idaho Technology, Inc., Idaho Falls, Idaho) en condiciones de PCR touchdown modificadas con baja concentración de cebador: 95 °C / 3 min (activación Taq); 5 ciclos de 94 °C / 30 s (desnaturalización), 72 °C / 4 min (hibridación) y 72 °C / 15 s (extensión); repetir 5 ciclos con temperatura de hibridación disminuida en 1 °C hasta 71 °C; iterar una disminución de 1 °C y 5 ciclos hasta que la temperatura de hibridación sea de 64 °C (un total de 45 ciclos).

En modos de realización particulares se usa la Taq polimerasa Hot Start, ya que el cebado fuera de diana y la actividad enzimática son suficientemente altas durante la preparación de la reacción para que, de lo contrario, solo se viera producto de dímero de cebador.

Rendimiento de STARSEQ con materiales de referencia ERCC

La figura 21A muestra la abundancia de señal medida de las dianas de ERCC en las muestras A, B, C y D. Los puntos representan la mediana de las mediciones de ERCC de esas preparaciones de biblioteca con al menos 15 lecturas de secuenciación tanto para DN como para PI. Las unidades del eje X se derivan de la bibliografía de productos de Ambion para la concentración conocida de controles de enriquecimiento de ERCC, protocolos de preparación de material de proyecto de SEQC y un rendimiento de transcripción inversa supuesto de un 100 % para cada diana.

La figura 21B muestra gráficos de diferencias de datos de la figura 21A ordenados numéricamente por ID de ERCC. Cada diana de ERCC representada se midió al menos una vez en las cuatro muestras A-D. Con fines de claridad, ERCC-170 se resalta en naranja en la figura 21A y la figura 21B.

En la figura 21C, las muestras C y D representan una mezcla 3:1 y 1:3, respectivamente, de ARN total de las muestras A y B. Estas relaciones se usaron para calcular las mediciones esperadas para las muestras C y D (eje x) a partir de las mediciones de A y B, y se representan frente a mediciones reales de las muestras C y D (eje y) (n = 52).

En la figura 21D, los puntos representan la desviación estándar en las mediciones de las dianas de ERCC en muestras SEQC A, B, C y D, para los ensayos con al menos dos puntos de dilución de PI que tenían al menos 15 lecturas de secuenciación tanto para DN como para PI. La línea roja representa la desviación estándar esperada basada en una distribución de muestreo de Poisson más una desviación estándar de replicación técnica de referencia de 0,08.

La figura 21E muestra curvas ROC para detectar el cambio con la correspondiente área bajo la curva (ABC) con intervalos de confianza del 95 %. Las curvas ROC se derivan de la comparación de subgrupos de relación diferencial de dianas de ERCC en muestras: A frente a B, A frente a C, A frente a D, B frente a C, B frente a D y C frente a D. Los resultados para el cambio de 1,1 veces representan un intervalo de subgrupos de relación diferencial [1,05 - 1,174] (controles n = 100, pruebas n = 96); 1,25 [1,175 - 1,374] (controles n = 163, pruebas n = 163); 1,5 [1,375 - 1,74] (controles n = 229, pruebas n = 227); 2,0 [1,75 - 2,49] (controles n = 229, pruebas n = 223); ≥4,0 [2,5 - 10,0] (controles n = 286, prueba n = 290).

Rendimiento de STARSEQ con dianas de ADNc endógenos

La abundancia absoluta de señales de dianas de ADNc en la muestra A en unidades de copias por preparación de biblioteca se midieron en días separados, diferentes sitios (OU = Universidad de Ohio; UTMC = Universidad de Toledo Medical Center) y diferentes preparaciones de transcripción inversa (RT1 y RT2). Los puntos representan la mediana de las mediciones de ERCC de esas preparaciones de biblioteca con al menos 15 lecturas de secuenciación tanto para DN como para PI. La figura 22A muestra el efecto interdía (n = 88). La figura 22B muestra el efecto interdía e intersitio (n = 81). La figura 22C muestra el efecto interdía e interbiblioteca (n = 92). La figura 22D muestra el efecto interdía, intersitio e interbiblioteca (n = 80). Las figuras 22E-22F muestran que las muestras C y D representan una mezcla 3:1 y 1:3, respectivamente, de ARN total de las muestras A y B. Estas relaciones se usaron para calcular las mediciones esperadas para las muestras C y D (eje x) a partir de las mediciones de A y B, y se representan frente a mediciones reales de las muestras C (n = 86) y D (n = 90) (eje y).

Comparación de plataformas de STARSEQ con respecto a qPCR TaqMan y a RNA-Seq Illumina.

El promedio de las diferencias para las mediciones de las muestras A y B entre STARSEQ y qPCR TaqMan (figura 24 que muestra diagramas de diferencias entre las mediciones de TaqMan y STARSEQ) o RNA-Seq Illumina (figura 25 que muestra diagramas de diferencias entre mediciones de RNA-Seq Illumina y STARSEQ) se determinó para cada diana endógena. Esta diferencia se resta de las mediciones de qPCR TaqMan o RNA-Seq Illumina para las muestras C y D y se representa gráficamente (eje x) frente a las mediciones de STARSEQ de C y D (eje y).

Las mediciones de STARSEQ representan la mediana de la medición de preparaciones de biblioteca que tenían por lo menos 15 lecturas de secuenciación tanto para DN como para PI. La figura 26A muestra una comparación de qPCR TaqMan con STARSEQ (n = 292). La figura 26B muestra una comparación de RNA-Seq Illumina con STARSEQ (n = 340).

Rendimiento del ensayo

El rendimiento de medición del ensayo se evaluó en las muestras SEQC A, B, C y D para ERCC, así como en las dianas de ADNc endógeno, como se muestra en la figura 26. Las dianas endógenas también se evaluaron frente al

control de ADNg (véase la figura 18B).

Se produjeron mediciones negativas verdaderas cuando se secuenció un número suficiente de patrones internos competitivos (secuenciados al menos 15 veces), pero se observó una plantilla natural insuficiente a todas las concentraciones de enriquecimiento del patrón interno. Un límite superior de expresión para estos ensayos todavía se puede calcular como [1/(recuentos de secuenciación de PI)] x concentración de PI en la preparación de la biblioteca con la concentración de PI más baja presente. Estas mediciones representan mediciones negativas verdaderas y el límite inferior de cuantificación exacta se puede determinar a partir de estos datos.

Los ensayos fallidos son mediciones en las que la "profundidad de secuenciación era demasiado baja" tanto para DN como para PI. Estos representan fallos verdaderos del ensayo (ni el patrón natural ni el interno se secuenciaron al menos 15 veces). De esta forma, las mezclas de PI competitivos pueden controlar los falsos negativos.

Adición de códigos de barras y adaptadores de secuenciación

Un conjunto de cebadores de fusión fueron diseñados con su extremo 3' complementario a las colas de secuencias APEX-2 universales añadidas durante la PCR competitiva múltiple. Estos cebadores de fusión tienen una cola con una secuencia de código de barras / índice de cuatro nucleótidos y, 5' con respecto a eso, un adaptador de secuenciación directa o inversa de amplicón de torrente de iones (figura 26). Tanto en los cebadores de secuenciación directa como inversa se incluyó intencionadamente un código de barras para indexar dos veces cada muestra y reducir la probabilidad de indexación falsa de una lectura de secuenciación; ambos códigos de barras deben coincidir. Para cada reacción de inserción de código de barras, se preparó un volumen de reacción de 10 µl que contenía: 1 µl de producto de PCR competitiva múltiple, 1 µl de cebador de código de barras directo o inverso 1 µM, 1 µl de dNTP 2 mM, 1 µl de tampón de reacción 10x de Idaho Technology con MgCl2 30 mM, 0,1 µl de Taq polimerasa GoTaq Hot Start de Promega (5 u/µl) y 4,9 µl de agua libre de RNAsa. Cada reacción de código de barras se realizó en un termociclador de aire RapidCycler (Idaho Technology, Inc. Idaho Falls, Idaho) en las siguientes condiciones: 95 ℃ / 3 min (activación Taq); 15 ciclos de 94 ℃ / 5 s (desnaturalización), 58 ℃ / 10 s (hibridación) y 72 ℃ / 15 s (extensión). Los recipientes de reacción se eliminan inmediatamente y se mantienen a 4 ℃ durante todas las etapas posteriores. El objetivo durante esta etapa es evitar la heterodimerización del producto con código de barras. Dependiendo del tipo de heterodimerización, pueden surgir errores de alineación tras la secuenciación a partir de llamadas de bases de secuenciación falsas con la consiguiente disminución en la precisión y exactitud de la medición. Las bibliotecas de secuenciación por PCR competitiva múltiple a las que se les ha incorporado un código de barras se cuantifican entonces individualmente en un Bioanalizador Agilent 2100 usando chips de ADN con reactivos del kit DNA 1000 de acuerdo con el protocolo del fabricante (Agilent Technologies Deutschland GmbH, Waldbronn, Alemania). Las bibliotecas de secuenciación con códigos de barras únicos se mezclan luego en una relación estequiométrica conocida a fin de optimizar el porcentaje de lecturas de secuenciación que eventualmente recibirá cada biblioteca; en la mayoría de los casos se usa 1:1.

"Negativo verdadero" en STARSEQ frente a TaqMan y RNA-Seq

Las 26 mediciones de STARSEQ tenían datos suficientes para informar de una medición por debajo del límite. De las 26 mediciones, TaqMan informó de no detección (ND) para 14, y RNA-Seq informó de ND para 1 (véase la figura 27). Debido a que STARSEQ pudo detectar PI, pero no cuantificar con exactitud la presencia de DN, estas representan detecciones de falsos negativos para TaqMan y RNA-Seq. Se calcularon mediciones por debajo del límite como [1/(recuentos de secuenciación de PI)] x concentración de PI cargada en la preparación de la biblioteca.

Desviación estándar de las mediciones de ERCC.

La figura 28 muestra que la desviación estándar (DE) de las diferencias se calcula a partir de los datos presentados en la figura 21. La DE intraensayo intramuestra se calcula a partir de la mediana de la DE intraensayo en cada muestra A-D. La DE intraensayo intermuestra se calcula a partir de la mediana de la DE intraensayo en las muestras A-D. La DE interensayo intermuestra se calcula a partir de la mediana de la DE interensayo de los residuos en las muestras A-D. Dado que la DE se informa en valores de Log_{10}, es más o menos equivalente a informar del coeficiente de variación (CV).

Purificación y secuenciación del producto

En ciertos modos de realización, es necesario durante la purificación de bibliotecas de secuenciación con códigos de barras que un sistema no use desnaturalizantes fuertes o sales caotrópicas, tales como clorhidrato o tiocianato de guanidina. Estos agentes dan como resultado la heterodimerización de la plantilla en dirección 3', llamadas de bases de secuenciación falsas y errores de alineación tras la secuenciación. Por esta razón, cada mezcla de bibliotecas de secuenciación con código de barras se purificó utilizando geles de Agarosa al 2 % E-Gel SizeSelect de Life Technologies, que no informan del uso de desnaturalizantes o sales caotrópicas, y se pueden ejecutar en una sala refrigerada para evitar la desnaturalización térmica durante la separación por electroforesis. Las bibliotecas de secuenciación purificadas se cuantificaron luego usando el kit de cuantificación de bibliotecas KAPA para plataformas de secuenciación de torrente de iones (Kapa Biosystems). En base a esta cuantificación, las bibliotecas

se diluyeron de manera adecuada y se prepararon para el servicio de secuenciación Ion Torrent PGM de acuerdo con las recomendaciones del fabricante en el Centro Médico de la Universidad de Toledo (UTMC), Toledo, OH y la Universidad de Ohio (OU), Athens, OH.

Procesamiento de archivos FASTQ

Los datos de secuenciación en bruto de un servicio NGS se proporcionaron de nuevo en formato FASTQ. Las lecturas de secuenciación se extrajeron y cada lectura de secuenciación se analizó en 3 archivos FASTQ separados: 1) regiones de código de barras directo (*query-barcode.fastq*) y 2) código de barras inverso (*query-revbarcode.fastq*), así como 3) porción central del amplicón (*query-subject.fastq*) correspondiente a la región interna para los sitios de cebado específico de la diana donde deben existir seis sustituciones de nucleótidos entre la DN y el PI competitivo coincidente.

BFAST de secuencias frente a la base de datos de sujetos

Cada uno de los tres archivos FASTQ se alinearon con bases de datos FASTA de referencia conocidas, correspondientes a si se trataba de un código de barras (*barcode.fa*) o de una región de amplicón (*subject.fa*) utilizando la herramienta de búsqueda rápida y exacta similar a BLAT (BFAST, versión 0.7.0a), con generación de archivo en formato de mapa/alineación de secuencia (SAM). La coincidencia BFAST con las bases de datos de índice y la generación del archivo SAM se realizaron para los archivos FASTQ recortados que contenían 1) código de barras directo, 2) código de barras inverso y 3) secuencias de sujeto de amplicón capturadas.

Intercalación de recuentos de secuencias

Cada uno de los tres archivos SAM de 1) código de barras directo, 2) código de barras inverso y 3) región de amplicón se fusionaron entonces en una tabla hash de un lenguaje práctico de extracción e informes (PERL) utilizando la ID de lectura de secuenciación como clave para la coincidencia (http://www.perl.org/). En función del alineamiento de códigos de barras y amplicones, cada lectura de secuenciación se agrupó en una matriz correspondiente a la concentración de entrada de PI para una preparación de muestra determinada, y si se llamaba DN o PI por alineación BFAST. Si las llamadas de alineación de códigos de barras directos e inversos no coincidían, la lectura de secuenciación no se agrupó. La tabla hash resultante de lecturas de secuenciación agrupadas se genera en formato delimitado por comas y se procesa como se describe en la sección Métodos estadísticos.

Medición de la abundancia relativa

Se requirieron al menos 14 lecturas de secuenciación para cada DN y PI. La dilución correcta se determinó en base al cambio en la relación DN:PI en múltiples dianas de ensayo y en múltiples enriquecimientos de patrones internos diluidos en serie. La dilución del patrón interno se multiplicó entonces por la relación DN:PI. Cada ensayo tenía múltiples mediciones por ensayo debido a las diluciones múltiples de patrón interno. Si el STDEV de estas mediciones tenía una varianza inferior a 10 veces, se aceptó la mediana de estas mediciones. Las medidas correctas se basaron en el sesgo sistemático del ensayo predeterminado de las concentraciones de patrones internos. La población de estas mediciones se normalizó a una mediana poblacional.

Criterios de inclusión/exclusión de mediciones de STARSEQ

Cada diana natural (ADNg o ADNc) se midió con relación a su patrón interno respectivo dentro de una concentración de valoración cruzada del ISM (figura 18). Un umbral empírico de al menos 15 lecturas de secuenciación para la diana natural (DN) y para el respectivo patrón interno (PI) competitivo fue el criterio de inclusión/exclusión óptimo para considerar una relación DN:PI válida (potencia > 80 %; tasa de error de tipo 1 < 0,05; para detectar un cambio en la relación DN:PI de 2 veces) (figura 18). Para aquellos ensayos en los que más de una medición cumplía los criterios anteriores, un coeficiente de variación (CV) de >1000 % entre las mediciones desencadenó la exclusión para esa medición de ensayo en esa muestra particular.

Métodos estadísticos: estimación de la concentración de la diana natural

Para cada diana génica y réplica técnica con la concentración de entrada de cada mezcla de PI indexada con el subíndice i, se calculó una estimación de la concentración de la diana natural (CN_i) basándose en los recuentos de secuencia observados/agrupados de la diana natural (DN_i) y del patrón interno (PI_i), así como la concentración inicial conocida (en unidades de copias de plantilla por preparación de biblioteca) del patrón interno (CP_i):

$$\log_{10} NC_i = \log_{10} \frac{NT_i}{IS_i} + \log_{10} SC_i$$

El método óptimo determinado empíricamente y el parámetro de control de calidad para la estimación de la cantidad de integración fue, 1) la mediana ($CN_{mediana}$) de las mediciones de réplicas técnicas de CN_i que tienen, 2) al menos

15 recuentos de secuenciación tanto para DN_i como para PI_i, y 3) el coeficiente de variación (CV) en CN_i de menos de 1,00 en una escala logarítmica de base 10. Esto se seleccionó para tener un muestreo suficiente de una diana dada para permitir la detección de un cambio de 2 veces en la abundancia entre las dianas con una tasa de error de tipo 1 menor de 0,05 y una tasa de error de tipo 2 menor de 0,20.

Ejemplo 4

Ejemplos no limitantes de aplicaciones

En el presente documento se describe un método para obtener un índice numérico que indica que un estado biológico comprende proporcionar 2 muestras correspondientes a cada uno de un primer estado biológico y un segundo estado biológico; medir y/o enumerar una cantidad de cada uno de 2 ácidos nucleicos en cada una de las 2 muestras; proporcionar las cantidades como valores numéricos que son directamente comparables entre varias muestras; calcular matemáticamente los valores numéricos correspondientes a cada uno de los estados biológicos primero y segundo; y determinar un cálculo matemático que discrimine los dos estados biológicos. Los estados biológicos primero y segundo, como se usan en el presente documento, corresponden a dos estados biológicos que se han de comparar, tales como dos estados fenotípicos que se deben distinguir. Los ejemplos no limitantes incluyen, por ejemplo, tejido no patológico (normal) vs. tejido patológico; un cultivo que muestra una respuesta terapéutica a los fármacos vs. un cultivo que muestra menos de la respuesta terapéutica a los fármacos; un sujeto que muestra una respuesta adversa al fármaco vs. un sujeto que muestra una respuesta menos adversa; un grupo de sujetos tratados vs. un grupo de sujetos no tratados, etc.

Un "estado biológico", tal como se usa en el presente documento, se puede referir a un estado fenotípico, por ejemplo, un fenotipo clínicamente relevante u otro trastorno metabólico de interés. Los estados biológicos pueden incluir, por ejemplo, un fenotipo de enfermedad, una predisposición a un estado de enfermedad o un estado sin enfermedad; una respuesta terapéutica al fármaco o predisposición a dicha respuesta, una respuesta adversa al fármaco (por ejemplo, toxicidad del fármaco) o una predisposición a dicha respuesta, una resistencia a un fármaco o una predisposición a mostrar dicha resistencia, etc. El índice numérico obtenido puede actuar como un biomarcador, por ejemplo, correlacionando con un fenotipo de interés. El fármaco puede ser un fármaco antitumoral. En ciertos modos de realización, el uso del método descrito en el presente documento puede proporcionar medicina personalizada.

En algunos de los métodos descritos en el presente documento, el estado biológico corresponde a un nivel de expresión normal de un gen. Cuando el estado biológico no corresponde a los niveles normales, por ejemplo, cayendo fuera de un intervalo deseado, puede estar indicado un estado no normal, por ejemplo, de enfermedad.

Un índice numérico que discrimina un estado biológico particular, por ejemplo, una enfermedad o trastorno metabólico, se puede usar como biomarcador para el trastorno dado y/o para trastornos relacionados con el mismo. Por ejemplo, el estado biológico indicado puede ser al menos uno de un trastorno relacionado con la angiogénesis, un trastorno relacionado con antioxidantes, una afección relacionada con la apoptosis, un trastorno relacionado con el sistema cardiovascular, un trastorno relacionado con el ciclo celular, un trastorno relacionado con la estructura celular, un trastorno relacionado con citoquinas, un trastorno relacionado con la respuesta de defensa, un trastorno relacionado con el desarrollo, un trastorno relacionado con la diabetes, un trastorno relacionado con la diferenciación, un trastorno relacionado con la replicación y/o reparación del ADN, un trastorno relacionado con células endoteliales, un trastorno relacionado con receptores hormonales, un trastorno relacionado con el receptor de folato, un trastorno relacionado con la inflamación, un trastorno relacionado con el metabolismo intermedio, un trastorno relacionado con el transporte de membrana, un trastorno relacionado con la neurotransmisión, un trastorno relacionado con el cáncer, un trastorno relacionado con el metabolismo oxidativo, un trastorno relacionado con la maduración de proteínas, un trastorno relacionado con la transducción de señales, un trastorno relacionado con la respuesta al estrés, un trastorno relacionado con la estructura del tejido, un trastorno relacionado con el factor de transcripción, un trastorno relacionado con el transporte y un trastorno relacionado con el metabolismo xenobiótico. Los genes de enzimas antioxidantes y del metabolismo xenobiótico se pueden evaluar en células humanas; expresión génica de células endoteliales microvasculares; expresión de genes de transporte de membrana; resistencia inmunitaria; control de la transcripción de la expresión de receptores de hormonas; y patrones de expresión génica con resistencia a fármacos en carcinomas y tumores.

El uno o más de los ácidos nucleicos que se han de medir están asociados con uno de los estados biológicos en mayor grado que el otro u otros. Por ejemplo, uno o más de los ácidos nucleicos que se han de evaluar se asocian con un primer estado biológico y no con un segundo estado biológico.

Se puede decir que un ácido nucleico está "asociado con" un estado biológico particular cuando el ácido nucleico está positiva o negativamente asociado con el estado biológico. Por ejemplo, se puede decir que un ácido nucleico está "asociado positivamente" con un primer estado biológico cuando el ácido nucleico aparece en cantidades superiores en un primer estado biológico comparado con un segundo estado biológico. Como ilustración, se puede decir que los genes altamente expresados en las células cancerosas en comparación con las células no cancerosas están positivamente asociados con el cáncer. Por otro lado, se puede decir que un ácido nucleico presente en

cantidades más bajas en un primer estado biológico comparado con un segundo estado biológico está negativamente asociado con el primer estado biológico.

El ácido nucleico que se ha de medir y/o enumerar puede corresponder a un gen asociado con un fenotipo particular. La secuencia del ácido nucleico puede corresponder a las regiones transcritas, expresadas y/o reguladoras del gen (por ejemplo, una región reguladora de un factor de transcripción, por ejemplo, un factor de transcripción para corregulación).

En algunos métodos descritos en el presente documento, las cantidades expresadas de más de 2 genes se miden y se usan para proporcionar un índice numérico indicativo de un estado biológico. Por ejemplo, en algunos casos, los patrones de expresión de genes múltiples se usan para caracterizar un estado fenotípico dado, por ejemplo, un fenotipo clínicamente relevante. Se pueden medir y usar cantidades expresadas de al menos aproximadamente 5 genes, al menos aproximadamente 10 genes, al menos aproximadamente 20 genes, al menos aproximadamente 50 genes o al menos aproximadamente 70 genes para proporcionar un índice numérico indicativo de un estado biológico. En algunos de los métodos descritos en el presente documento se pueden medir y usar cantidades expresadas de menos de aproximadamente 90 genes, menos de aproximadamente 100 genes, menos de aproximadamente 120 genes, menos de aproximadamente 150 genes o menos de aproximadamente 200 genes para proporcionar un índice numérico indicativo de un estado biológico.

La determinación de qué cálculo matemático se ha de utilizar para proporcionar un índice numérico indicativo de un estado biológico se puede lograr mediante cualquier método conocido en la técnica, por ejemplo, en las técnicas matemáticas, estadísticas y/o computacionales. La determinación del cálculo matemático puede implicar un uso de software. Por ejemplo, se puede usar un software de aprendizaje automático.

