

19



OFICINA ESPAÑOLA DE
PATENTES Y MARCAS

ESPAÑA



11 Número de publicación: **3 014 093**

51 Int. Cl.:

C12Q 1/6816 (2008.01)

C12Q 1/6841 (2008.01)

C12Q 1/6806 (2008.01)

C12Q 1/6869 (2008.01)

C12Q 1/6844 (2008.01)

12

TRADUCCIÓN DE PATENTE EUROPEA

T3

86 Fecha de presentación y número de la solicitud internacional: **29.07.2015 PCT/US2015/042556**

87 Fecha y número de publicación internacional: **04.02.2016 WO16018960**

96 Fecha de presentación y número de la solicitud europea: **29.07.2015 E 15827358 (1)**

97 Fecha y número de publicación de la concesión europea: **19.02.2025 EP 3175023**

54 Título: **Método para determinar ácidos nucleicos**

30 Prioridad:

30.07.2014 US 201462031062 P

15.09.2014 US 201462050636 P

03.04.2015 US 201562142653 P

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

16.04.2025

73 Titular/es:

**PRESIDENT AND FELLOWS OF HARVARD
COLLEGE (100.00%)**

**17 Quincy Street
Cambridge, MA 02138, US**

72 Inventor/es:

**ZHUANG, XIAOWEI;
CHEN, KOK-HAO;
BOETTIGER, ALISTAIR;
MOFFITT, JEFFREY R. y
WANG, SIYUAN**

74 Agente/Representante:

QUIRÓS MARÍN, María

ES 3 014 093 T3

Aviso: En el plazo de nueve meses a contar desde la fecha de publicación en el Boletín Europeo de Patentes, de la mención de concesión de la patente europea, cualquier persona podrá oponerse ante la Oficina Europea de Patentes a la patente concedida. La oposición deberá formularse por escrito y estar motivada; sólo se considerará como formulada una vez que se haya realizado el pago de la tasa de oposición (art. 99.1 del Convenio sobre Concesión de Patentes Europeas).

DESCRIPCIÓN

Método para determinar ácidos nucleicos

5 **Campo**

La presente invención se refiere generalmente a sistemas y métodos para determinar u obtener imágenes de ácidos nucleicos, por ejemplo, dentro de células. En algunas realizaciones, puede determinarse el transcriptoma de una célula.

10

Antecedentes

La hibridación fluorescente *in situ* de una sola molécula (smFISH) es un poderoso método para detectar moléculas de ARNm individuales en células. La alta eficiencia de detección y el gran rango dinámico de este método proporcionan un detalle exquisito del estado de expresión, la distribución espacial dentro de células y tejidos intactos, y la variación entre células de ARNm individuales. Tales enfoques han sido esenciales para muchas ideas recientes sobre la comprensión de la regulación y expresión génica. Sin embargo, una limitación fundamental de smFISH es su bajo rendimiento, normalmente de tan sólo unos pocos genes a la vez. Este bajo rendimiento se debe a la falta de sondas distinguibles con las que marcar las células y al coste de producción de grandes cantidades de sondas marcadas requeridas para una tinción de alta eficiencia. Por tanto, son necesarias mejoras en la detección de moléculas de ARNm.

E. Lubeck *et al.* divulgan en "Single-cell *in situ* RNA profiling by sequential hybridization" (en: Nature Methods, vol. 11, n.º 4, págs. 360 y 361, abril de 2014) la obtención de perfiles de ARN *in situ* de una sola célula por medio de una técnica de hibridación secuencial. El enfoque propone exponer iterativamente una muestra a un conjunto de sondas de hibridación que tienen un solo tipo de fluoróforo en una iteración. Se obtienen imágenes de la muestra y se eliminan las sondas. En una siguiente iteración, se expone la muestra a otro conjunto de sondas de hibridación que tienen un solo tipo de fluoróforo, pero distinto, en comparación con la etapa anterior.

El documento WO 03/003810 A2 divulga la detección y cuantificación de analitos en mezclas complejas, basándose en la técnica específica de electroestiramiento de ADN. En tal ADN sometido a estiramiento, deben detectarse secuencias diana espacialmente separadas. El propósito es identificar individualmente múltiples marcadores en una sola molécula de sonda de ADN.

El documento US 2014/0031243 A1 divulga la detección de especies moleculares en células mediante obtención de imágenes de superresolución y marcaje combinatorio.

Sumario

Los problemas mencionados se resuelven mediante el contenido de las reivindicaciones independientes. En las reivindicaciones dependientes se definen realizaciones preferidas adicionales.