Los valores numéricos matemáticamente calculados se pueden referir al uso de cualquier ecuación, operación, fórmula y/o regla para interactuar valores numéricos, por ejemplo, una suma, diferencia, producto, cociente, potencia logarítmica y/u otro cálculo matemático. Un índice numérico se puede calcular dividiendo un numerador por un denominador, donde el numerador corresponde a una cantidad de un ácido nucleico y el denominador corresponde a una cantidad de otro ácido nucleico. En algunos casos, el numerador corresponde a un gen positivamente asociado con un estado biológico dado y el denominador corresponde a un gen asociado negativamente con el estado biológico. En algunos de los métodos descritos se puede usar más de un gen asociado positivamente con el estado biológico que se evalúa y más de un gen asociado negativamente con el estado biológico que se evalúa. Por ejemplo, se puede derivar un índice numérico que comprenda valores numéricos para los genes asociados positivamente en el numerador y valores numéricos para un número equivalente de los genes asociados negativamente en el denominador. En dichos índices numéricos equilibrados, los valores numéricos de los ácidos nucleicos de referencia se cancelan. Los valores numéricos equilibrados pueden neutralizar los efectos de la variación en la expresión de los genes que proporcionan los ácidos nucleicos de referencia. En algunos métodos, un índice numérico se calcula mediante una serie de una o más funciones matemáticas.

Se pueden comparar, por ejemplo, distinguir, más de 2 estados biológicos. Por ejemplo, se pueden proporcionar muestras a partir de una gama de estados biológicos, por ejemplo, correspondientes a diferentes estadios de progresión de la enfermedad, por ejemplo, diferentes estadios de cáncer. Las células en diferentes estadios de cáncer, por ejemplo, incluyen una célula no cancerosa frente a una célula cancerosa sin metástasis frente a una célula con metástasis de un paciente determinado en varios momentos durante el curso de la enfermedad. Se pueden usar células cancerosas de varios tipos de cáncer, incluyendo, por ejemplo, un cáncer de vejiga, un cáncer de huesos, un tumor cerebral, un cáncer de mama, un cáncer de colon, un cáncer de sistema endocrino, un cáncer gastrointestinal, un cáncer ginecológico, cáncer de cabeza y cuello, una leucemia, un cáncer de pulmón, un linfoma, una metástasis, un mieloma, tejido neoplásico, un cáncer pediátrico, un cáncer de pene, un cáncer de próstata, un sarcoma, un cáncer de piel, un cáncer testicular, un cáncer de tiroides y un cáncer del tracto urinario. Se pueden desarrollar biomarcadores para predecir qué antineoplásico puede funcionar mejor para un determinado tipo de cáncer, por ejemplo, en un paciente en particular.

Una célula no cancerosa puede incluir una célula de hematoma y/o tejido cicatricial, así como parénquima morfológicamente normal de pacientes sin cáncer, por ejemplo, pacientes sin cáncer relacionados o no relacionados con un paciente con cáncer. Las células no cancerosas también pueden incluir parénquima morfológicamente normal de pacientes con cáncer, por ejemplo, de un sitio cercano al sitio del cáncer en el mismo tejido y/o el mismo órgano; de un sitio más alejado del sitio del cáncer, por ejemplo, en un tejido y/u órgano diferente en el mismo sistema de órganos, o de un sitio aún más alejado, por ejemplo, en un órgano diferente y/o un sistema de órganos diferente.

Los índices numéricos obtenidos se pueden proporcionar como una base de datos. Los índices numéricos y/o sus bases de datos pueden encontrar uso en diagnóstico, por ejemplo, en el desarrollo y la aplicación de pruebas clínicas.

Aplicaciones diagnósticas

En el presente documento también se describe un método para identificar un estado biológico. En algunos casos, el método comprende medir y/o enumerar una cantidad de cada uno de 2 ácidos nucleicos en una muestra, proporcionando las cantidades como valores numéricos; y usar los valores numéricos para proporcionar un índice numérico, a través del cual el índice numérico indica el estado biológico.

Un índice numérico que indica un estado biológico se puede determinar cómo se describió anteriormente. La muestra se puede obtener de una muestra biológica, por ejemplo, una muestra biológica recogida de un sujeto al que hay que tratar. El sujeto puede estar en un entorno clínico, que incluye, por ejemplo, un hospital, una consulta de un proveedor de atención médica, una clínica y/u otro establecimiento de atención médica y/o investigación. Las cantidades de ácido(s) nucleico(s) de interés en la muestra se pueden medir y/o enumerar.

Cuando se va a evaluar un número dado de genes, se pueden obtener datos de expresión para ese número dado de genes simultáneamente. Al comparar el patrón de expresión de ciertos genes con los de una base de datos, se puede determinar un antineoplásico al que probablemente respondería un tumor con ese patrón de expresión génica.

Los métodos descritos en el presente documento se pueden usar para cuantificar un gen exógeno normal en presencia de gen endógeno mutado. Usando cebadores que abarcan la región eliminada, se puede amplificar selectivamente y cuantificar la expresión de un gen normal transfectado y/o un gen constitutivo anormal.

Los métodos descritos en el presente documento se pueden usar para determinar niveles de expresión normales, por ejemplo, proporcionando valores numéricos correspondientes a los niveles de expresión de un transcrito del gen normal. Dichos métodos se pueden usar para indicar un estado biológico normal, al menos con respecto a la expresión del gen evaluado.

Los niveles normales de expresión se pueden referir al nivel de expresión de un transcrito en condiciones normalmente no asociados con una enfermedad, trauma, y/u otra lesión celular. Los niveles de expresión normales se pueden proporcionar como un número, o preferentemente como un intervalo de valores numéricos que corresponden a un intervalo de expresión normal de un gen particular, por ejemplo, dentro de +/- un porcentaje de error experimental. Un valor numérico obtenido para un ácido nucleico dado en una muestra, por ejemplo, un ácido nucleico correspondiente a un gen particular, se puede comparar con valores numéricos normales establecidos, por ejemplo, mediante comparación con los datos en una base de datos proporcionada en el presente documento. Como los valores numéricos pueden indicar el número de moléculas del ácido nucleico en la muestra, esta comparación puede indicar si el gen se está expresando en niveles normales o no.

El método se puede usar para identificar un estado biológico que comprende la evaluación de una cantidad de un ácido nucleico en una primera muestra, y proporcionar dicha cantidad como un valor numérico en el que dicho valor numérico es directamente comparable con un número de otras muestras. En algunos métodos, el valor numérico es potencialmente comparable directamente con un número ilimitado de otras muestras. Las muestras se pueden evaluar en diferentes momentos, por ejemplo, en días diferentes; en el mismo o diferentes experimentos en el mismo laboratorio; y/o en diferentes experimentos en diferentes laboratorios.

Terapéutica

En el presente documento también se divulga un método para mejorar el desarrollo de fármacos. Por ejemplo, el uso de una mezcla estandarizada de patrones internos, una base de datos de valores numéricos y/o una base de datos de índices numéricos se puede usar para mejorar el desarrollo de fármacos.

En algunos métodos, la modulación de la expresión génica se mide y/o se enumera en una o más de estas etapas, por ejemplo, para determinar el efecto de un fármaco candidato. Por ejemplo, un fármaco candidato (por ejemplo, identificado en una etapa dada) se puede administrar a una entidad biológica. La entidad biológica puede ser cualquier entidad capaz de albergar un ácido nucleico, como se describió anteriormente, y se puede seleccionar apropiadamente en función de la etapa de desarrollo del fármaco. Por ejemplo, en la etapa de identificación principal, la entidad biológica puede ser un cultivo in vitro. En la etapa de un ensayo clínico, la entidad biológica puede ser un paciente humano.

El efecto del fármaco candidato sobre la expresión génica se puede evaluar entonces, por ejemplo, usando cualquiera de los métodos divulgados en el presente documento. Por ejemplo, una muestra de ácido nucleico se puede recoger de la entidad biológica y las cantidades de ácidos nucleicos de interés se pueden medir y/o enumerar. Por ejemplo, las cantidades se pueden proporcionar como valores numéricos y/o índices numéricos. Entonces se puede comparar una cantidad con otra cantidad de ese ácido nucleico en una etapa diferente del desarrollo del fármaco; y/o a valores numéricos y/o índices en una base de datos. Esta comparación puede proporcionar información para alterar el proceso de desarrollo del fármaco de una o más maneras.

La alteración de una etapa del desarrollo del fármaco se puede referir a realizar uno o más cambios en el proceso de desarrollo de un fármaco, preferentemente para reducir el tiempo y/o el gasto para el desarrollo del fármaco. Por

ejemplo, la alteración puede comprender la estratificación de un ensayo clínico. La estratificación de un ensayo clínico se puede referir, por ejemplo, a la segmentación de una población de pacientes dentro de un ensayo clínico y/o a la determinación de si un individuo particular puede o no participar en el ensayo clínico y/o continuar a una fase posterior del ensayo clínico. Por ejemplo, los pacientes se pueden segmentar según una o más características de su composición genética determinada. Por ejemplo, considere un valor numérico obtenido en una etapa preclínica, por ejemplo, de un cultivo in vitro que se encuentra que corresponde a la ausencia de respuesta a un fármaco candidato. En la etapa de ensayo clínico, los sujetos que muestran un valor numérico igual o similar pueden quedar exentos de participar en el ensayo. De esta forma, el proceso de desarrollo de fármacos ha sido alterado en consecuencia, ahorrando tiempo y costes.

Kits

Los controles internos de amplificación (IAC) / patrones internos (PI) competitivos descritos en el presente documento se pueden ensamblar y proporcionar en forma de kits. El kit puede proporcionar el IAC y los reactivos necesarios para realizar una PCR, incluida PCR múltiple y secuenciación de nueva generación (NGS). El IAC se puede proporcionar en una única forma concentrada en la que se conoce la concentración o diluirse en serie en solución a al menos una de varias concentraciones de trabajo conocidas.

Los kits pueden incluir PI de 150 dianas endógenas identificadas, como se describe en el presente documento, o PI de 28 dianas de ERCC, como se describe en el presente documento, o ambos. Estos PI se pueden proporcionar en una solución que permita que el PI permanezca estable durante varios años.

Los kits también pueden proporcionar cebadores diseñados específicamente para amplificar el PI de 150 dianas endógenas, el PI de 28 dianas de ERCC, y sus dianas naturales correspondientes. Los kits también pueden proporcionar uno o más recipientes llenos con uno o más reactivos de PCR necesarios, que incluyen pero no se limitan a dNTP, tampón de reacción, polimerasa Taq y agua libre de RNasa. Opcionalmente asociado con dicho(s) recipiente(s) es un aviso en la forma prescrita por una agencia gubernamental que regula la fabricación, uso o venta de IAC y reactivos asociados, cuyo aviso refleja la aprobación por parte de la agencia de fabricación, uso o venta para uso en investigación.

Los kits pueden incluir instrucciones apropiadas para preparar, ejecutar y analizar PCR, incluyendo PCR múltiple y NGS, usando el PI incluido en el kit. Las instrucciones pueden estar en cualquier formato adecuado, incluidos pero no limitados a materiales impresos, cintas de vídeo, discos legibles por ordenador o discos ópticos.

REIVINDICACIONES

1. Un método para reducir el sobremuestreo de dianas de ácidos nucleicos naturales sobrerrepresentadas y el error de muestreo estocástico asociado con la secuenciación profunda, que comprende:

i) preparar una mezcla que comprende un número conocido de moléculas de ácido nucleico de control interno de amplificación (IAC) competitivo correspondientes a cada diana de ácido nucleico natural; y

ii) mezclar la mezcla de IAC competitivos de la etapa i) con una muestra que contiene una diana de ácido nucleico natural antes de la preparación de una biblioteca para secuenciación, o antes de la secuenciación si no se requiere la preparación de la biblioteca;

en el que cada diana de ácido nucleico natural es similar a su respectivo IAC competitivo, con la excepción de uno o más cambios en la secuencia de ácido nucleico que son identificables con la secuenciación, y

en el que dichos cambios pueden incluir una o más deleciones, adiciones o alteraciones en el ordenamiento o la composición de los nucleótidos usados;

iii) evaluar la proporción de eventos de secuenciación entre la diana de ácido nucleico natural y su respectivo IAC competitivo, junto con el número conocido de moléculas de ácido nucleico de IAC competitivo introducidas en la muestra antes de la preparación de la biblioteca; y

iv) determinar cuantificablemente la cantidad original de cada diana de ácido nucleico natural en la muestra original antes de la preparación de la biblioteca y la secuenciación.

2. El método de la reivindicación 1, en el que el número conocido de la etapa i) comprende uno o más de: abundancia, concentración y cantidad.

3. El método de la reivindicación 1, en el que los eventos de secuenciación de la etapa iii) comprenden uno o más de: observaciones, recuentos y lecturas.

4. Un método para determinar cuantificablemente un número de copias de dianas de ácidos nucleicos naturales de una diana de ácido nucleico natural en una muestra, que comprende:

introducir un control interno de amplificación (IAC) competitivo en una muestra de dianas de ácidos nucleicos antes de la preparación de la biblioteca de manera que la diana de ácido nucleico natural y el IAC competitivo experimenten un error no sistemático y/o un sesgo específico del secuenciador de forma similar, en el que en un extremo de la preparación de la biblioteca de secuenciación, la proporción relativa entre cada diana natural única convergerá hacia una abundancia equimolar en la biblioteca.

5. El método de la reivindicación 4, en el que la abundancia equimolar comprende una abundancia uniforme en la biblioteca.

6. El método de cualquiera de las reivindicaciones 2 a 4, en el que, por cada reducción de 10 veces en el rango de abundancia entre las dianas de ácidos nucleicos naturales, se requieren 10 veces menos lecturas de secuenciación.

7. Un método que comprende:

i) evaluar una proporción de eventos de secuenciación entre al menos una diana de ácido nucleico natural y su patrón de control interno de amplificación (IAC) competitivo respectivo;

ii) evaluar el número original de moléculas de IAC competitivo introducidas en la muestra antes de la preparación de la biblioteca; y

iii) determinar el número original de moléculas para cada diana de ácido nucleico natural en la muestra antes de la preparación de la biblioteca y secuenciación multiplicando la proporción entre ácido nucleico natural y el IAC por el número de entrada de IAC competitivos originales.

8. El método de la reivindicación 7, en el que el método incluye usar la misma mezcla de moléculas de IAC competitivo en múltiples pruebas diferentes.

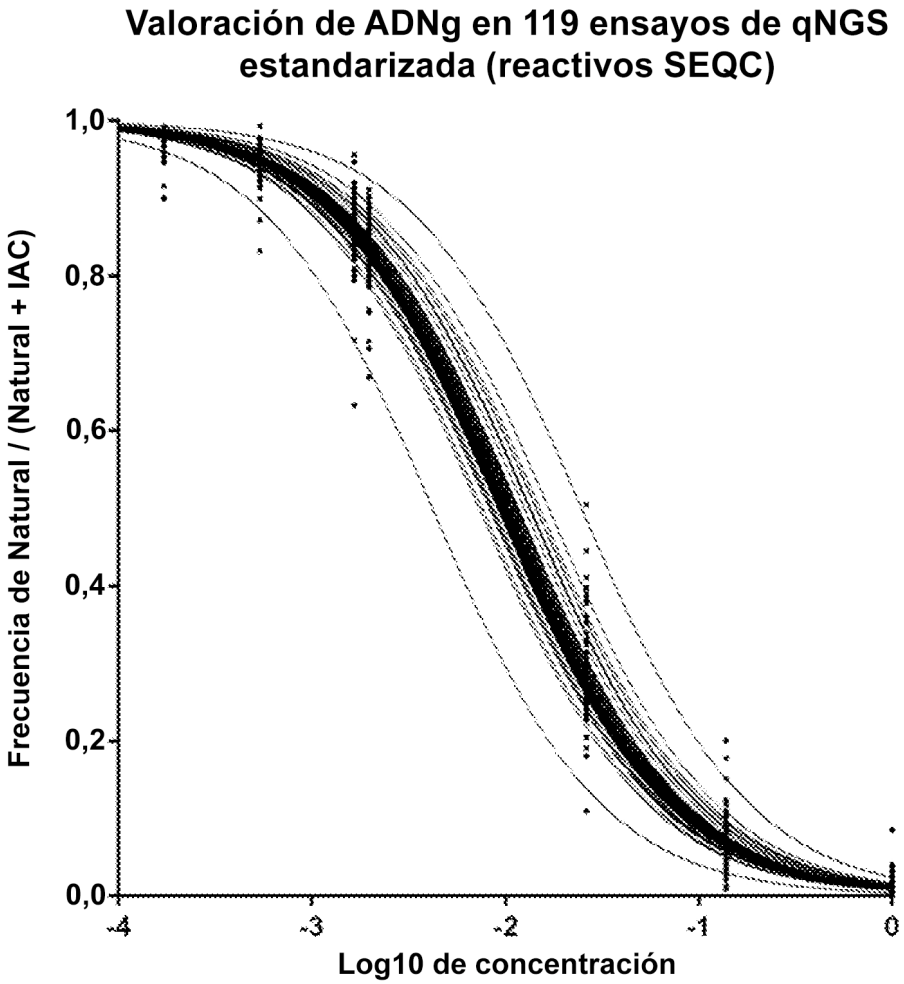
Valoración de ADNg en 119 ensayos de qNGS estandarizada (reactivos SEQC)
Frecuencia de Natural / (Natural + IAC)
1,0
0,8
0,6
0,4
0,2
0,0
-4
-3
-2
-1
0
Log10 de concentración

FIG. 1

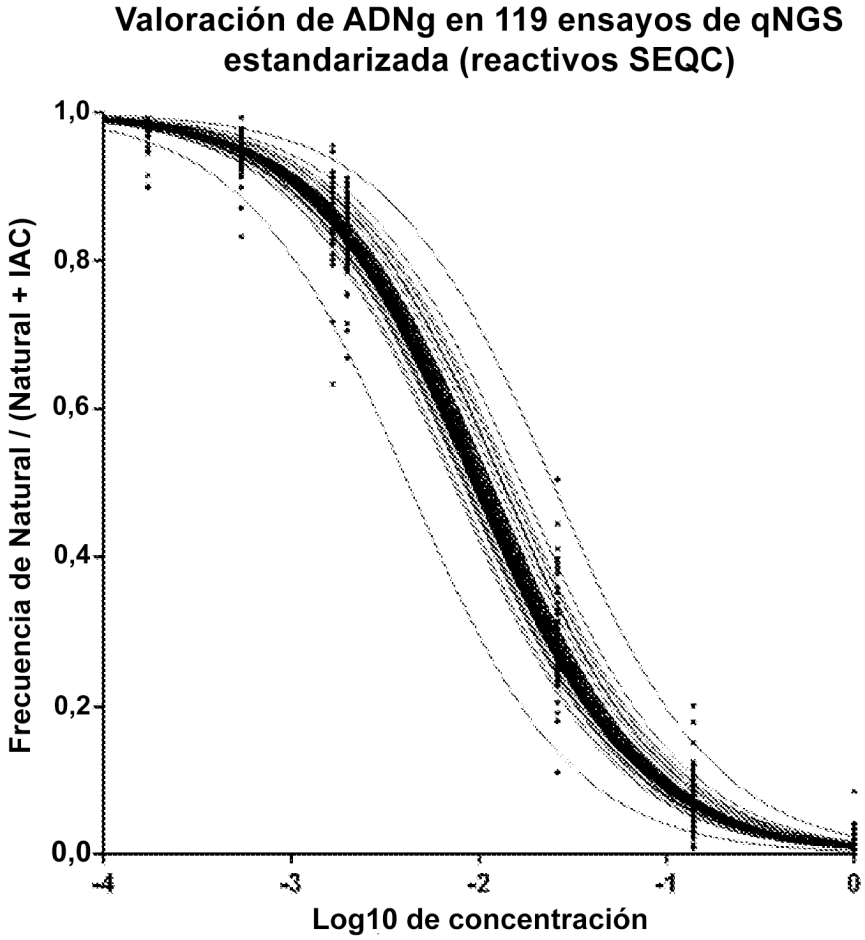
Valoración de ADNg en 119 ensayos de qNGS estandarizada (reactivos SEQC)
Frecuencia de Natural / (Natural + IAC)
1,0
0,8
0,6
0,4
0,2
0,0
-4
-3
-2
-1
0
Log10 de concentración