Breve descripción de los dibujos

Se describirán realizaciones no limitativas de la presente invención a modo de ejemplo con referencia a las figuras adjuntas, que son esquemáticas y no se pretende que estén dibujadas a escala. En las figuras, cada componente idéntico o prácticamente idéntico ilustrado está representado normalmente por un número único. Por motivos de claridad, no se marca cada componente en cada figura, ni se muestra cada componente de cada realización de la invención donde la ilustración no es necesaria para permitir la comprensión de la invención a los expertos habituales en la técnica. En las figuras:

las figuras 1A-1B ilustran un esquema de codificación para sondas de ácido nucleico, en determinadas realizaciones de la invención;

la figura 1C ilustra un esquema de codificación para sondas de ácido nucleico, en otro ejemplo;

las figuras 2A-2G ilustran la determinación de ARNm en una célula, en algunas realizaciones de la invención;

las figuras 3A-3B ilustran la determinación de ácidos nucleicos, según diversas realizaciones de la invención;

las figuras 4A-4B son un ejemplo no limitativo de múltiples secuencias de lectura distribuidas en una población de sondas de ácido nucleico diferentes, según determinadas realizaciones de la invención;

las figuras 5A-5E ilustran la determinación de ácidos nucleicos, según otra realización de la invención;

las figuras 6A-6H ilustran la determinación simultánea de múltiples especies de ácido nucleico en células, en

determinadas realizaciones de la invención;

las figuras 7A-7E muestran el ruido de expresión de genes y la covariación de la expresión entre diferentes genes determinados según algunas realizaciones de la invención;

las figuras 8A-8E ilustran la distribución espacial de ARN en células determinada según una realización de la invención;

las figuras 9A-9C ilustran la determinación simultánea de múltiples especies de ácido nucleico en células, en otra realización de la invención;

las figuras 10A-10B muestran la expresión entre diferentes genes determinada según aún otra realización de la invención;

la figura 11 es una descripción esquemática de marcaje combinatorio, según otra realización de la invención;

la figura 12 muestra descripciones esquemáticas de la distancia de Hamming, en otra realización de la invención;

la figura 13 ilustra la producción de una biblioteca de sondas, en todavía otra realización de la invención;

las figuras 14A-14B ilustran determinaciones de puntos fluorescentes, en otra realización de la invención;

las figuras 15A-15B ilustran que la corrección de error facilita la detección de ARN, en aún otra realización de la invención;

las figuras 16A-16B muestran la caracterización de tasas de identificaciones erróneas y tasas de identificaciones correctas, en una realización de la invención;

las figuras 17A-17D muestran la caracterización de tasas de identificaciones erróneas y tasas de identificaciones correctas, en otra realización de la invención;

las figuras 18A-18C muestran una comparación de experimentos, según una realización de la invención; y

las figuras 19A-9D ilustran la evaluación de errores y decodificación, en otra realización de la invención.

Breve descripción de las secuencias

SEQ ID NO: 1 es:

GTTGGCGACGAAAGCACTGCGATTGGAACCGTCCCAAGCGTTGCGCTTAATGGA
TCATCAATTTTGTCTCACTACGACGGTCAATCGCGCTGCATACTTGCCTCGGTGCG
GACAAACGAGG;

SEQ ID NO: 2 es CGCAACGCTTGGGACGGTTCCAATCGGATC;

SEQ ID NO: 3 es CGAATGCTCTGGCCTCGAACGAACGATAGC;

SEQ ID NO: 4 es ACAAATCCGACCAGATCGGACGATCATGGG;

SEQ ID NO: 5 es CAAGTATGCAGCGCGATTGACCGTCTCGTT;

SEQ ID NO: 6 es TCGTCTGCTGGCTAGCACGGCACGCAAAT;

SEQ ID NO: 7 es AAGTCGTACGCCGATGCGCAGCAATTCACT;

SEQ ID NO: 8 es CGAAACATCGGCCACGGTCCCGTTGAACTT;

SEQ ID NO: 9 es ACGAATCCACCGTCCAGCGCGTCAAACAGA;

SEQ ID NO: 10 es CGCGAAATCCCCGTAACGAGCGTCCCTTGC;

ES 3 014 093 T3

SEQ ID NO: 11 es GCATGAGTTGCCTGGCGTTGCGACGACTAA;

SEQ ID NO: 12 es CCGTCGTCTCCGGTCCACCGTTGCGCTTAC;