FIG. 2A

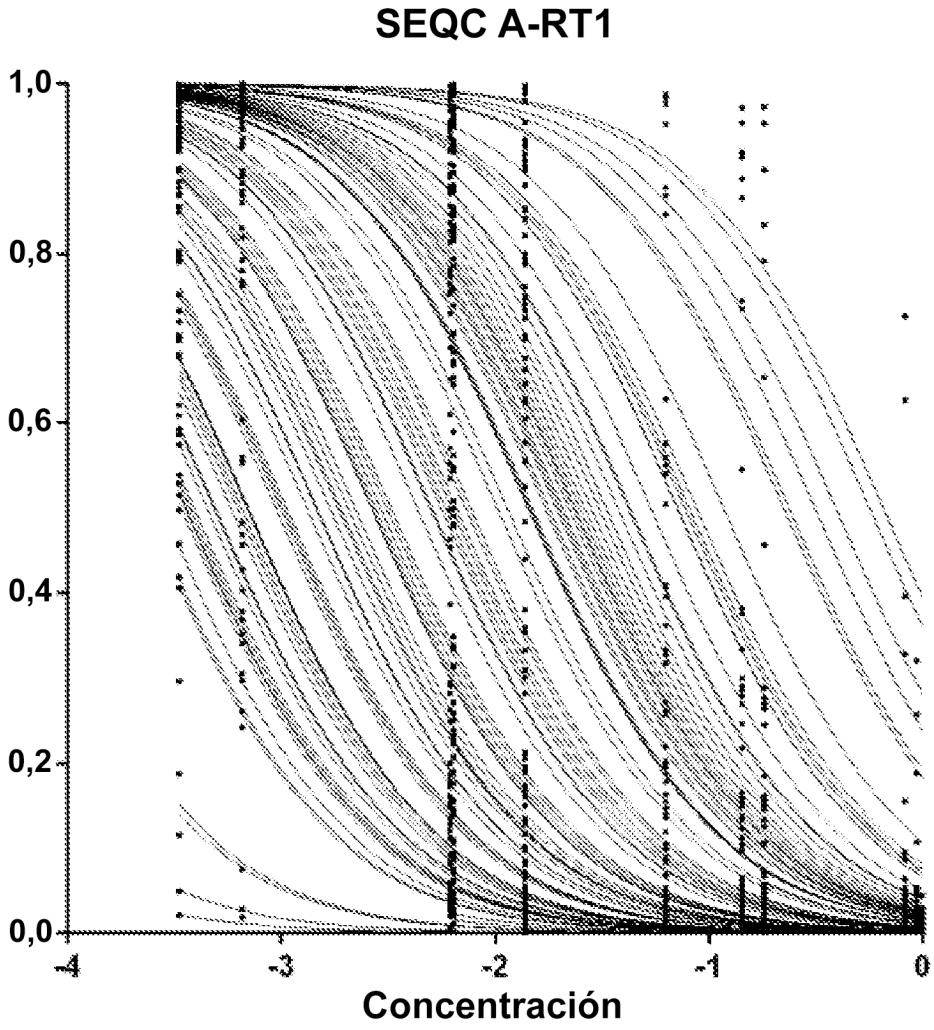

FIG. 2B

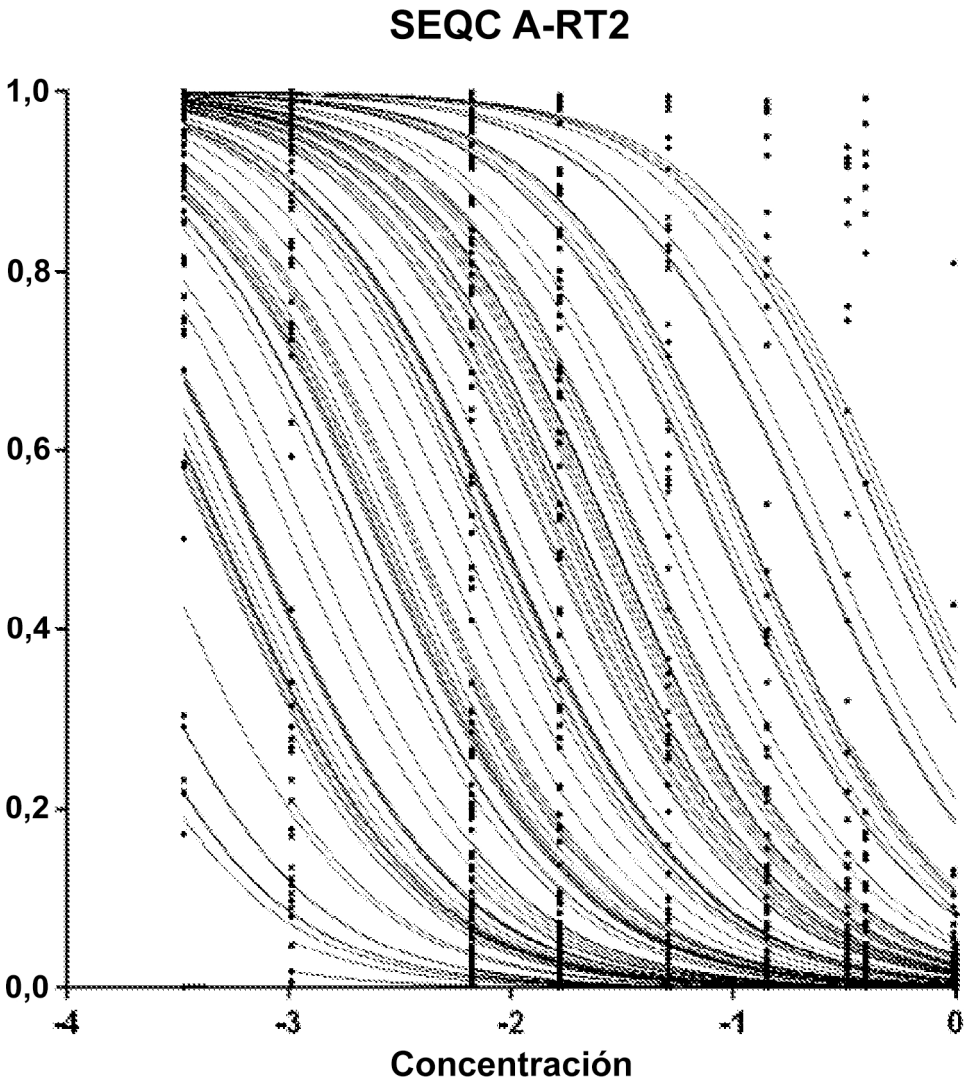

FIG. 2C

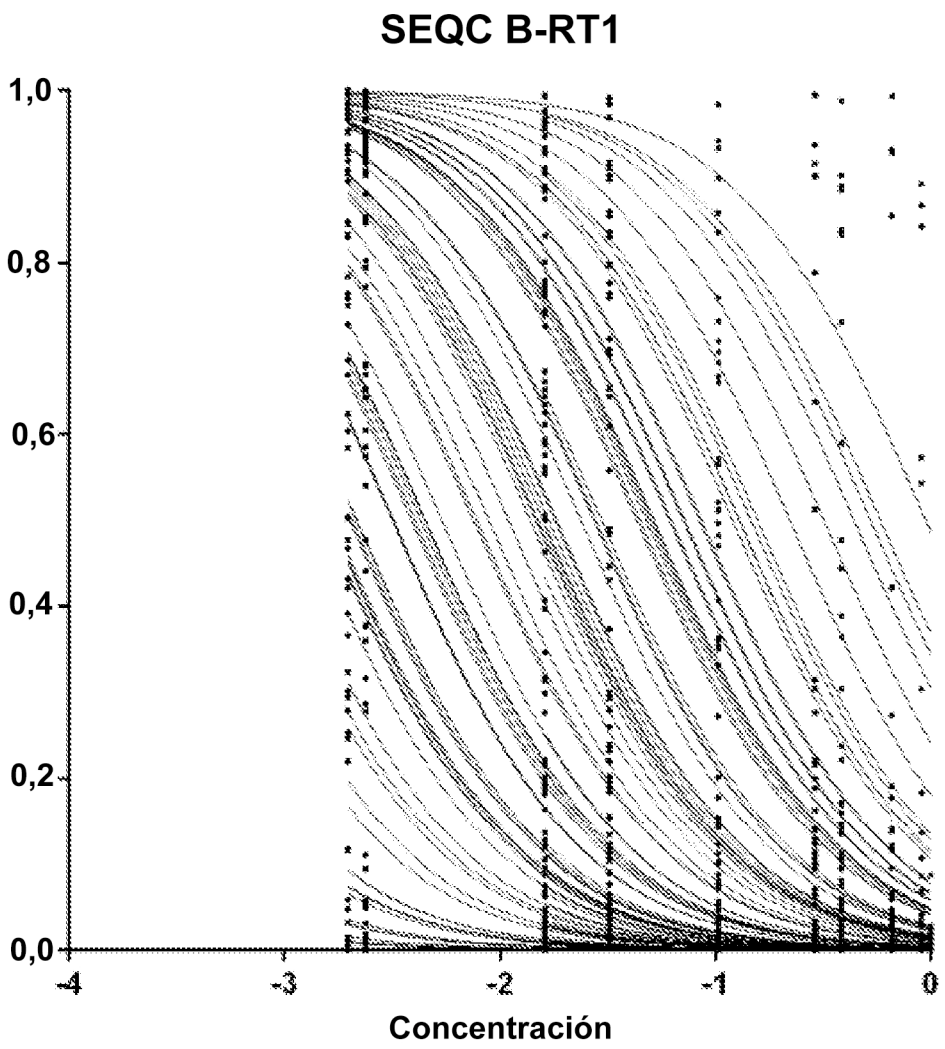

FIG. 2D

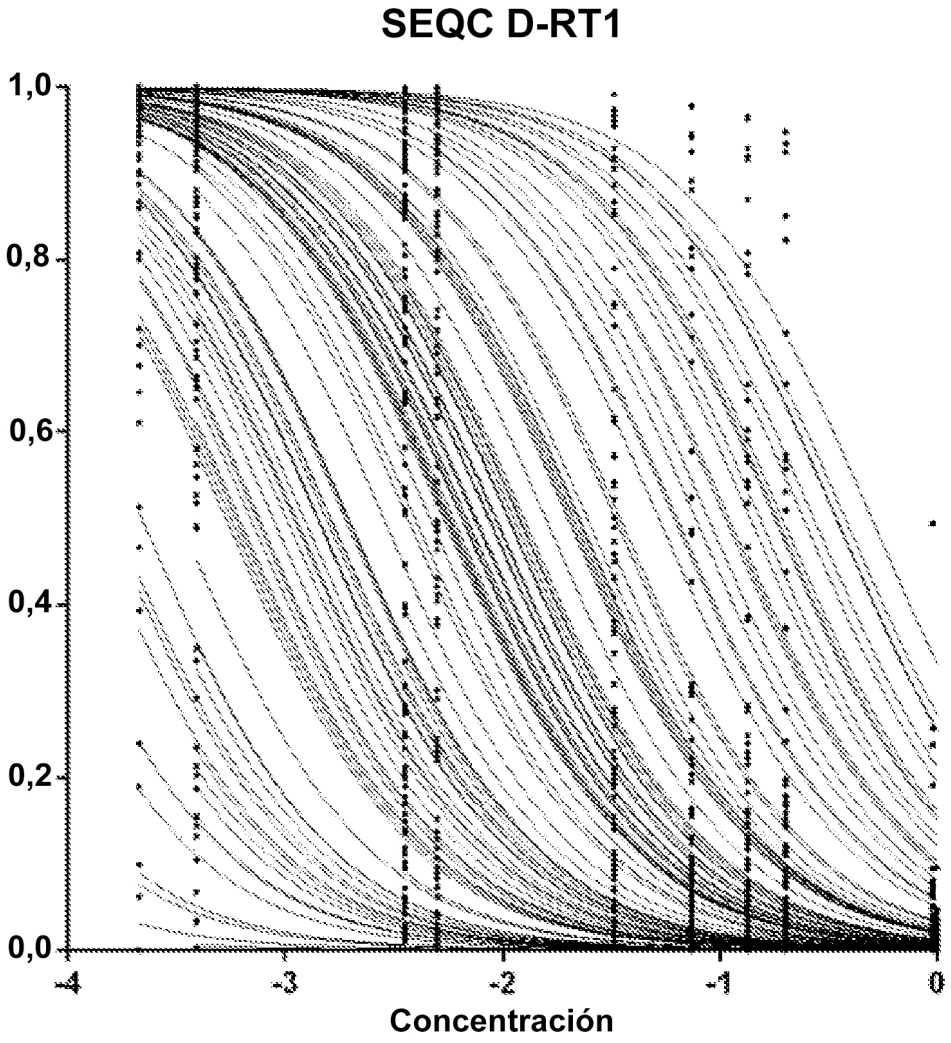

FIG. 2E

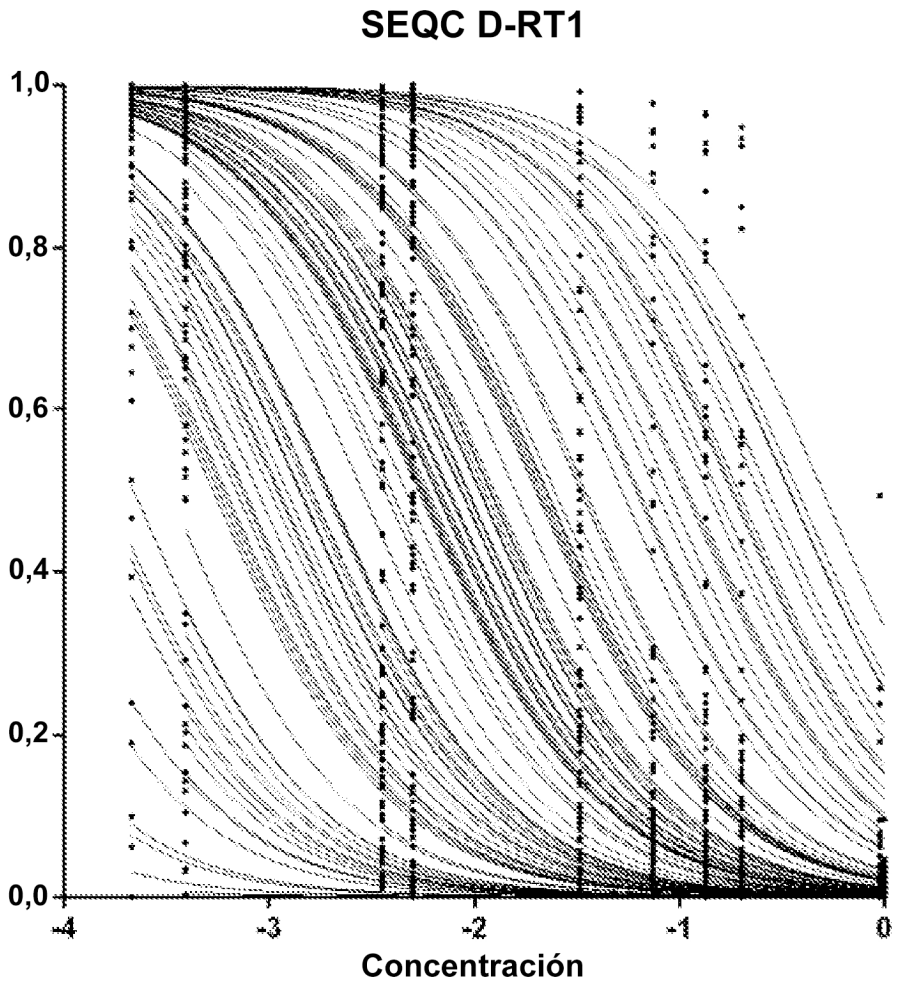

FIG. F

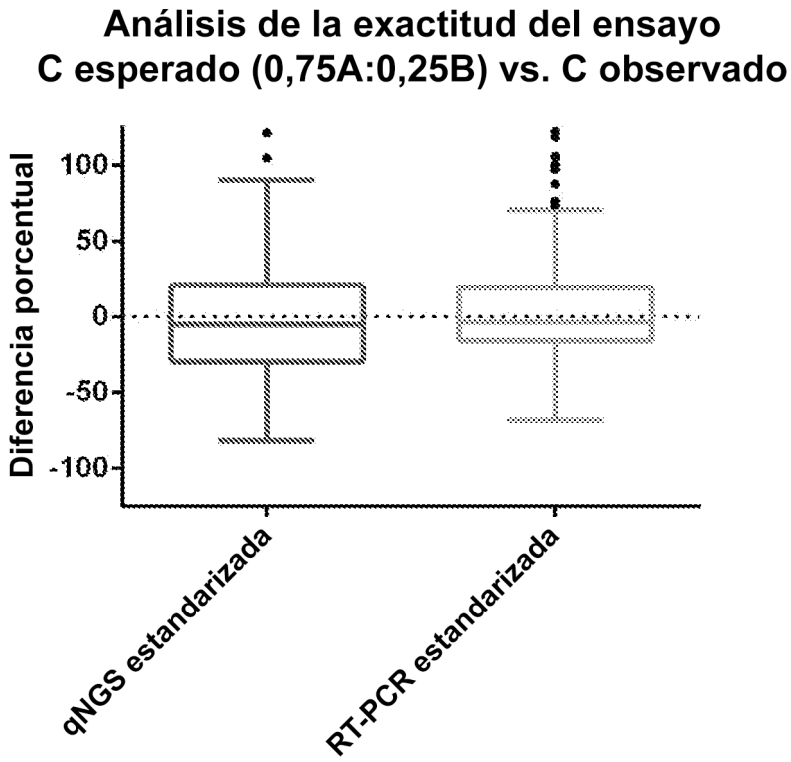
Análisis de la exactitud del ensayo
C esperado (0,75A:0,25B) vs. C observado
Diferencia porcentual
100
50
0
-50
-100
qNGS estandarizada
RT-PCR estandarizada

FIG. 3A

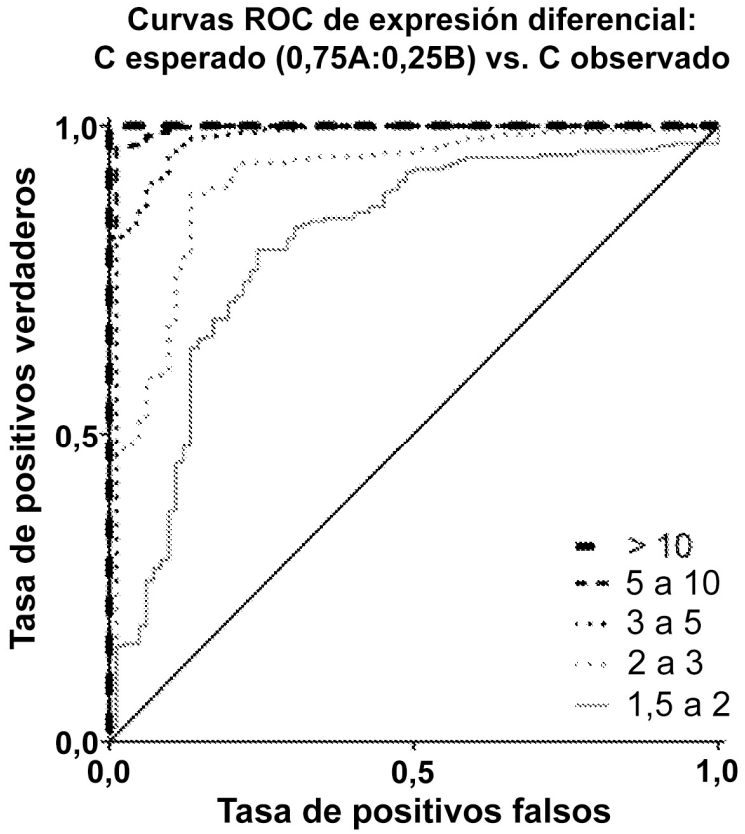

FIG. 4A

Cambio detectado (veces)	Exactitud	CI del 95 %
> 10	0,99	(0,99-1,00)
5 a 10	0,99	(0,99-1,00)
3 a 5	0,97	(0,95-0,99)
2 a 3	0,91	(0,87-0,94)
1,5 a 3	0,80	(0,74-0,86)

FIG. 4B

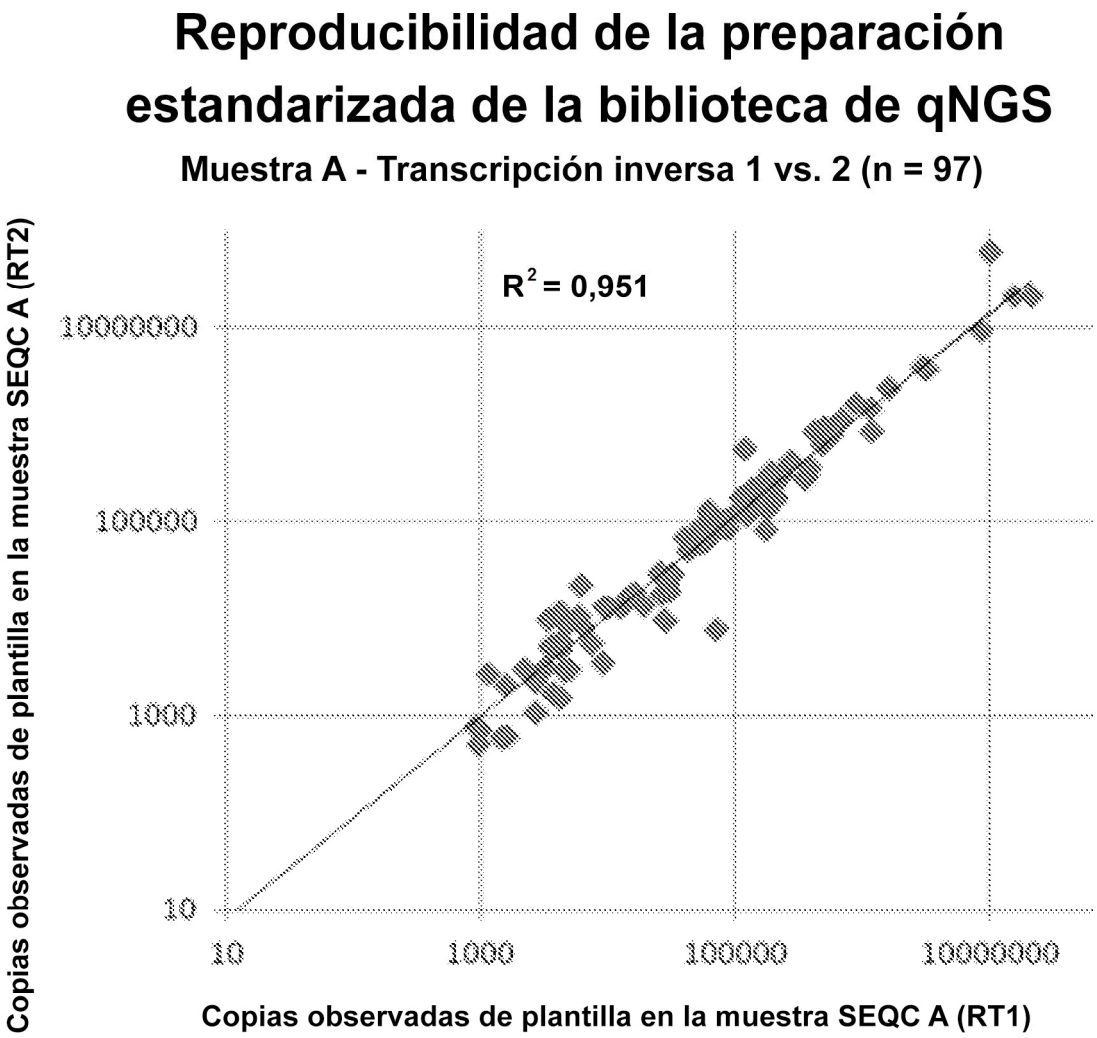
Reproducibilidad de la preparación estandarizada de la biblioteca de qNGS
Muestra A - Transcripción inversa 1 vs. 2 (n = 97)
R2 = 0,951
10000000
100000
1000
10
10
1000
100000
10000000
Copias observadas de plantilla en la muestra SEQC A (RT2)
Copias observadas de plantilla en la muestra SEQC A (RT1)

FIG. 5

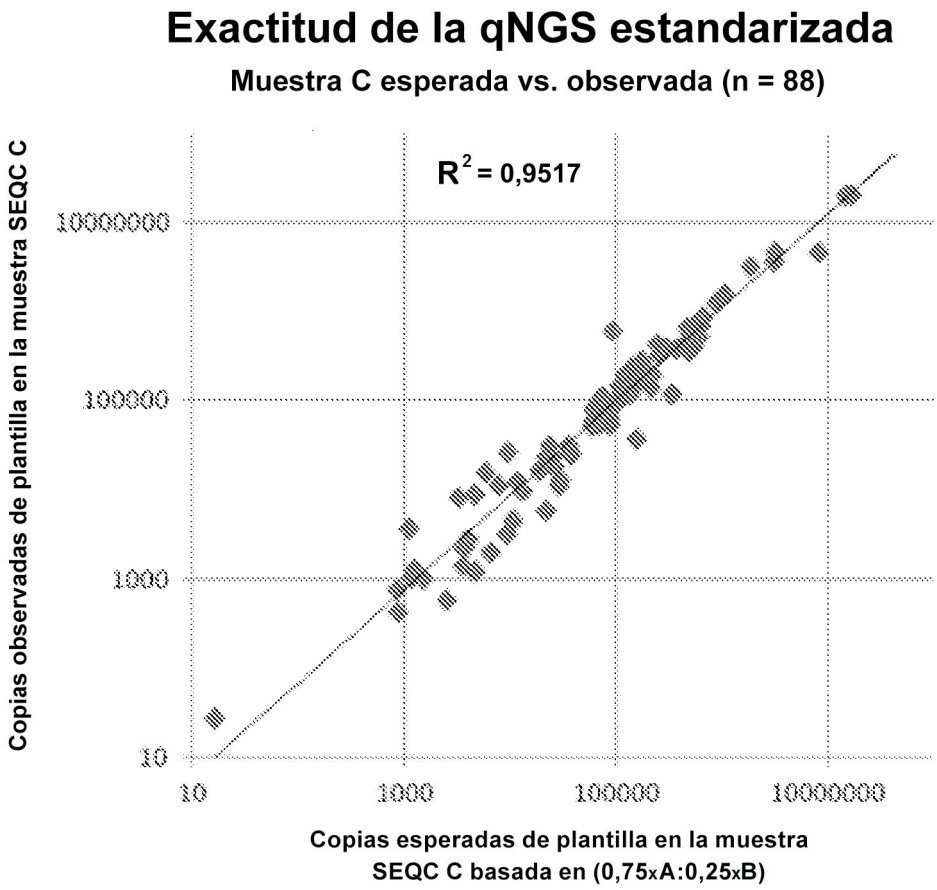
Exactitud de la qNGS estandarizada
Muestra C esperada vs. observada (n = 88)
R2 = 0,9517
Copias observadas de plantilla en la muestra SEQC C
10000000
100000
1000
10
10
1000
100000
10000000
Copias esperadas de plantilla en la muestra
SEQC C basada en (0,75xA:0,25xB)