5 SEQ ID NO: 13 es GGCCAATGGCCCAGGTCCGTACGCAATTT;

SEQ ID NO: 14 es TTGATCGAATCGGAGCGTAGCGGAATCTGC;

10 SEQ ID NO: 15 es CGCGCGGATCCGCTTGTCGGGAACGGATAC;

SEQ ID NO: 16 es GCCTCGATTACGACGGATGTAATTCGGCCG;

SEQ ID NO: 17 es GCCCGTATTCCCGCTTGCGAGTAGGGCAAT;

15 SEQ ID NO: 18 es GTTGGTCGGCACTTGGGTGC;

SEQ ID NO: 19 es CGATGCGCCAATTCCGGTTC;

SEQ ID NO: 20 es CGCGGGCTATATGCGAACCG;

20 SEQ ID NO: 21 es TAATACGACTCACTATAGGGAAAGCCGGTTCATCCGGTGG;

SEQ ID NO: 22 es TAATACGACTCACTATAGGGTGATCATCGCTCGCGGGTTG;

25 SEQ ID NO: 23 es TAATACGACTCACTATAGGGCGTGGAGGGCATAACAACGC;

SEQ ID NO: 24 es CGCAACGCTTGGGACGGTTCCAATCGGATC/3Cy5Sp/;

SEQ ID NO: 25 es CGAATGCTCTGGCCTCGAACGAACGATAGC/3Cy5Sp/;

30 SEQ ID NO: 26 es ACAAATCCGACCAGATCGGACGATCATGGG/3Cy5Sp/;

SEQ ID NO: 27 es CAAGTATGCAGCGCGATTGACCGTCTCGTT/3Cy5Sp/;

35 SEQ ID NO: 28 es GCGGGAAGCACGTGGATTAGGGCATCGACC/3Cy5Sp/;

SEQ ID NO: 29 es AAGTCGTACGCCGATGCGCAGCAATTCCT/3Cy5Sp/;

SEQ ID NO: 30 es CGAAACATCGGCCACGGTCCCGTTGAACTT/3Cy5Sp/;

40 SEQ ID NO: 31 es ACGAATCCACCGTCCAGCGCGTCAAACAGA/3Cy5Sp/;

SEQ ID NO: 32 es CGCGAAATCCCCGTAACGAGCGTCCCTTGC/3Cy5Sp/;

45 SEQ ID NO: 33 es GCATGAGTTGCCTGGCGTTGCGACGACTAA/3Cy5Sp/;

SEQ ID NO: 34 es CCGTCGTCTCCGGTCCACCGTTGCGCTTAC/3Cy5Sp/;

SEQ ID NO: 35 es GGCCAATGGCCCAGGTCCGTACGCAATTT/3Cy5Sp/;

50 SEQ ID NO: 36 es TTGATCGAATCGGAGCGTAGCGGAATCTGC/3Cy5Sp/;

SEQ ID NO: 37 es CGCGCGGATCCGCTTGTCGGGAACGGATAC/3Cy5Sp/;

55 SEQ ID NO: 38 es CGCGCGGATCCGCTTGTCGGGAACGGATAC/3Cy5Sp/; y

SEQ ID NO: 39 es GCCCGTATTCCCGCTTGCGAGTAGGGCAAT/3Cy5Sp/.

Descripción detallada

La presente invención se refiere generalmente a sistemas y métodos para determinar u obtener imágenes de ácidos nucleicos, por ejemplo, dentro de células. En las realizaciones, puede determinarse el transcriptoma de un célula.

5 Determinadas realizaciones se refieren a determinar ácidos nucleicos, tales como ARNm, dentro de células a resoluciones relativamente altas. En las realizaciones, puede aplicarse una pluralidad de sondas de ácido nucleico a una muestra, y determinarse su unión dentro de la muestra, por ejemplo, usando fluorescencia, para determinar ubicaciones de las sondas de ácido nucleico dentro de la muestra. En las realizaciones, las palabras de código

10 pueden estar basadas en la unión de la pluralidad de sondas de ácido nucleico, y en algunos casos, las palabras de código pueden definir un código de corrección de error para reducir o prevenir la identificación errónea de los ácidos nucleicos. En determinados casos, puede identificarse un número relativamente grande de dianas diferentes usando un número relativamente pequeño de marcadores, por ejemplo, usando diversos enfoques combinatorios.