FIG. 6

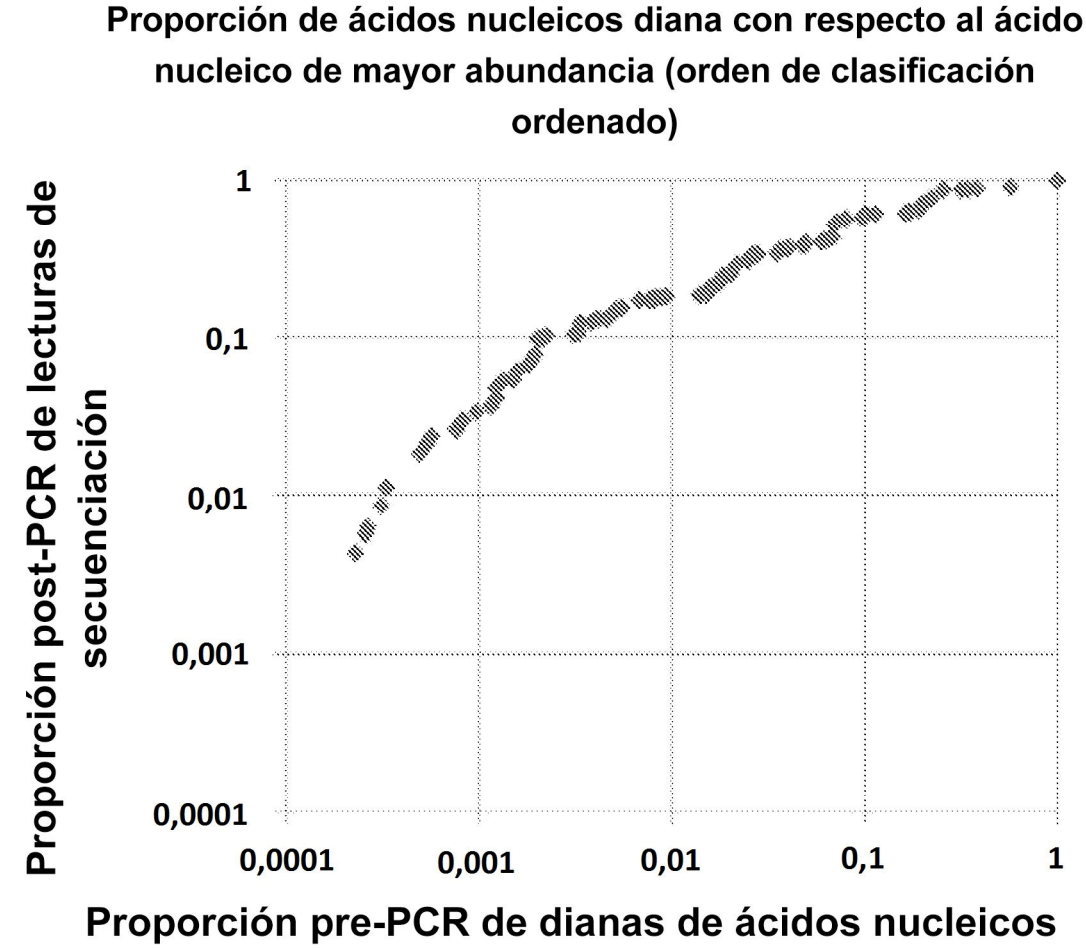
Proporción de ácidos nucleicos diana con respecto al ácido nucleico de mayor abundancia (orden de clasificación ordenado)
Proporción post-PCR de lecturas de secuenciación
1
0,1
0,01
0,001
0,0001
0,0001
0,001
0,01
0,1
1
Proporción pre-PCR de dianas de ácidos nucleicos

FIG. 7

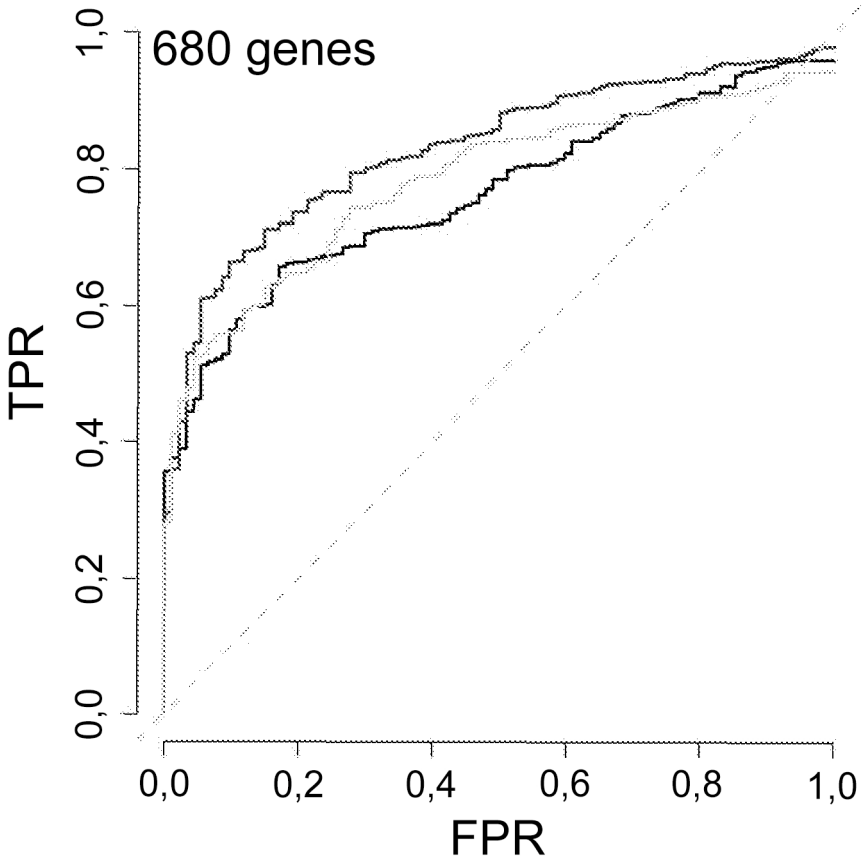

FIG. 8

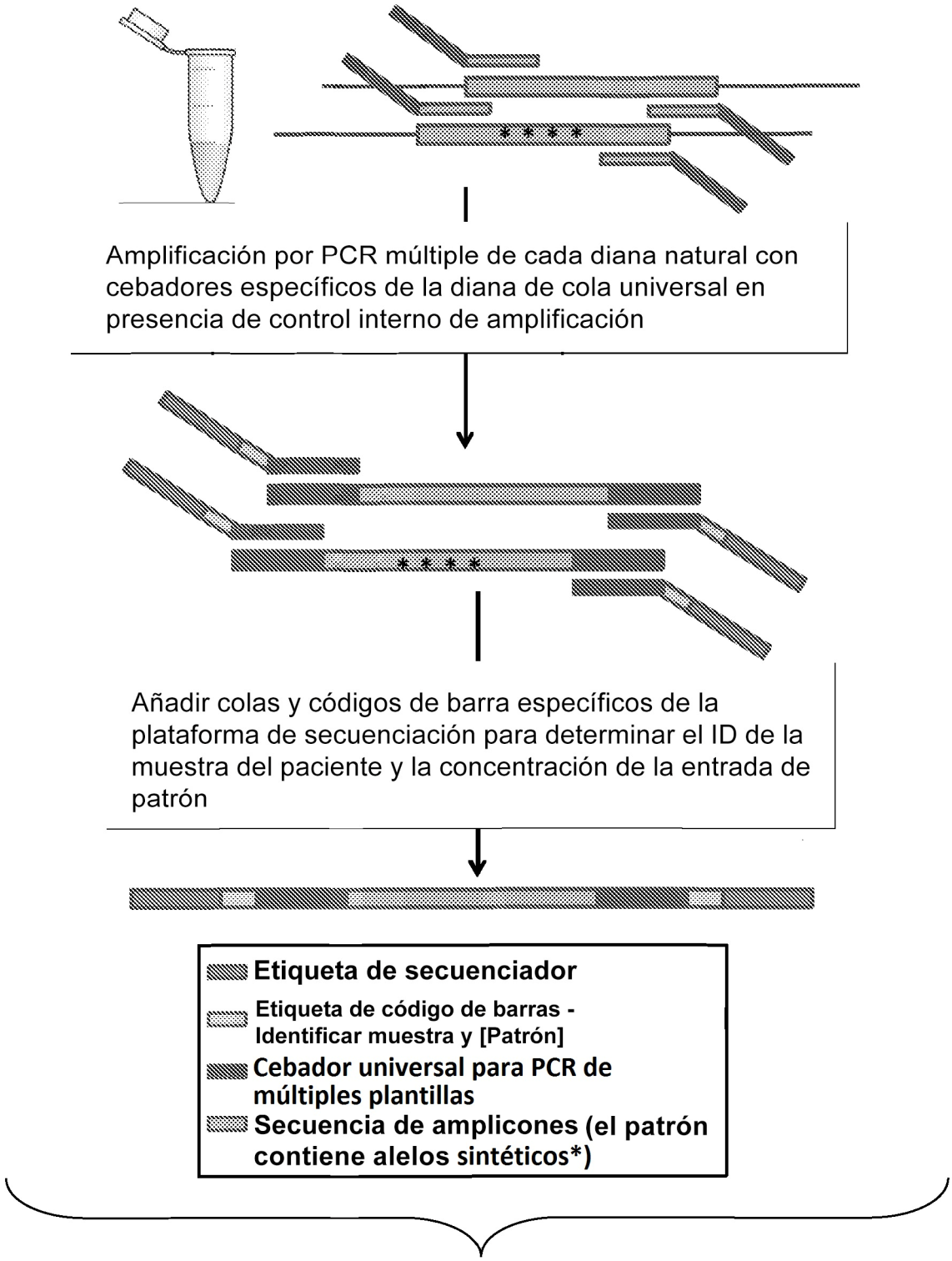
Amplificación por PCR múltiple de cada diana natural con cebadores específicos de la diana de cola universal en presencia de control interno de amplificación
Añadir colas y códigos de barra específicos de la plataforma de secuenciación para determinar el ID de la muestra del paciente y la concentración de la entrada de patrón
Etiqueta de secuenciador
Etiqueta de código de barras - Identificar muestra y [Patrón]
Cebador universal para PCR de múltiples plantillas
Secuencia de amplicones (el patrón contiene alelos sintéticos*)

FIG. 9

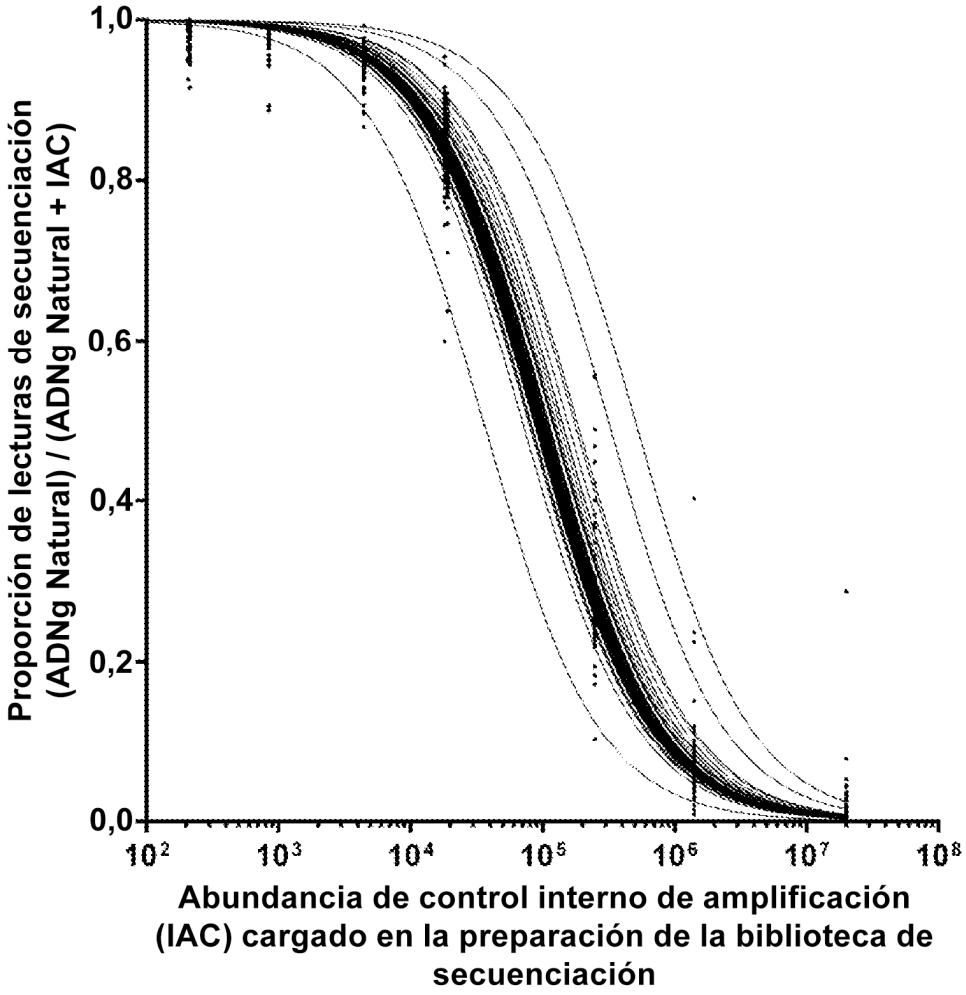
Proporción de lecturas de secuenciación (ADNg Natural) / (ADNg Natural + IAC)
1,0
0,8
0,6
0,4
0,2
0,0
10²
10³
10⁴
10⁵
10⁶
10⁷
10⁸
Abundancia de control interno de amplificación (IAC) cargado en la preparación de la biblioteca de secuenciación

FIG. 10

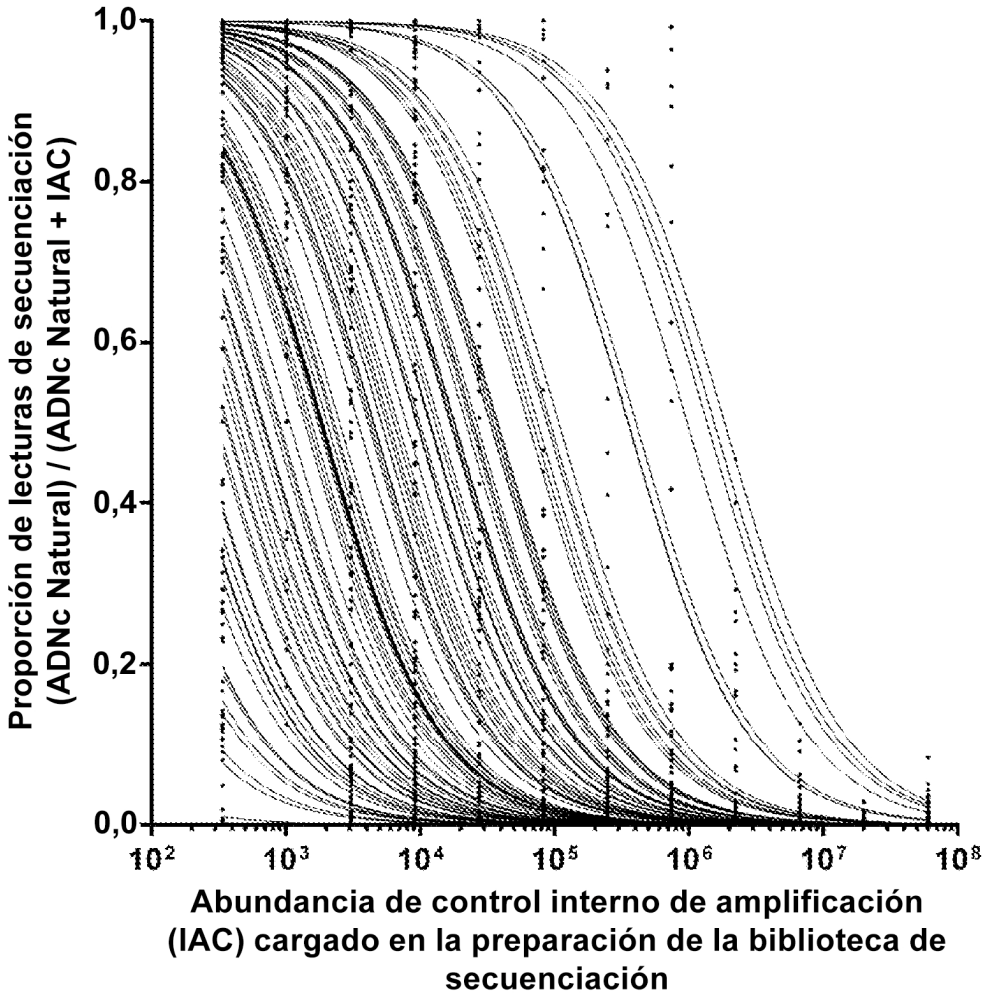
Proporción de lecturas de secuenciación (ADNc Natural) / (ADNc Natural + IAC)
1,0
0,8
0,6
0,4
0,2
0,0
10^2
10^3
10^4
10^5
10^6
10^7
10^8
Abundancia de control interno de amplificación (IAC) cargado en la preparación de la biblioteca de secuenciación

FIG. 11

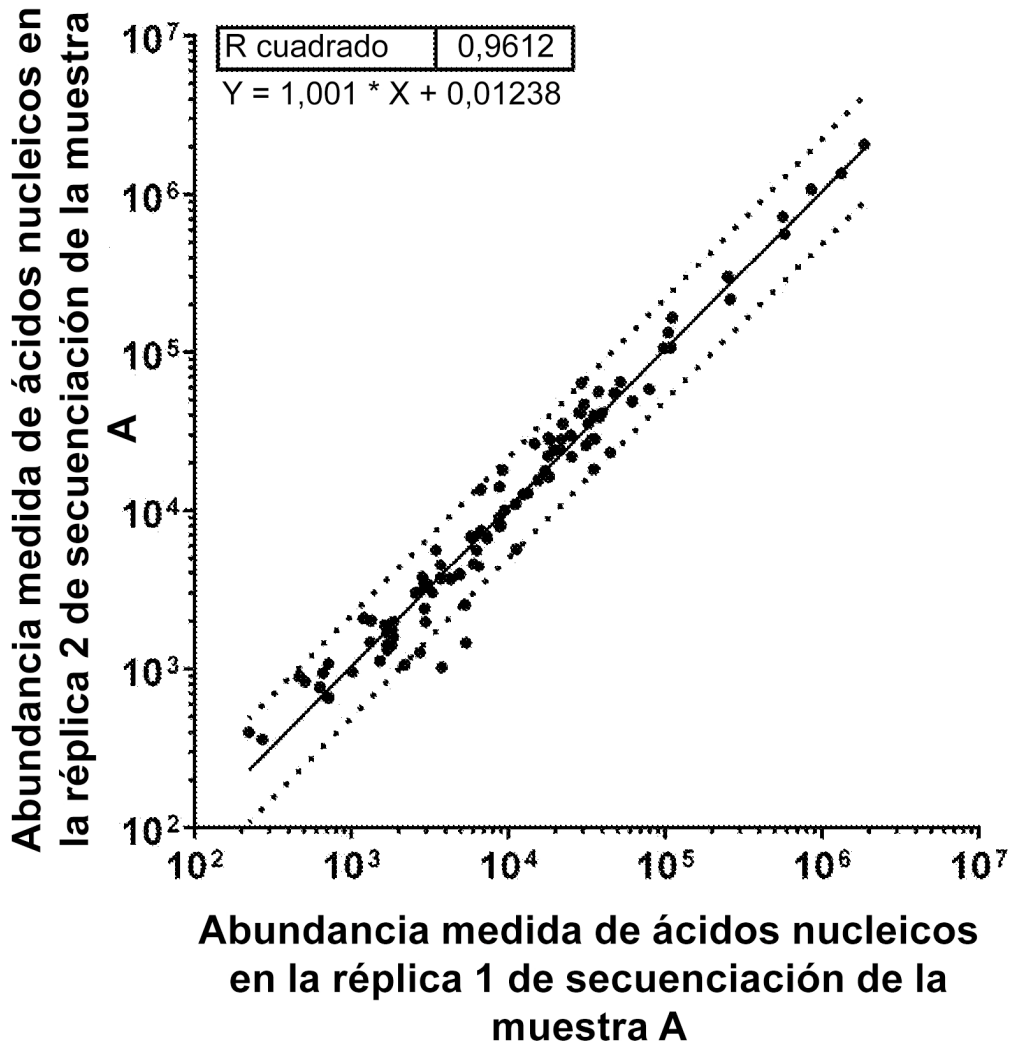
R cuadrado 0,9612
Y = 1,001 * X + 0,01238
Abundancia medida de ácidos nucleicos en la réplica 2 de secuenciación de la muestra A
Abundancia medida de ácidos nucleicos en la réplica 1 de secuenciación de la muestra A
10^2
10^3
10^4
10^5
10^6
10^7

FIG. 12

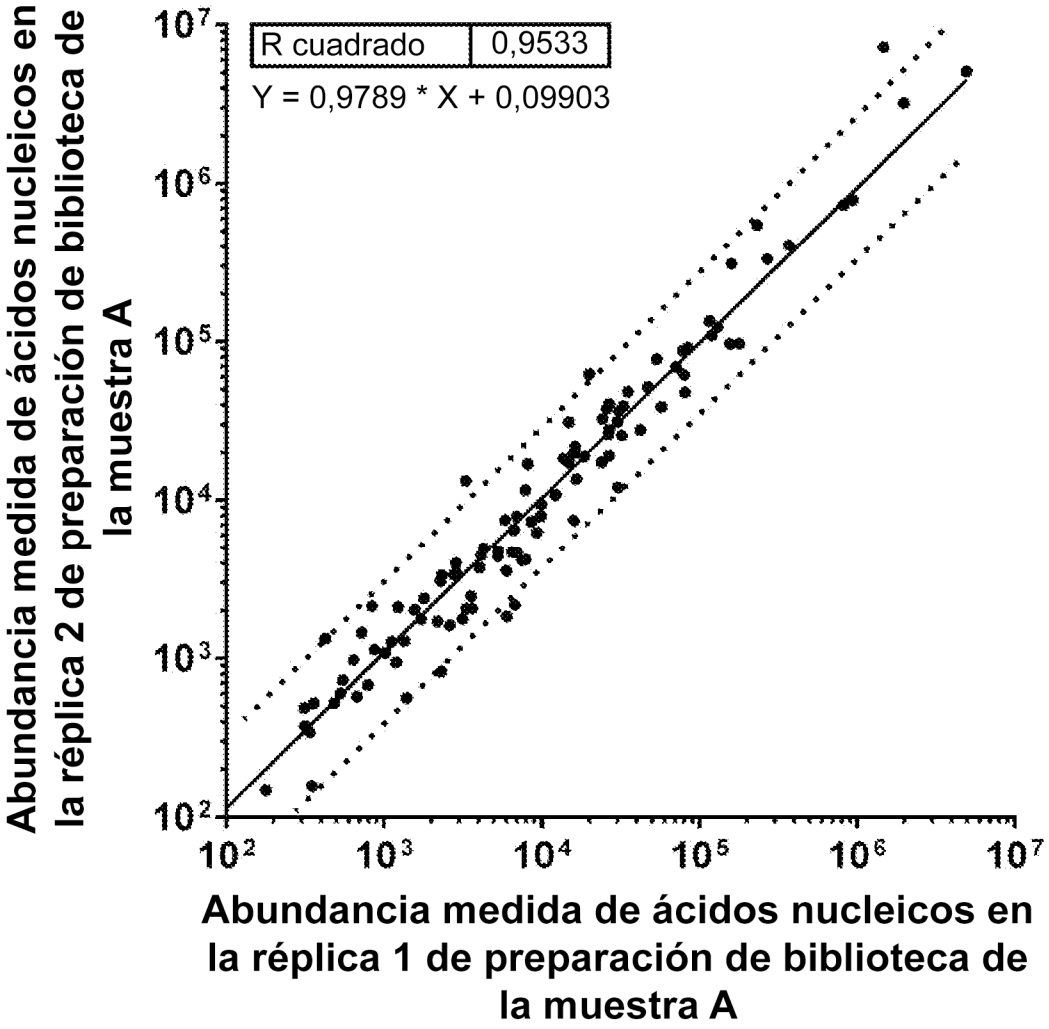
R cuadrado 0,9533
Y = 0,9789 * X + 0,09903
Abundancia medida de ácidos nucleicos en la réplica 2 de preparación de biblioteca de la muestra A
Abundancia medida de ácidos nucleicos en la réplica 1 de preparación de biblioteca de la muestra A
10^2
10^3
10^4
10^5
10^6
10^7

FIG. 13

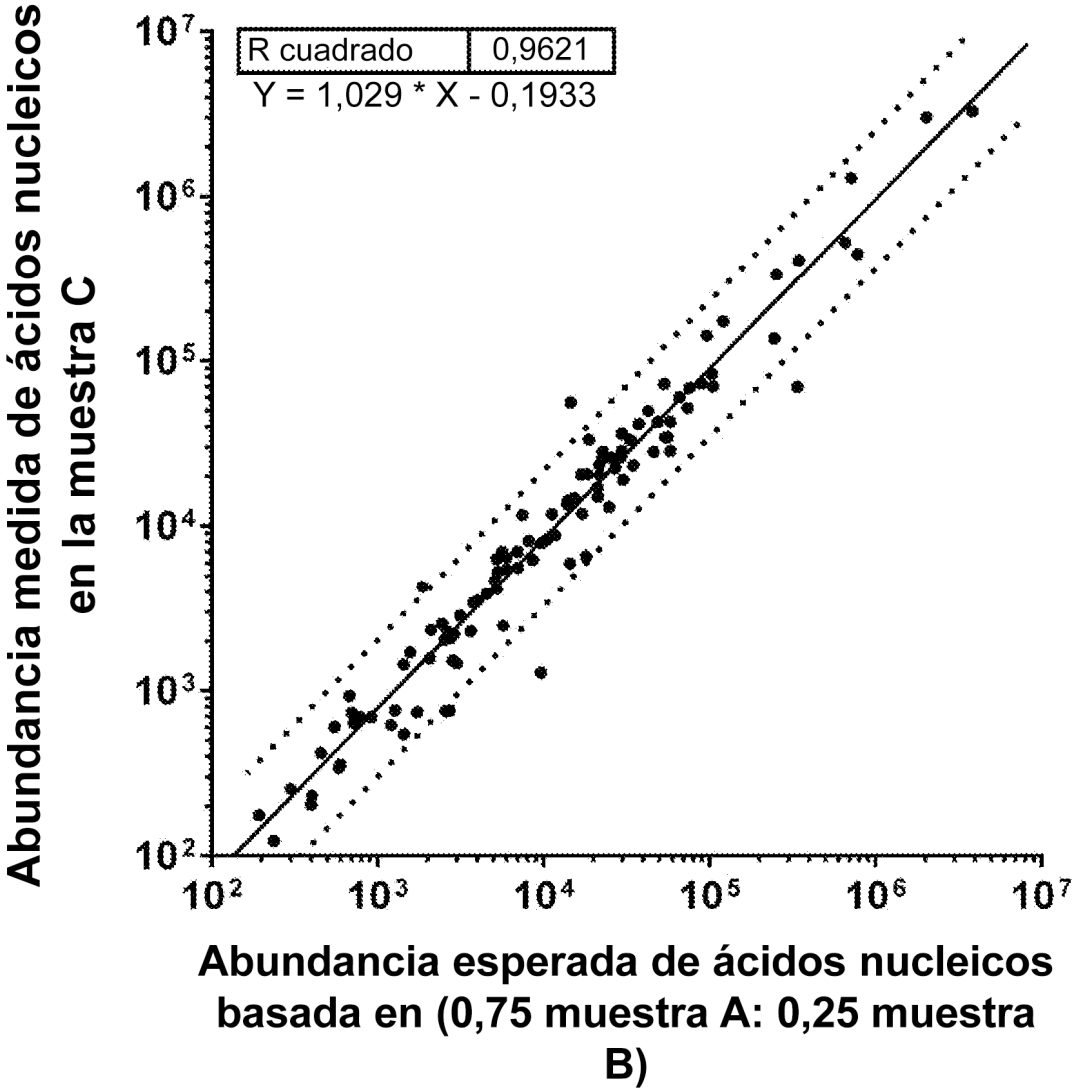
R cuadrado
0,9621
Y = 1,029 * X - 0,1933
Abundancia medida de ácidos nucleicos en la muestra C
Abundancia esperada de ácidos nucleicos basada en (0,75 muestra A: 0,25 muestra B)
10^2
10^3
10^4
10^5
10^6
10^7

FIG. 14A

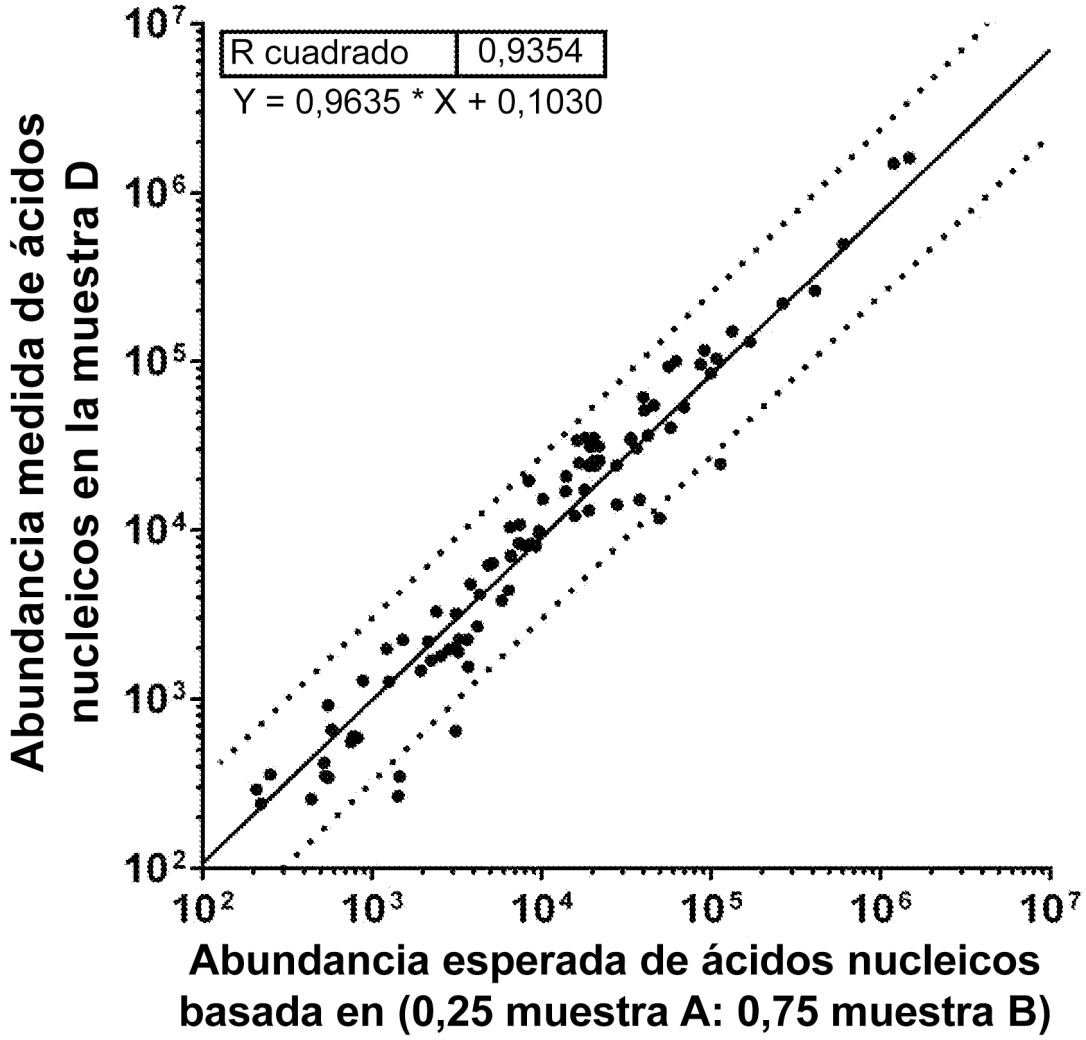
R cuadrado
0,9354
Y = 0,9635 * X + 0,1030
Abundancia medida de ácidos nucleicos en la muestra D
Abundancia esperada de ácidos nucleicos basada en (0,25 muestra A: 0,75 muestra B)
10^2
10^3
10^4
10^5
10^6
10^7

FIG. 14B

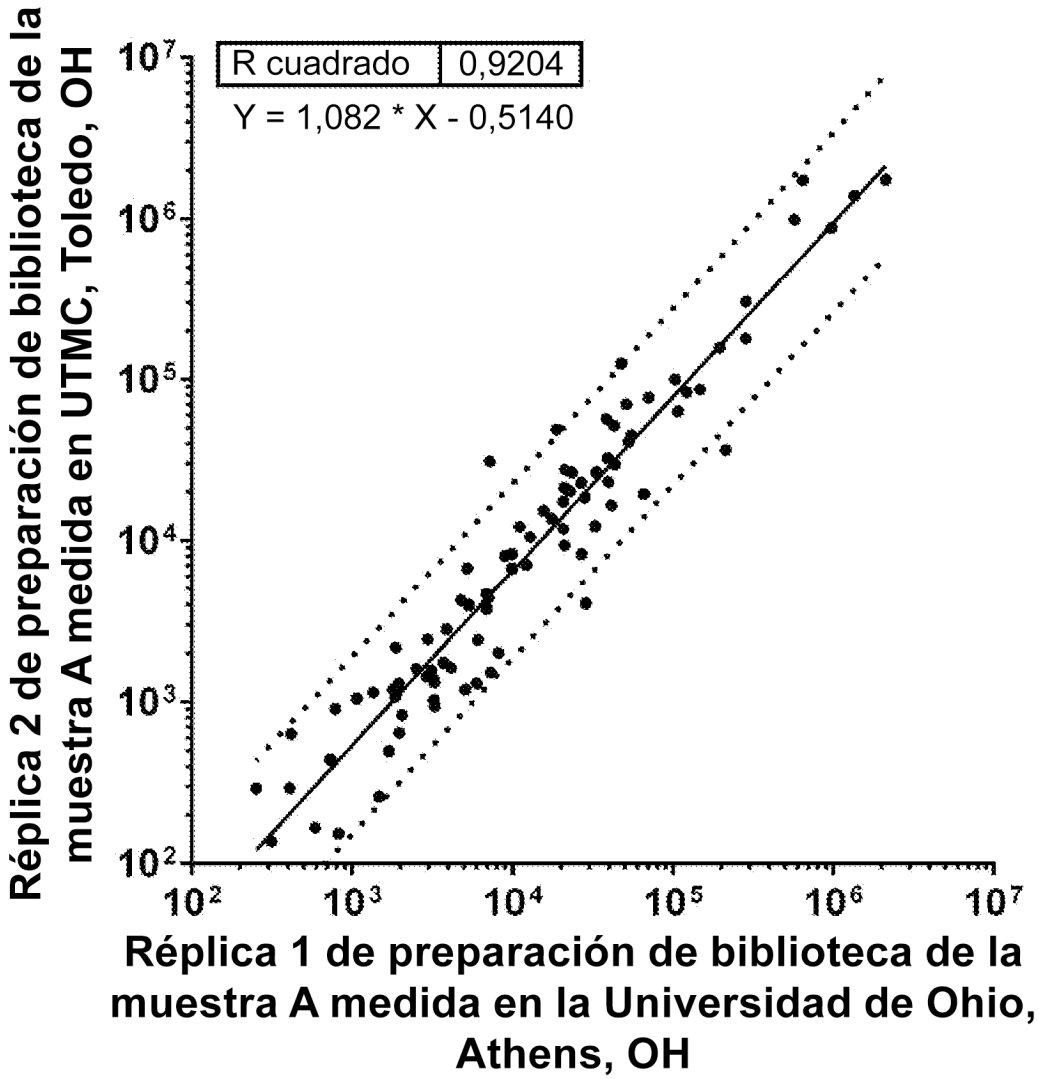
R cuadrado 0,9204
Y = 1,082 * X - 0,5140
Réplica 2 de preparación de biblioteca de la muestra A medida en UTMC, Toledo, OH
Réplica 1 de preparación de biblioteca de la muestra A medida en la Universidad de Ohio, Athens, OH
10^2
10^3
10^4
10^5
10^6
10^7

FIG. 15

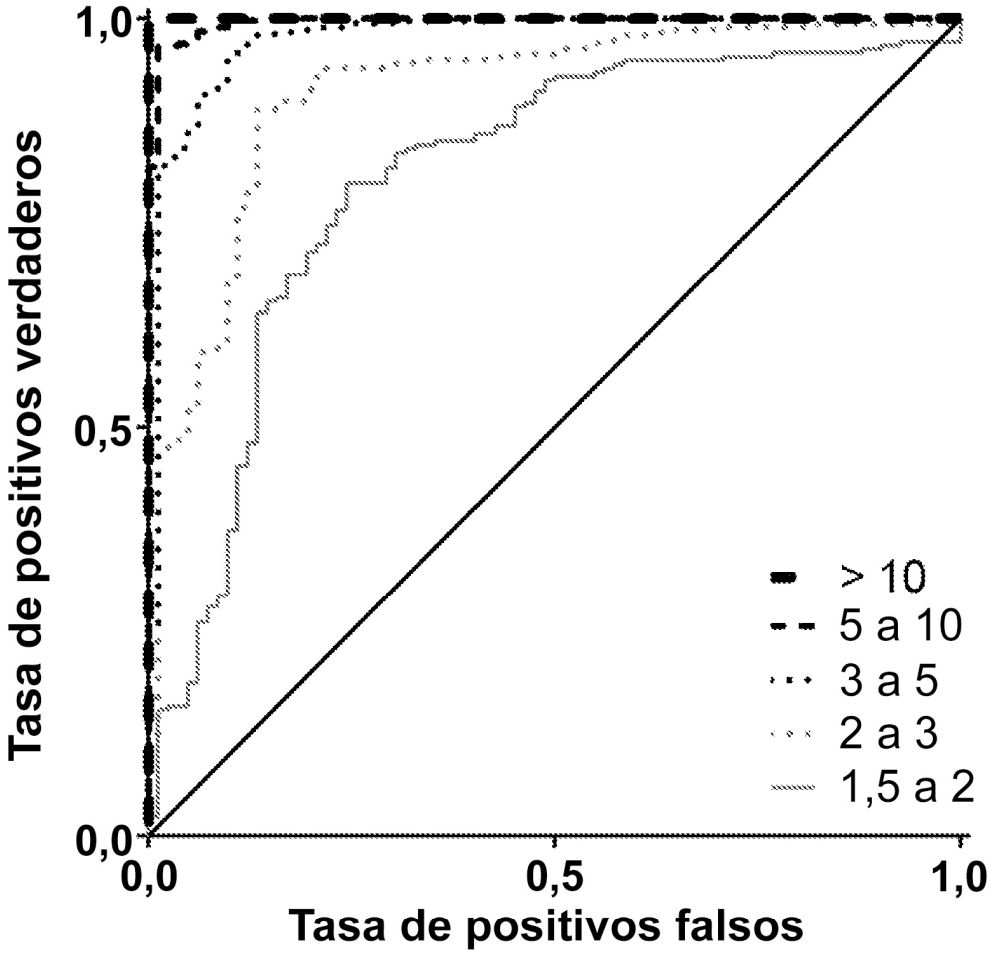

FIG. 16A

Cambio detectado	Exactitud	CI del 95 %
> 10	0,99	(0,99-1,00)
5 a 10	0,99	(0,99-1,00)
3 a 5	0,97	(0,95-0,99)
2 a 3	0,91	(0,87-0,94)
1,5 a 2	0,80	(0,74-0,86)

Fig. 4B

FIG. 16B

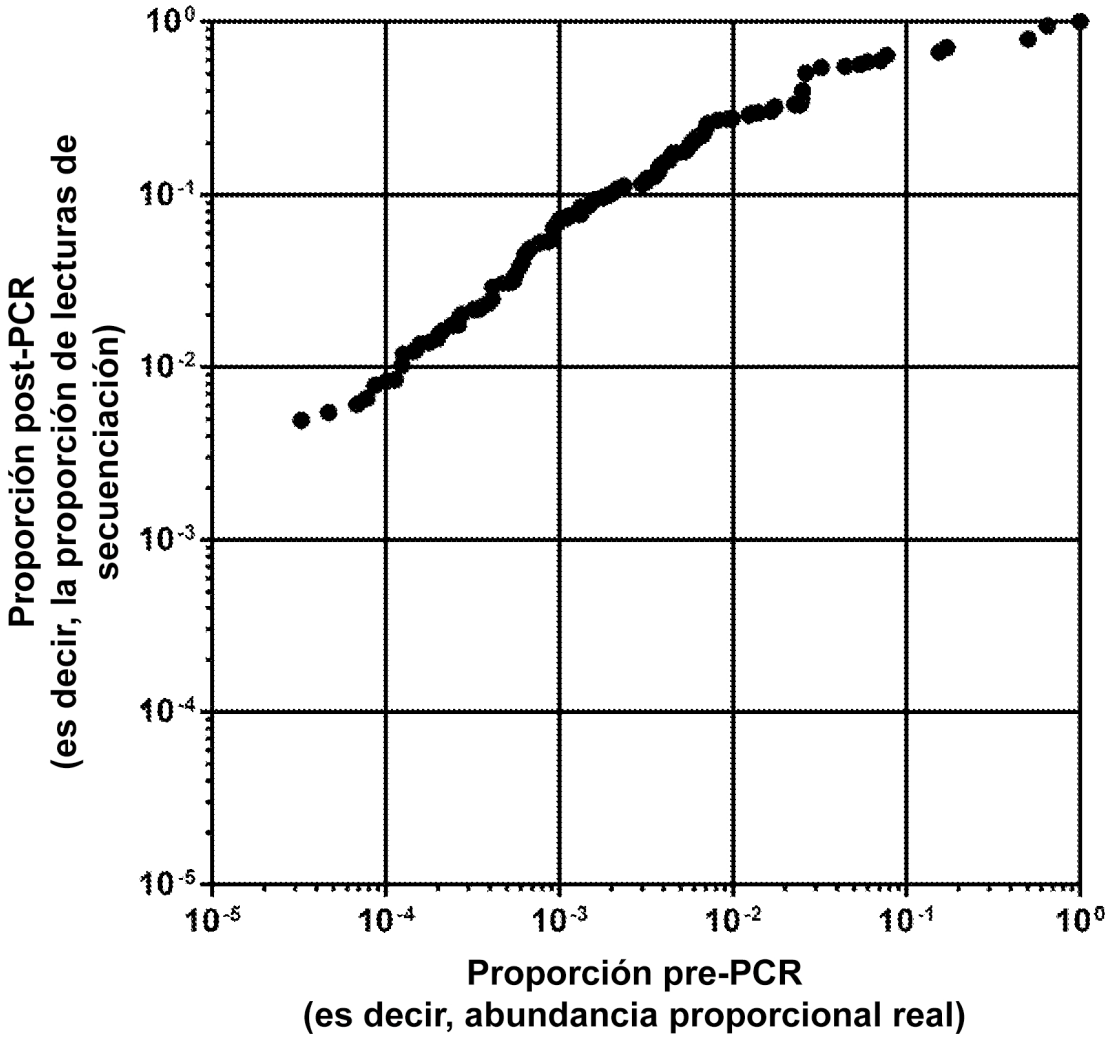
Proporción post-PCR
(es decir, la proporción de lecturas de secuenciación)
10^0
10^-1
10^-2
10^-3
10^-4
10^-5
10^-5
10^-4
10^-3
10^-2
10^-1
10^0
Proporción pre-PCR
(es decir, abundancia proporcional real)