Ahora se comentan dos enfoques de ejemplo. Sin embargo, debe entenderse que estos se presentan a modo de explicación y no de limitación; se comentan otros aspectos y realizaciones con mayor detalle en el presente documento. En un método de ejemplo, se usan sondas primarias (también denominadas sondas de codificación) y sondas secundarias (también denominadas sondas de lectura), donde las sondas primarias codifican para "palabras de código" y se unen a ácidos nucleicos diana en la muestra, y las sondas secundarias se usan para leer las palabras de código a partir de las sondas primarias. En otro método de ejemplo, se dividen una pluralidad de sondas

20 primarias diferentes que contienen palabras de código en tantas agrupaciones independientes como posiciones haya en las palabras de código, de manera que cada agrupación de sondas primarias corresponde a un valor determinado en una posición determinada de las palabras de código (por ejemplo, un "uno" en la primera posición como en "1001").

Ahora se describe el primer ejemplo con respecto a la figura 3A. Tal como se comentará con más detalle a continuación, en otras realizaciones, también pueden usarse otras configuraciones. En este primer ejemplo, se usa una serie de sondas de ácido nucleico para determinar ácidos nucleicos dentro de una célula u otra muestra, por ejemplo, de manera cualitativa o cuantitativa. Por ejemplo, pueden identificarse que están presentes o ausentes ácidos nucleicos, y/o pueden determinarse los números o las concentraciones de determinados ácidos nucleicos

30 dentro de la célula u otra muestra. En algunos casos, pueden determinarse las posiciones de las sondas dentro de la célula u otra muestra a resoluciones relativamente altas, y en algunos casos, a resoluciones mejores que la longitud de onda de la luz visible.

Este ejemplo se refiere generalmente a detectar espacialmente ácidos nucleicos dentro de una célula u otra muestra, por ejemplo, a resoluciones relativamente altas. Por ejemplo, los ácidos nucleicos pueden ser ARNm, u otros ácidos nucleicos descritos en el presente documento. En un conjunto de realizaciones, pueden determinarse los ácidos nucleicos dentro de la célula administrando o aplicando sondas de ácido nucleico a la célula. En algunos casos, usando enfoques combinatorios, puede determinarse un número relativamente grande de ácidos nucleicos usando un número relativamente pequeño de marcadores diferentes en las sondas de ácido nucleico. Por tanto, por

40 ejemplo, puede usarse un número relativamente pequeño de experimentos para determinar un número relativamente grande de ácidos nucleicos en una muestra, por ejemplo, debido a la unión simultánea de las sondas de ácido nucleico a diferentes ácidos nucleicos en la muestra.

En un conjunto de realizaciones, se aplica una población de sondas de ácido nucleico primarias a la célula (u otra muestra) que es capaz de unirse a los ácidos nucleicos que se sospecha que están presentes dentro de la célula. Después de eso, secuencialmente, se añaden sondas de ácido nucleico secundarias que pueden unirse a, o interactuar de otro modo con, parte de los ácidos nucleicos primarios y se determinan, por ejemplo, usando técnicas de obtención de imágenes tales como microscopía de fluorescencia (por ejemplo, microscopía de fluorescencia convencional), STORM (microscopía de reconstrucción óptica estocástica) u otras técnicas de obtención de

50 imágenes. Después de la obtención de imágenes, se inactivan o eliminan las sondas de ácido nucleico secundarias, y se añade una sonda de ácido nucleico secundaria diferente a la muestra. Esto se repite múltiples veces con múltiples sondas de ácido nucleico secundarias diferentes. Puede usarse el patrón de unión de las diversas sondas de ácido nucleico secundarias para determinar las sondas de ácido nucleico primarias en ubicaciones dentro de la célula u otra muestra, que pueden usarse para determinar ARNm u otros ácidos nucleicos que están presentes.

Por ejemplo, tal como se muestra en la figura 3A, se expone una población de ácidos 10 nucleicos dentro de una célula (representados en este caso por los ácidos 11, 12, y 13 nucleicos) a una población de sondas 20 de ácido nucleico primarias, incluyendo las sondas 21 y 22. Las sondas de ácido nucleico primarias contienen, por ejemplo, una secuencia diana que puede reconocer un ácido nucleico (por ejemplo, una secuencia dentro del ácido 11 nucleico). Las sondas 21 y 22 pueden contener secuencias de direccionamiento iguales o diferentes, que pueden unirse a, o hibridarse con, ácidos nucleicos iguales o diferentes. Como un ejemplo, tal como se muestra en la figura 3A, la sonda 21 contiene una primera secuencia 25 de direccionamiento que dirige la sonda al ácido 11 nucleico, mientras que la sonda 22 contiene una segunda secuencia 26 de direccionamiento, no idéntica a la primera secuencia 25 de direccionamiento y que dirige la sonda al ácido 12 nucleico. La secuencia diana puede ser sustancialmente complementaria a al menos una porción de un ácido nucleico diana, y puede estar presente lo

65 suficiente de la secuencia diana de manera que pueda producirse la unión específica de la sonda de ácido nucleico

al ácido nucleico diana.