FIG. 17A

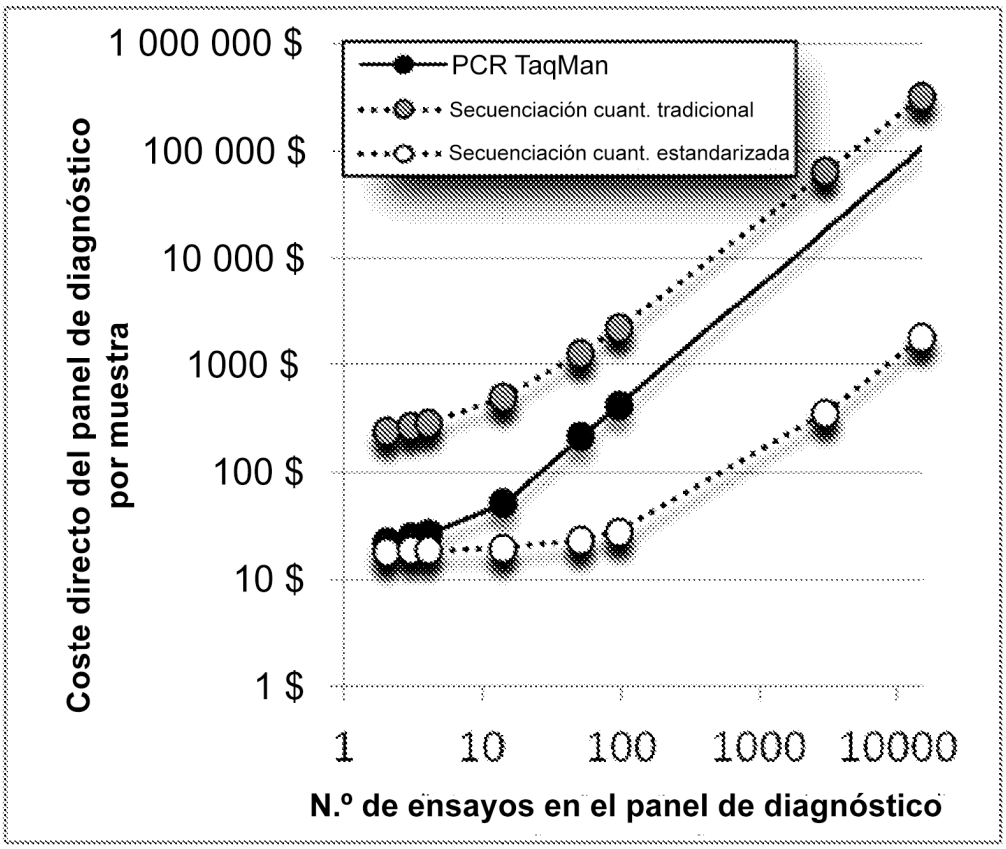
1 000 000 $
100 000 $
10 000 $
1000 $
100 $
10 $
1 $
PCR TaqMan
Secuenciación cuant. tradicional
Secuenciación cuant. estandarizada
Coste directo del panel de diagnóstico por muestra
1
10
100
1000
10000
N.º de ensayos en el panel de diagnóstico

FIG. 17B

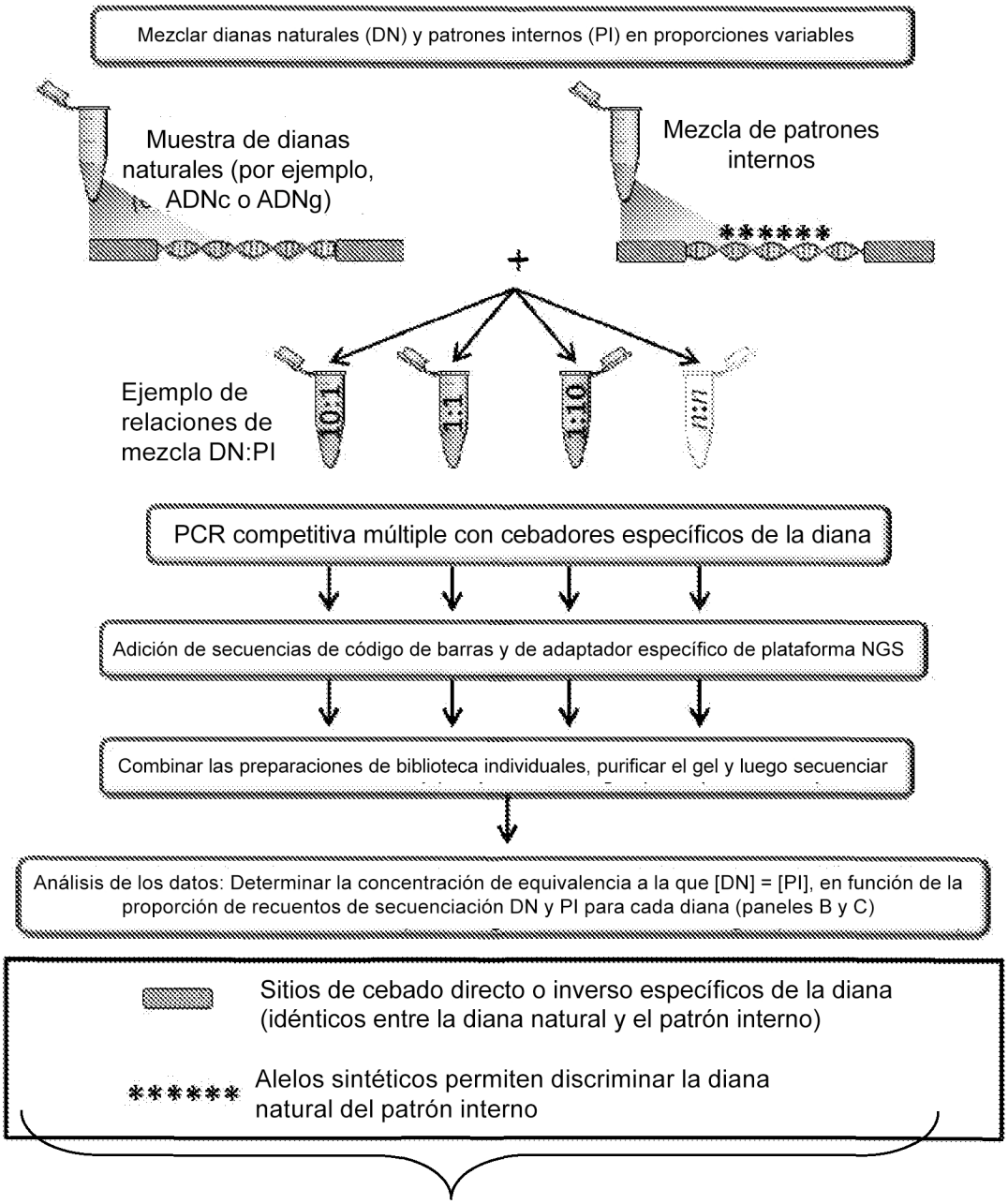
Mezclar dianas naturales (DN) y patrones internos (PI) en proporciones variables
Muestra de dianas naturales (por ejemplo, ADNc o ADNg)
Mezcla de patrones internos
Ejemplo de relaciones de mezcla DN:PI
10:1
1:1
1:10
n:n
PCR competitiva múltiple con cebadores específicos de la diana
Adición de secuencias de código de barras y de adaptador específico de plataforma NGS
Combinar las preparaciones de biblioteca individuales, purificar el gel y luego secuenciar
Análisis de los datos: Determinar la concentración de equivalencia a la que [DN] = [PI], en función de la proporción de recuentos de secuenciación DN y PI para cada diana (paneles B y C)
Sitios de cebado directo o inverso específicos de la diana (idénticos entre la diana natural y el patrón interno)
Alelos sintéticos permiten discriminar la diana natural del patrón interno

FIG. 18A

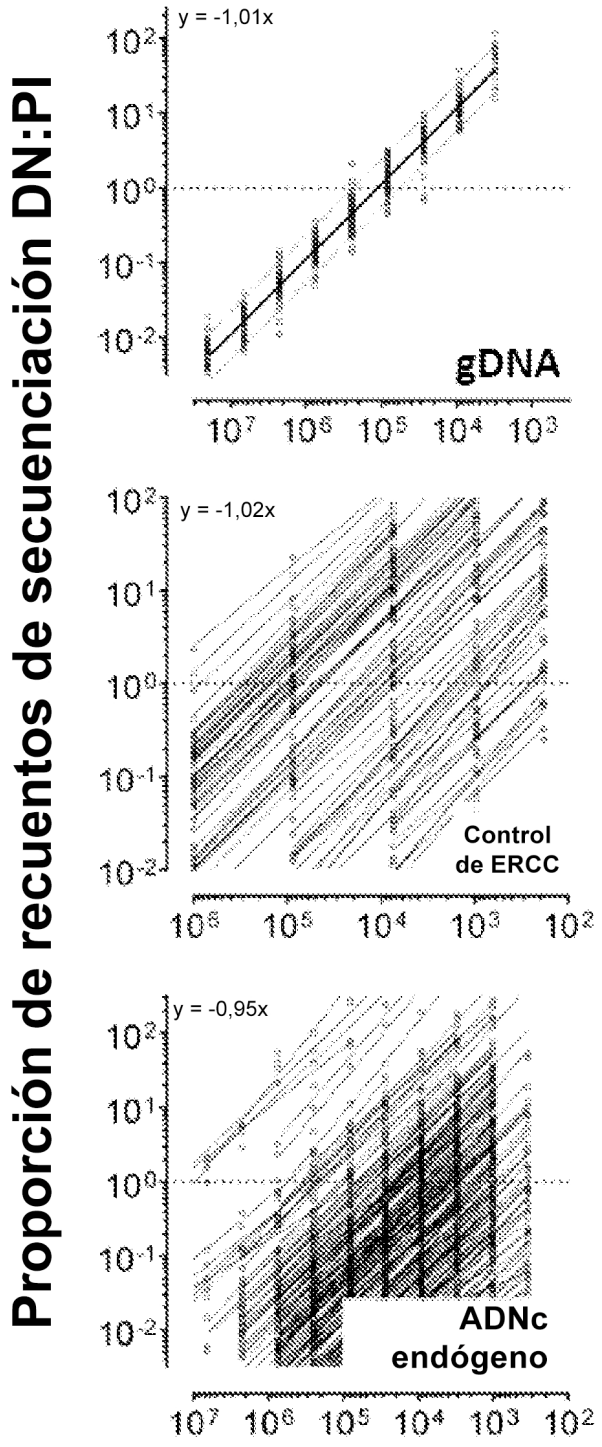
Proporción de recuentos de secuenciación DN:PI
y = -1,01x
gDNA
y = -1,02x
Control de ERCC
y = -0,95x
ADNc endógeno
Concentración de patrón interno (copias por preparación de biblioteca)

FIG. 18B

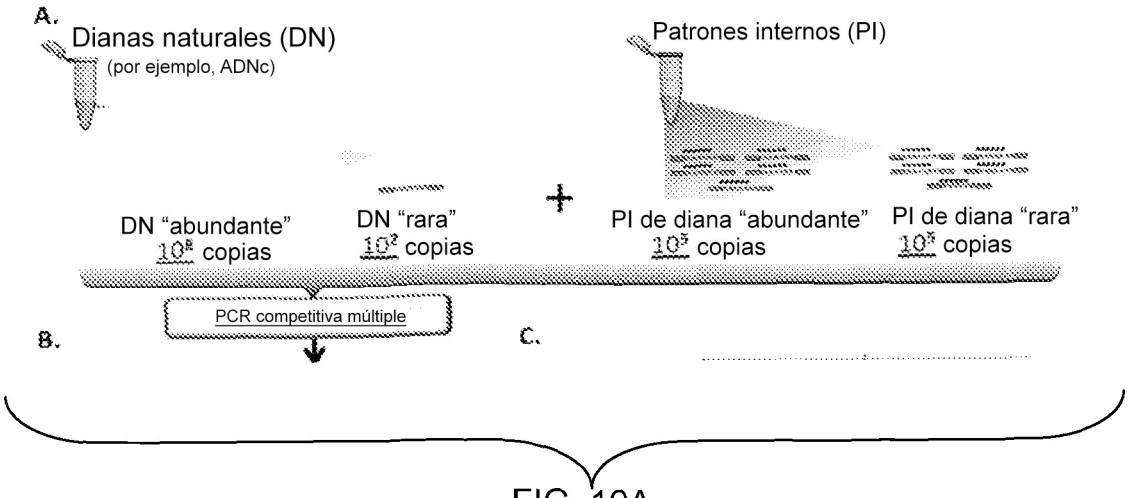

FIG. 19A

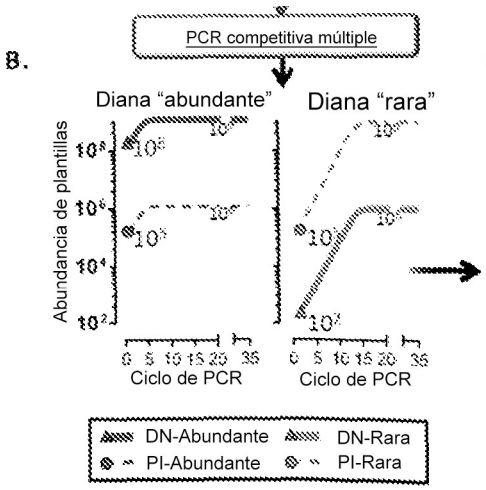

FIG. 19B

C.

	Antes de PCR		Después de PCR	
	Abundante	Rara	Abundante	Rara
Copias de diana natural (DN)	10^8	10^2	10^8	10^2
Copias de patrón interno (PI)	10^5	10^5	10^5	10^5
Proporción DN:PI	1000:1	1:1000	1000:1	1:1000
Intervalo de todas las plantillas	$(10^8/10^2)$ = **10^6**		$(10^8/10^5)$ = **10^3**	

FIG. 19C

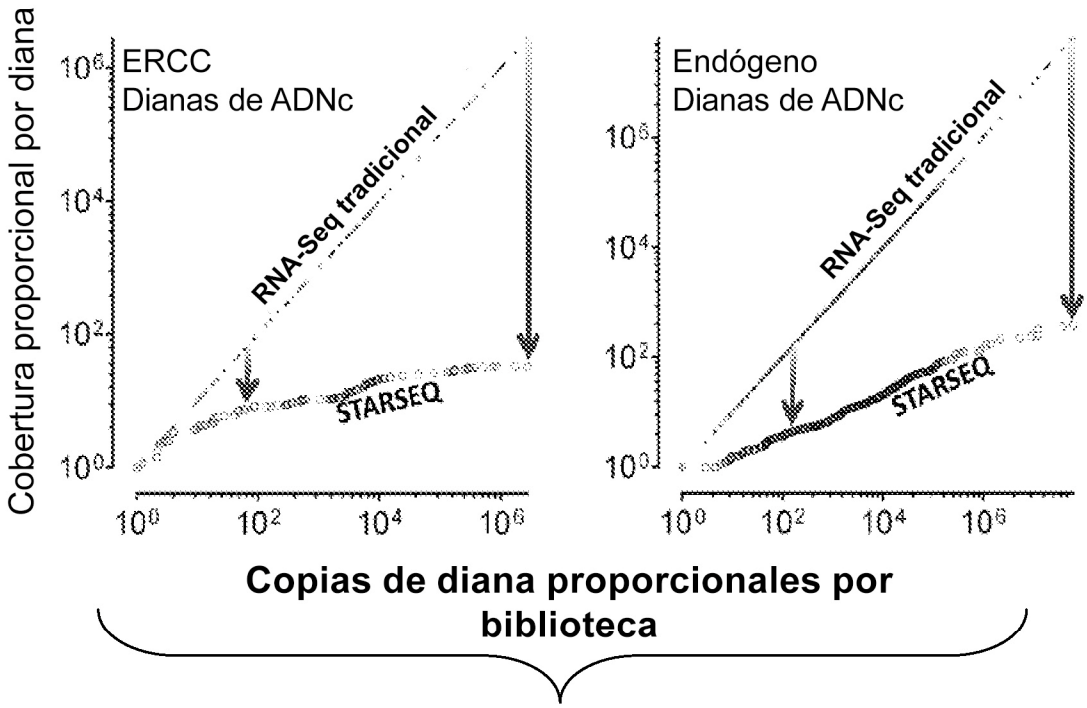

FIG. 20A

	Lecturas de secuenciación	
	ERCC	Endógeno
RNA-Seq tradicional	$1,0 \times 10^7$	$1,5 \times 10^8$
÷	÷	÷
STARSEQ	$1,5 \times 10^3$	$9,5 \times 10^3$
Reducción (veces)	**$6,9 \times 10^3$**	**$1,6 \times 10^4$**

FIG. 20B

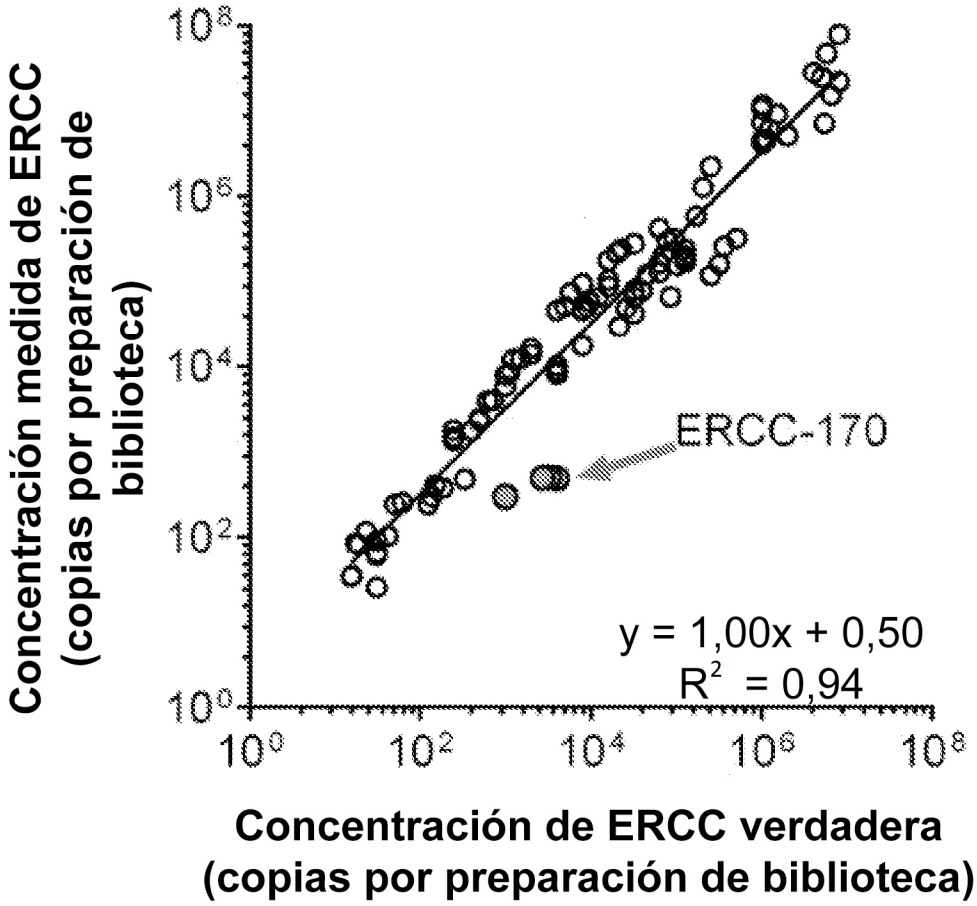
Concentración medida de ERCC (copias por preparación de biblioteca)
10^8
10^6
10^4
10^2
10^0
ERCC-170
y = 1,00x + 0,50
R^2 = 0,94
10^0
10^2
10^4
10^6
10^8
Concentración de ERCC verdadera (copias por preparación de biblioteca)

FIG. 21A

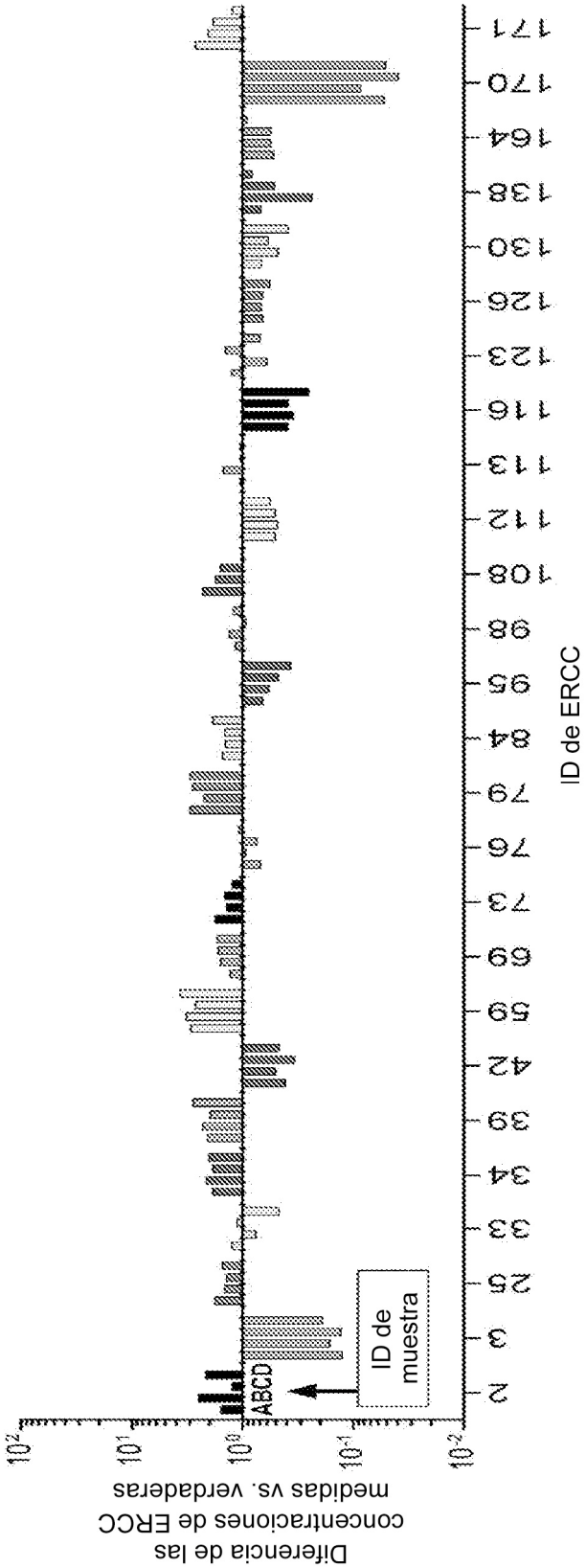
Diferencia de las concentraciones de ERCC medidas vs. verdaderas
10^{2}
10^{1}
10^{0}
10^{-1}
10^{-2}
ABCD
ID de muestra
2
3
25
33
34
39
42
59
69
73
76
79
84
95
98
108
112
113
116
123
126
130
138
164
170
171
ID de ERCC

FIG. 21B

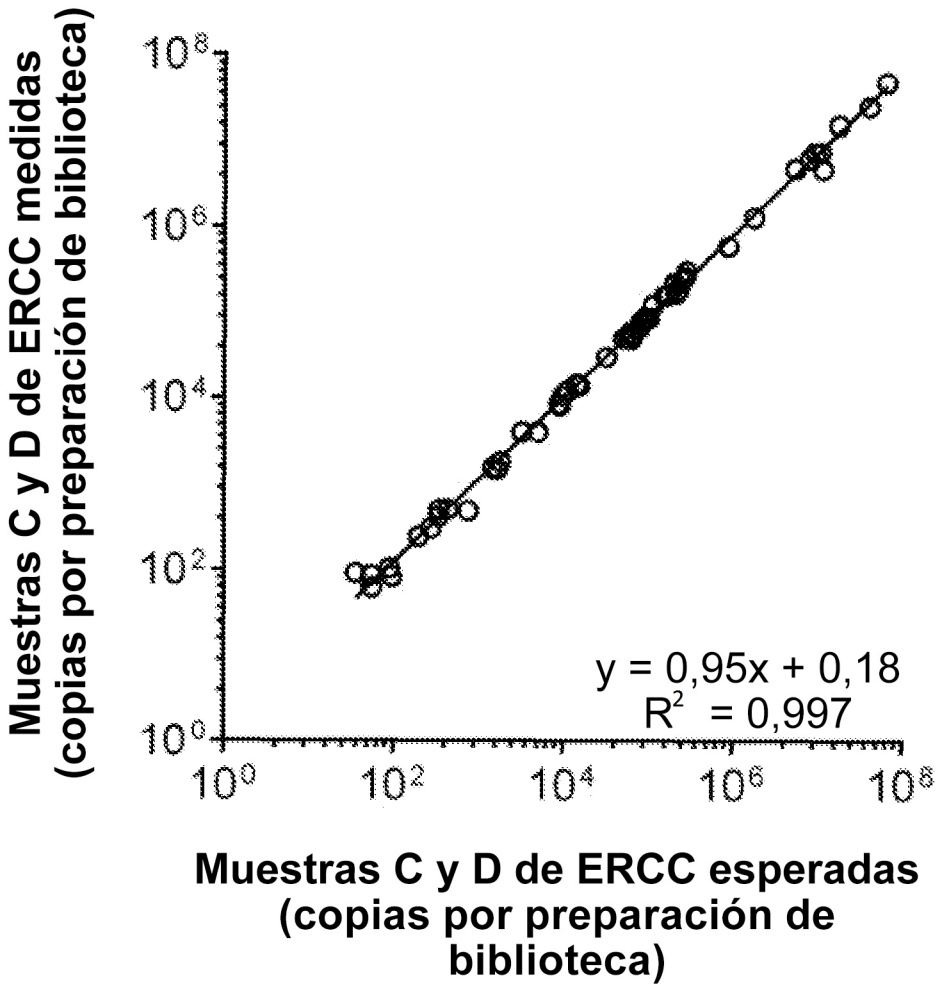
Muestras C y D de ERCC medidas (copias por preparación de biblioteca)
10^8
10^6
10^4
10^2
10^0
y = 0,95x + 0,18
R^2 = 0,997
10^0
10^2
10^4
10^6
10^8
Muestras C y D de ERCC esperadas (copias por preparación de biblioteca)

FIG. 21C

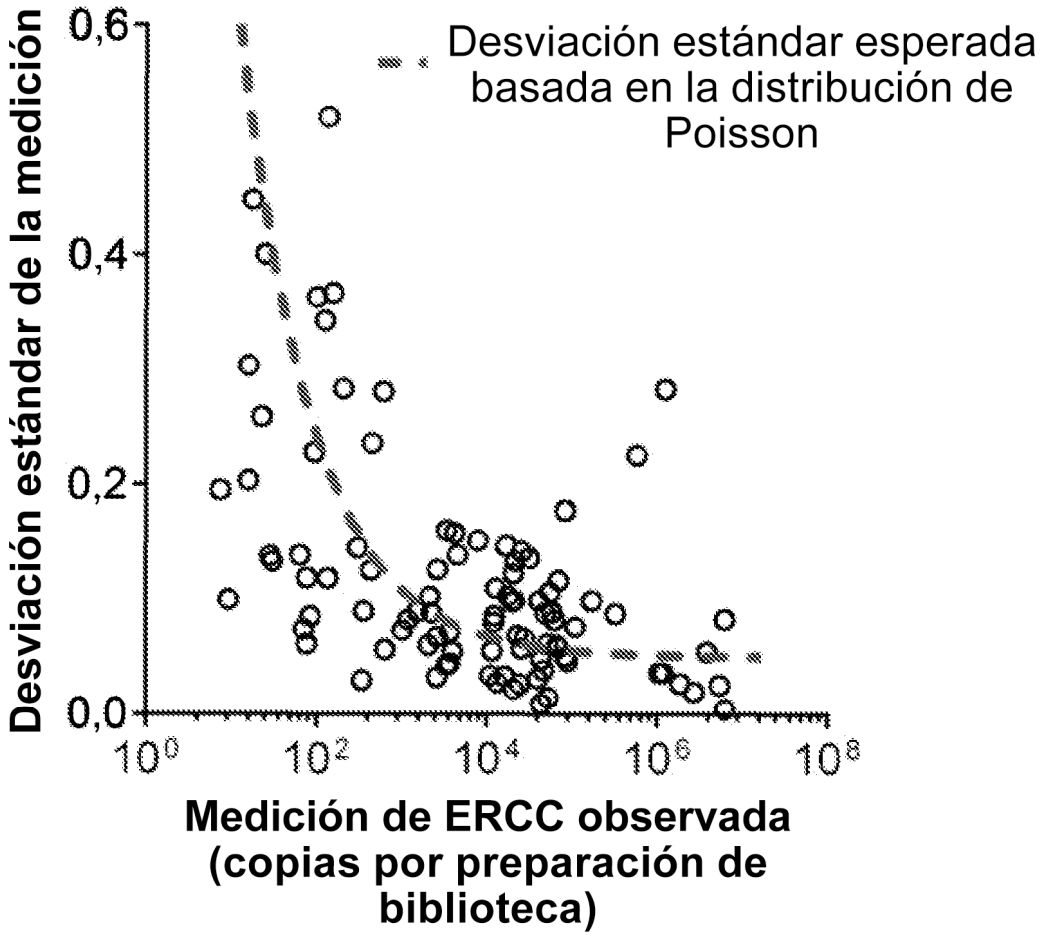
Desviación estándar esperada basada en la distribución de Poisson
Desviación estándar de la medición
0,6
0,4
0,2
0,0
10^0
10^2
10^4
10^6
10^8
Medición de ERCC observada (copias por preparación de biblioteca)

FIG. 21D

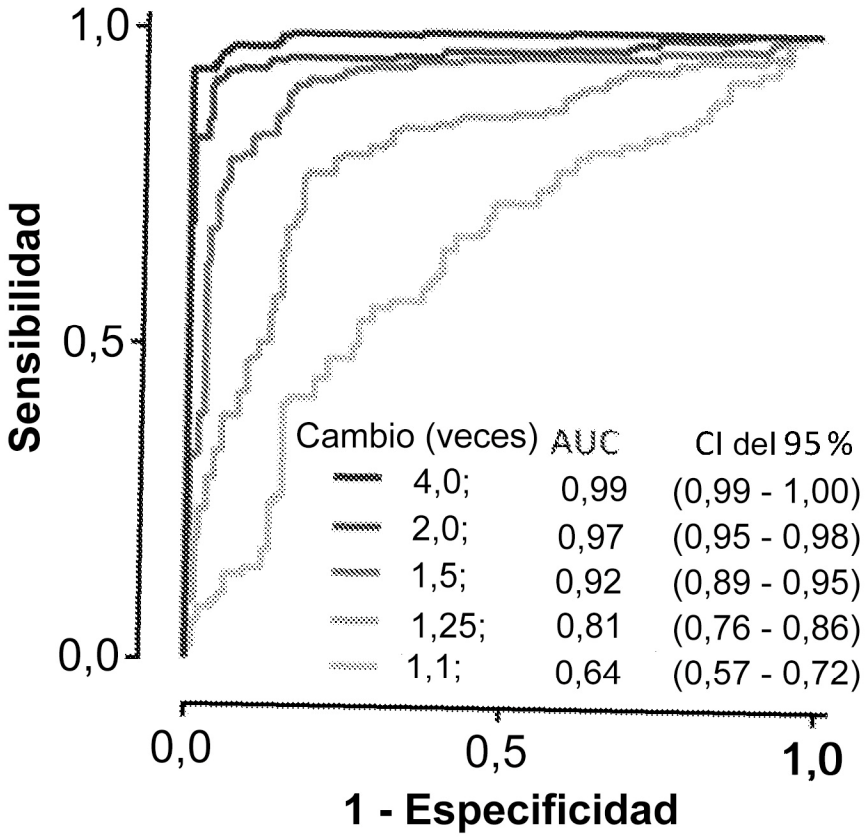
1,0
0,5
0,0
Sensibilidad
Cambio (veces)
AUC
CI del 95 %
4,0;
0,99
(0,99 - 1,00)
2,0;
0,97
(0,95 - 0,98)
1,5;
0,92
(0,89 - 0,95)
1,25;
0,81
(0,76 - 0,86)
1,1;
0,64
(0,57 - 0,72)
0,0
0,5
1,0
1 - Especificidad

FIG. 21E

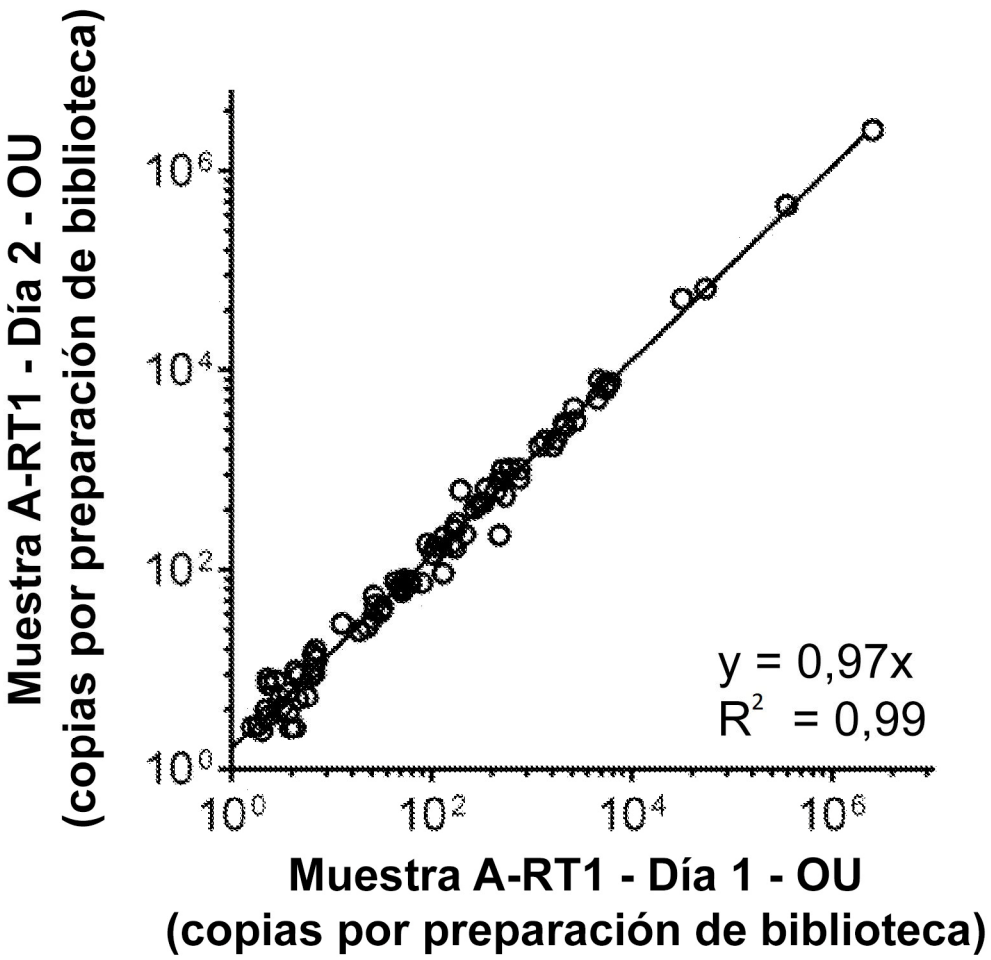
Muestra A-RT1 - Día 2 - OU
(copias por preparación de biblioteca)
10^6
10^4
10^2
10^0
y = 0,97x
R^2 = 0,99
10^0
10^2
10^4
10^6
Muestra A-RT1 - Día 1 - OU
(copias por preparación de biblioteca)

FIG. 22A

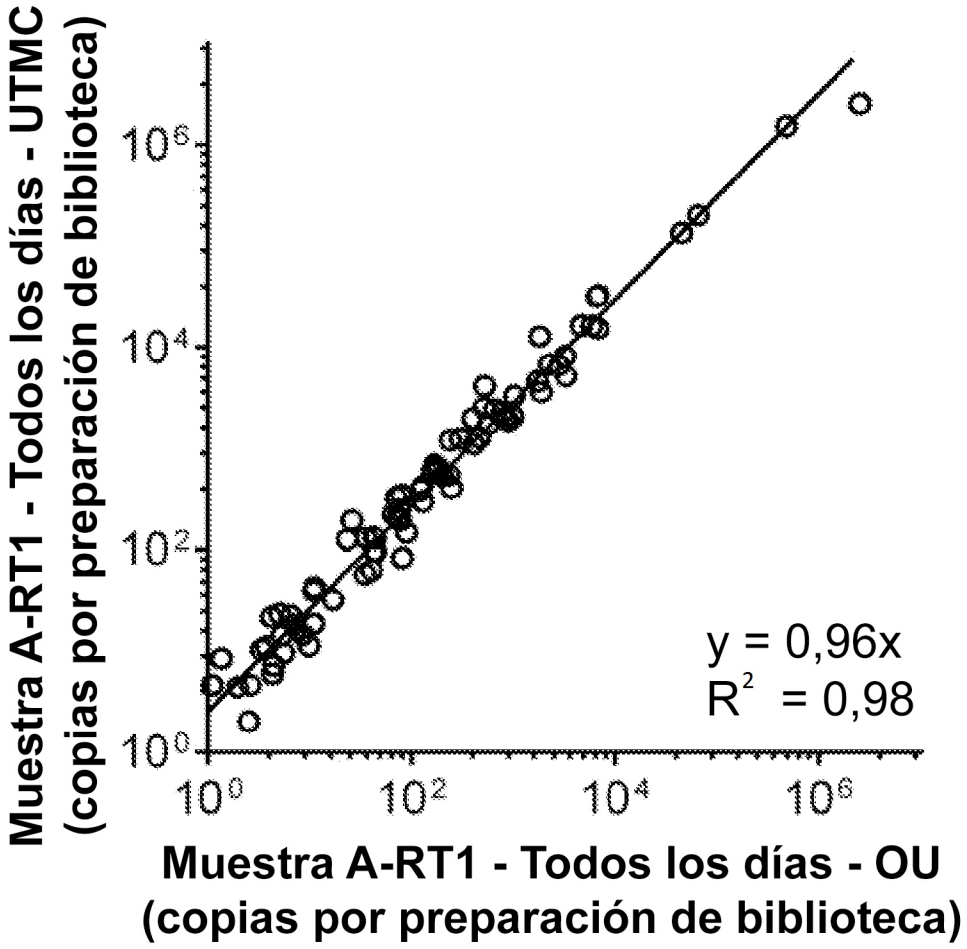
Muestra A-RT1 - Todos los días - UTMC
(copias por preparación de biblioteca)
10^6
10^4
10^2
10^0
y = 0,96x
R^2 = 0,98
10^0
10^2
10^4
10^6
Muestra A-RT1 - Todos los días - OU
(copias por preparación de biblioteca)

FIG. 22B

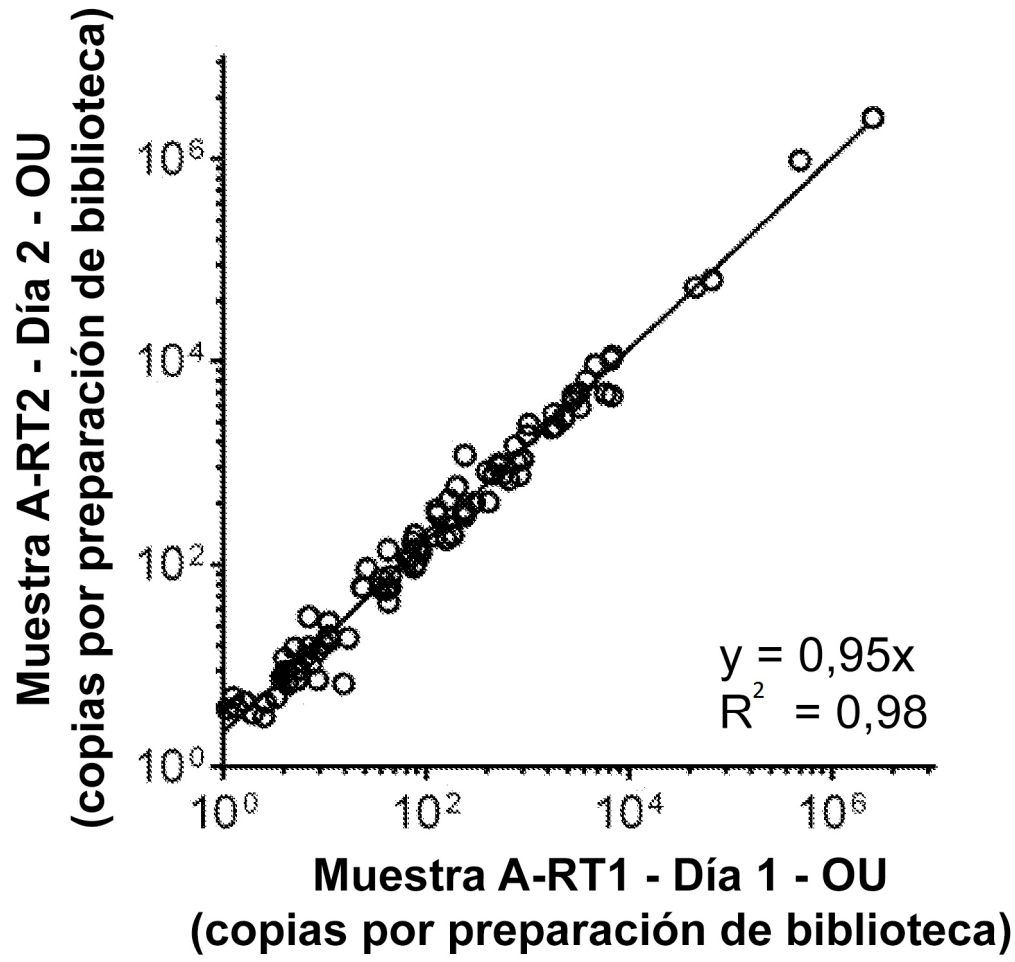
Muestra A-RT2 - Día 2 - OU
(copias por preparación de biblioteca)
10^6
10^4
10^2
10^0
y = 0,95x
R^2 = 0,98
10^0
10^2
10^4
10^6
Muestra A-RT1 - Día 1 - OU
(copias por preparación de biblioteca)

FIG. 22C

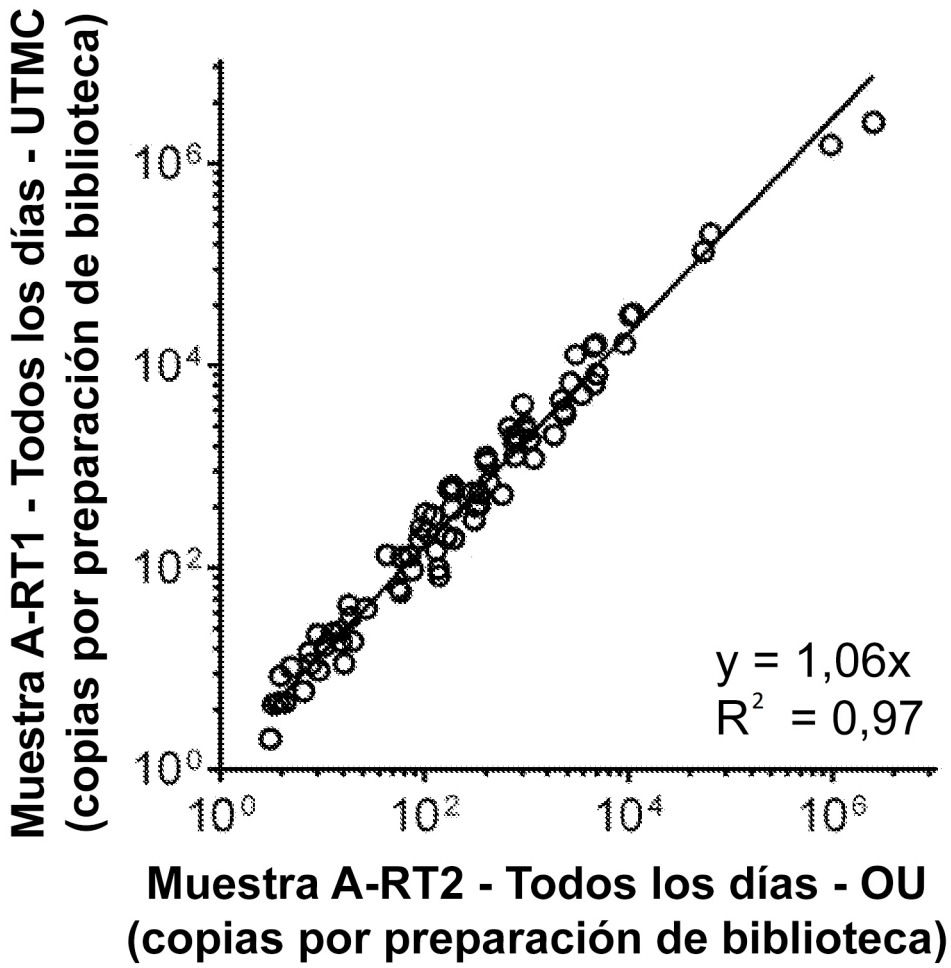
Muestra A-RT1 - Todos los días - UTMC
(copias por preparación de biblioteca)
10^6
10^4
10^2
10^0
y = 1,06x
R^2 = 0,97
10^0
10^2
10^4
10^6
Muestra A-RT2 - Todos los días - OU
(copias por preparación de biblioteca)

FIG. 22D

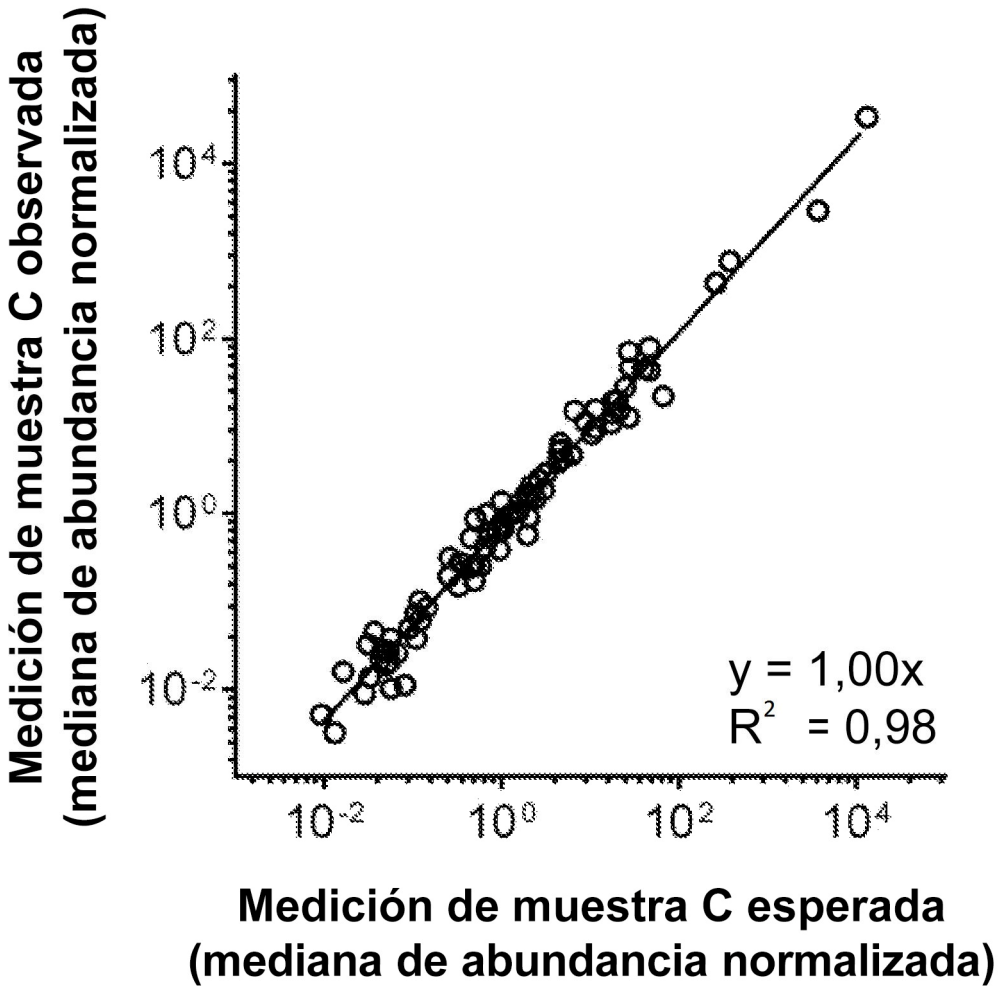
Medición de muestra C observada
(mediana de abundancia normalizada)
10^4
10^2
10^0
10^-2
y = 1,00x
R^2 = 0,98
10^-2
10^0
10^2
10^4
Medición de muestra C esperada
(mediana de abundancia normalizada)

FIG. 22E

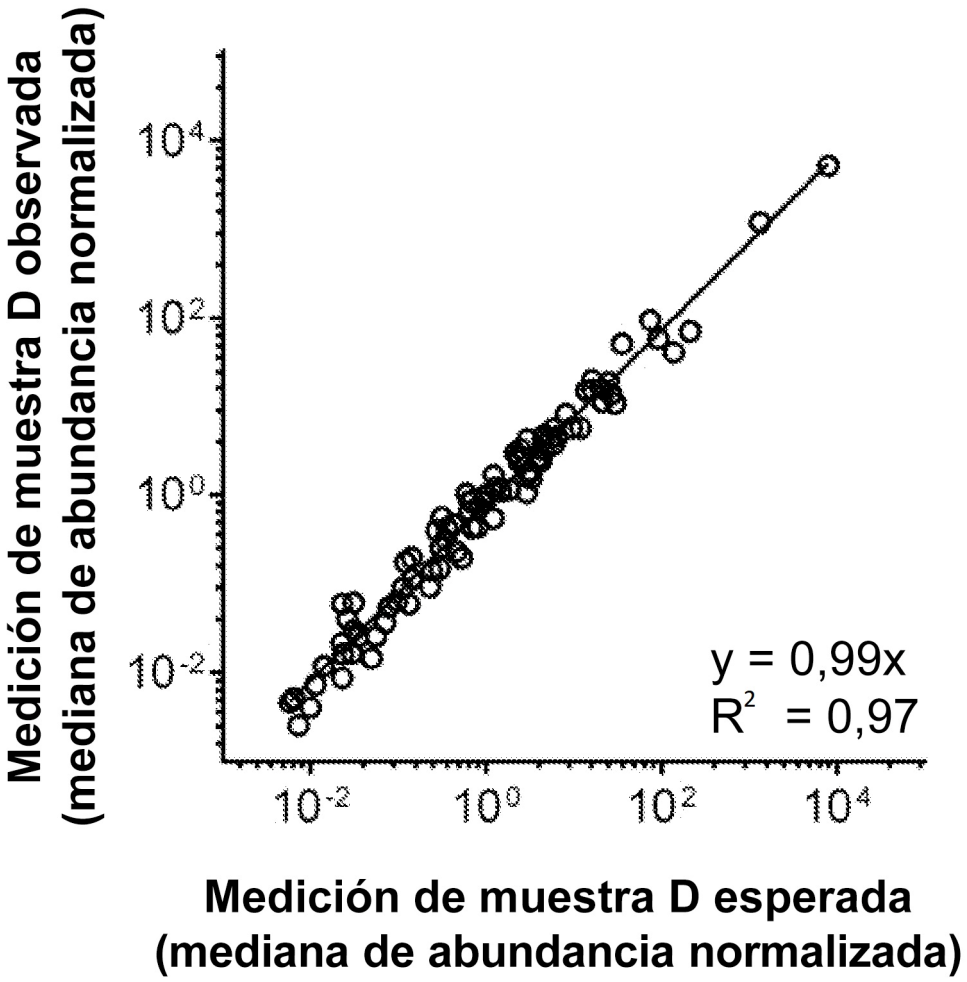
Medición de muestra D observada
(mediana de abundancia normalizada)
10^4
10^2
10^0
10^-2
y = 0,99x
R^2 = 0,97
10^-2
10^0
10^2
10^4
Medición de muestra D esperada
(mediana de abundancia normalizada)

FIG. 22F

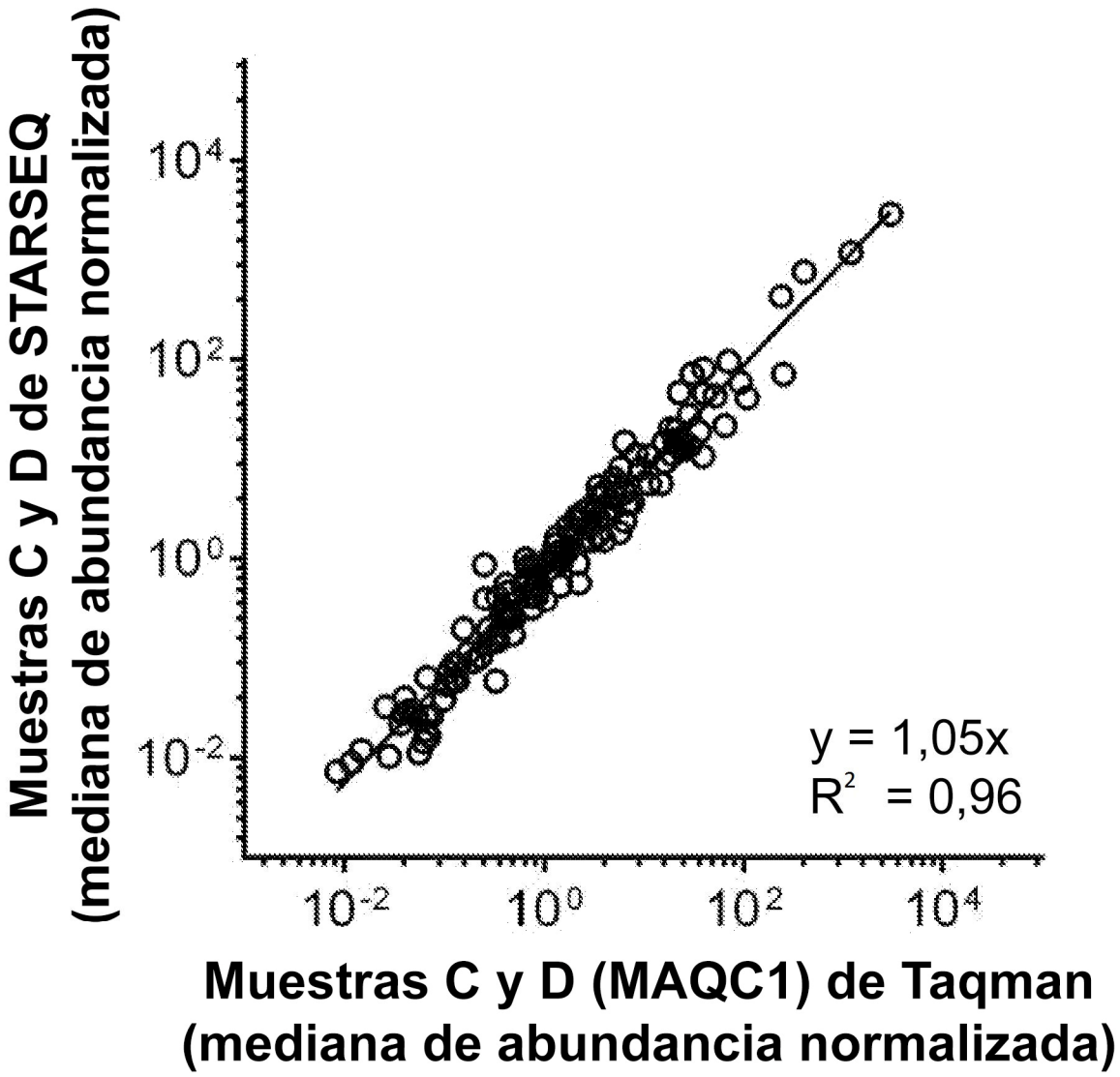
Muestras C y D de STARSEQ
(mediana de abundancia normalizada)
10^4
10^2
10^0
10^-2
y = 1,05x
R^2 = 0,96
10^-2
10^0
10^2
10^4
Muestras C y D (MAQC1) de Taqman
(mediana de abundancia normalizada)

FIG. 23A

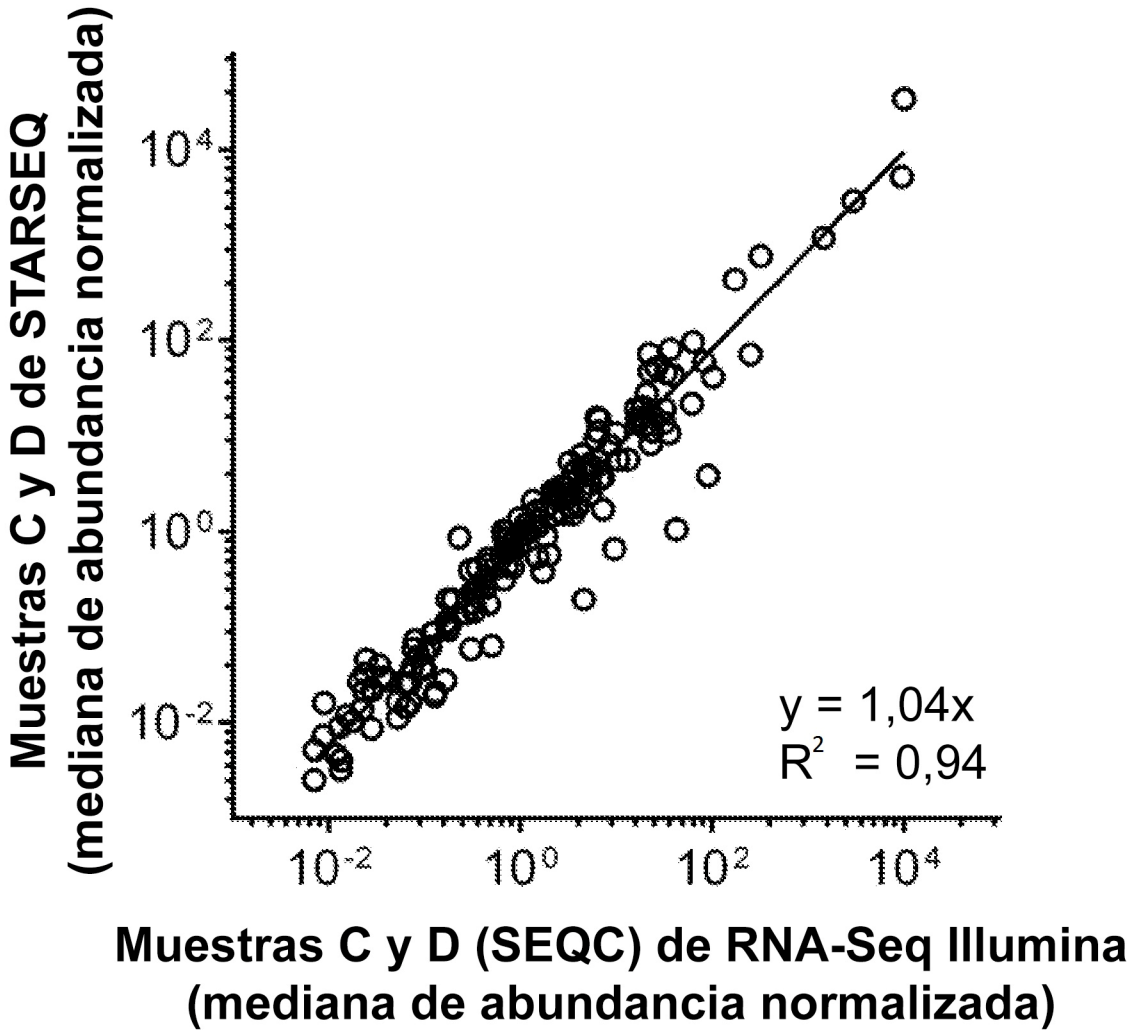
Muestras C y D de STARSEQ
(mediana de abundancia normalizada)
10^4
10^2
10^0
10^-2
y = 1,04x
R^2 = 0,94
10^-2
10^0
10^2
10^4
Muestras C y D (SEQC) de RNA-Seq Illumina
(mediana de abundancia normalizada)

FIG. 23B

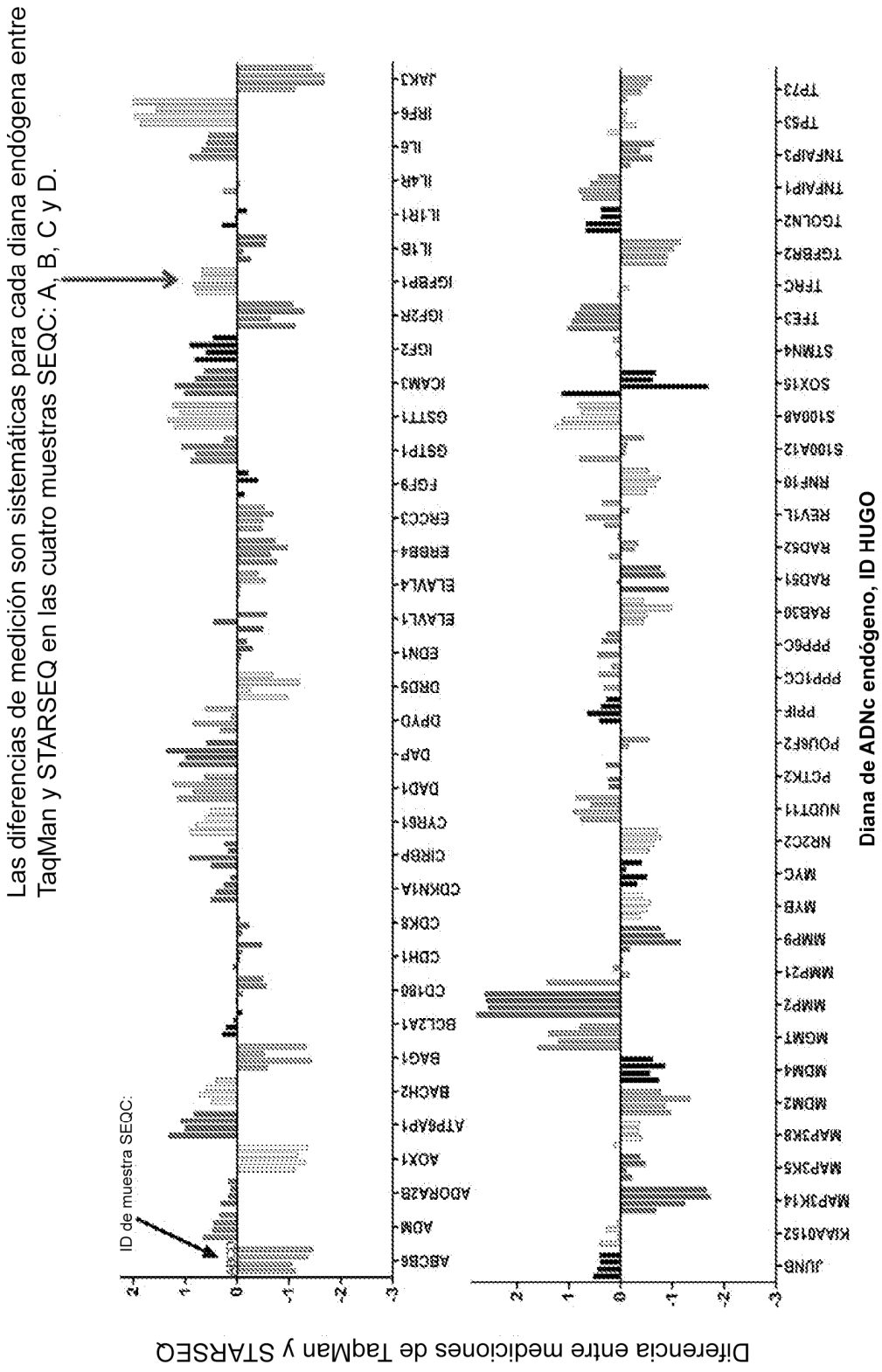
Las diferencias de medición son sistemáticas para cada diana endógena entre TaqMan y STARSEQ en las cuatro muestras SEQC: A, B, C y D.
ID de muestra SEQC:
Diferencia entre mediciones de TaqMan y STARSEQ
Diana de ADNc endógeno, ID HUGO

FIG. 24

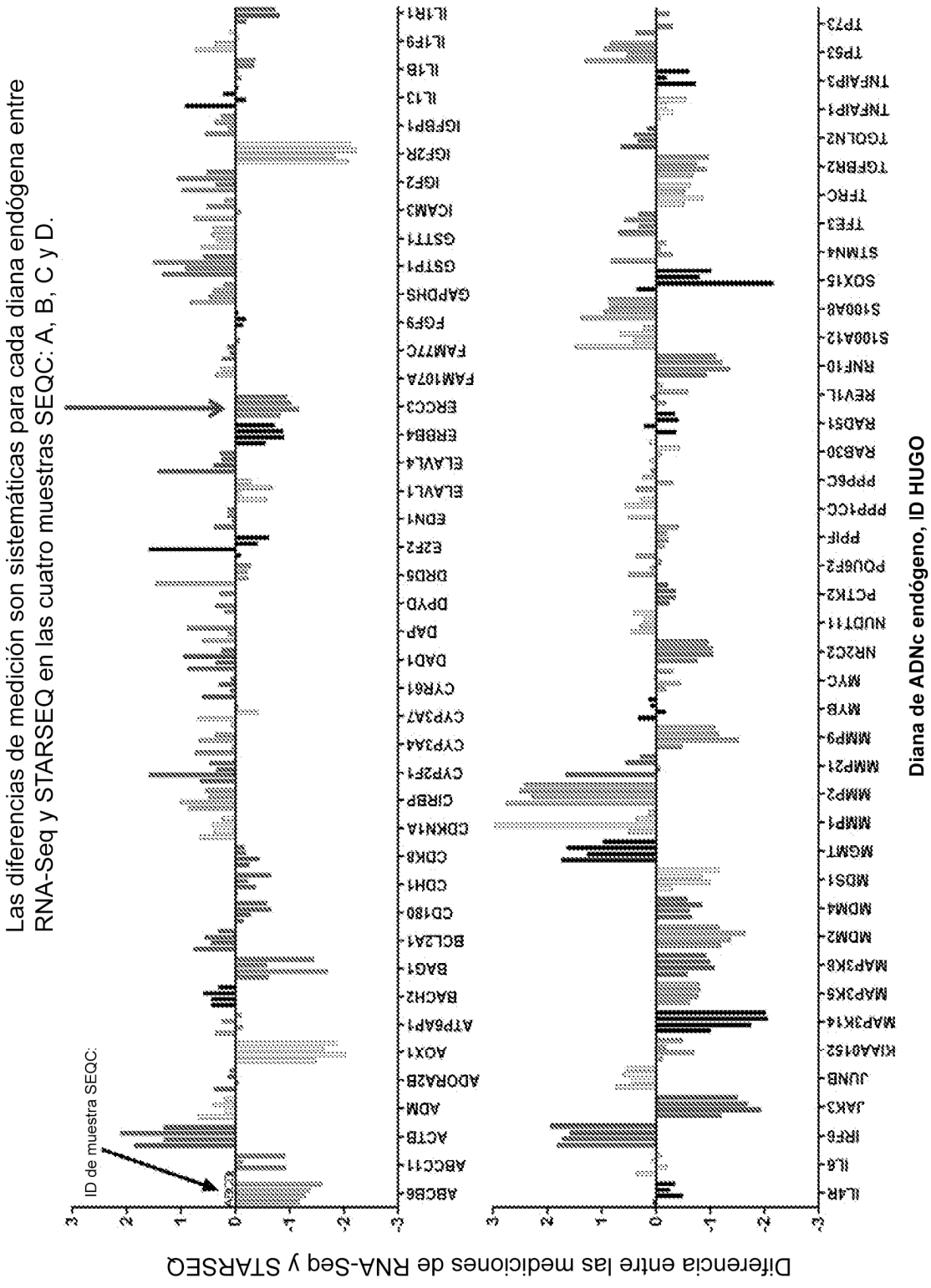
Las diferencias de medición son sistemáticas para cada diana endógena entre RNA-Seq y STARSEQ en las cuatro muestras SEQC: A, B, C y D.
ID de muestra SEQC:
Diferencia entre las mediciones de RNA-Seq y STARSEQ
Diana de ADNc endógeno, ID HUGO

FIG. 25

Mediciones de dianas ERCC	A	B	C	D
Válida	26	26	26	26
Negativo verdadero	0	0	0	0
Fallida	1	1	1	1

Mediciones de dianas endógenas	A	B	C	D	ADN_g
Válida	89	99	96	107	123
Negativo verdadero	4	12	6	4	0
Fallida	20	12	21	12	0

FIG. 26

ID HUGO y de muestra	STARSEQ	Taqman	RNA-Seq
RPL3L Muestra D	< -2,4879	-3,24	-2,90
ANXA13 Muestra C	< -2,127	-3,20	-2,25
MMP7 Muestra C	< -1,7003	-3,27	-2,23
KRT24 Muestra D	< -1,6589	-3,00	-2,31
KIAA0101 Muestra B	< -1,3856	-2,24	-1,64
KRT24 Muestra B	< -1,2609	-2,83	-1,88
STMN2 Muestra A	< -1,0849	-1,47	-1,92
IL18R1 Muestra B	< -0,938	-2,28	-0,81
DPP4 Muestra B	< -0,733	-1,64	-0,61
IL18R1 Muestra D	< -0,7132	-2,37	-1,10
KIT Muestra C	< -0,0946	-0,80	-0,18
KIT Muestra A	< 0,1187	-1,21	-0,53
IL1F6 Muestra B	< -1,9726	ND	ND
POU1F1 Muestra C	< -1,9169	ND	-3,95
POU1F1 Muestra B	< -1,8552	ND	-2,62
RPL3L Muestra B	< -1,8183	ND	-3,15
KRT24 Muestra C	< -1,6485	ND	-2,72
POU1F1 Muestra A	< -1,5431	ND	-4,18
CYP2C9 Muestra D	< -1,1907	ND	-3,41
ABCB11 Muestra B	< -1,1123	ND	-2,38
SERPINB7 Muestra B	< -1,0772	ND	-3,15
DLG7 Muestra B	< -0,8081	ND	-2,31
FOXA1 Muestra B	< -0,4969	ND	-2,42
CYP2C9 Muestra C	< 0,37	ND	-3,20
CYP2C9 Muestra B	< 0,6952	ND	-2,92
CYP2C9 Muestra A	< 0,7906	ND	-3,16

FIG. 27

DE intraensayo; intramuestra	0,08
DE intraensayo; intermuestra	0,18
DE interensayo; intermuestra	0,40

FIG. 28