Las sondas 20 de ácido nucleico primarias también contienen una o más secuencias de "lectura". En este ejemplo se usan dos de tales secuencias de lectura, aunque en otras realizaciones puede haber una, tres, cuatro, o más secuencias de lectura presentes dentro de una sonda de ácido nucleico primaria. Las secuencias de lectura pueden ser todas ellas independientemente iguales o diferentes. Además, en un conjunto de realizaciones, sondas de ácido nucleico diferentes pueden usar una o más secuencias de lectura comunes. Por ejemplo, puede combinarse combinatoriamente más de una secuencia de lectura en sondas de ácido nucleico diferentes, produciendo de ese modo un número relativamente grande de sondas de ácido nucleico diferentes que pueden identificarse por separado, aunque sólo se usa un número relativamente pequeño de secuencias de lectura. Por tanto, por ejemplo, en la figura 3A, la sonda 21 contiene las secuencias 27 y 29 de lectura, mientras que la sonda 22 contiene las secuencias 27 y 28 de lectura, donde las dos secuencias 27 de lectura son idénticas y diferentes de las secuencias 28 y 29 de lectura.

Después de haber introducido las sondas 20 de ácido nucleico primarias en la muestra y permitido que interaccionen con los ácidos 11, 12, y 13 nucleicos, se aplican una o más sondas 30 de ácido nucleico secundarias a la muestra para determinar las sondas de ácido nucleico primarias. Las sondas de ácido nucleico secundarias pueden contener una secuencia de reconocimiento capaz de reconocer una de las secuencias de lectura presentes dentro de la población de sondas de ácido nucleico primarias. Por ejemplo, la secuencia de reconocimiento puede ser sustancialmente complementaria a al menos una porción de la secuencia de lectura, de manera que la sonda de ácido nucleico secundaria es capaz de unirse a, o hibridarse con, la sonda de ácido nucleico primaria correspondiente. Por ejemplo, en este ejemplo, la secuencia 35 de reconocimiento es capaz de reconocer la secuencia 27 de lectura. Además, las sondas de ácido nucleico secundarias pueden contener una o más entidades 33 de señalización. Por ejemplo, una entidad de señalización puede ser una entidad fluorescente unida a la sonda, o una determinada secuencia de ácidos nucleicos que puede determinarse de cierta manera. Puede usarse más de una secuencia secundaria, por ejemplo, secuencialmente. Por ejemplo, tal como se muestra en esta figura, puede eliminarse la sonda 30 secundaria inicial (por ejemplo, tal como se comenta a continuación) y puede añadirse una sonda 31 secundaria nueva que contiene la secuencia 36 de reconocimiento capaz de reconocer la secuencia 28 de lectura y una o más entidades 33 de señalización. Esto también puede repetirse múltiples veces, por ejemplo, para determinar la secuencia 29 de lectura u otras secuencias de lectura que pueden estar presentes.

La ubicación de las sondas 30, 31, etc., de ácido nucleico secundarias se determina determinando la entidad 33 de señalización. Por ejemplo, si la entidad de señalización es fluorescente, entonces puede usarse microscopía de fluorescencia para determinar la entidad de señalización. En algunas realizaciones, puede usarse obtención de imágenes de una muestra para determinar la entidad de señalización a resoluciones relativamente altas, y en algunos casos, pueden usarse técnicas de obtención de imágenes de superresolución (por ejemplo, resoluciones mejores que la longitud de onda de la luz visible o el límite de difracción de la luz). Los ejemplos de técnicas de obtención de imágenes de superresolución incluyen STORM, u otras técnicas tal como se comentan en el presente documento. En algunos casos, por ejemplo, con determinadas técnicas de obtención de imágenes de superresolución tales como STORM, puede adquirirse más de una imagen de la muestra.

Puede aplicarse más de un tipo de sonda de ácido nucleico secundaria a una célula u otra muestra. Por ejemplo, puede aplicarse una primera sonda de ácido nucleico secundaria que puede reconocer una primera secuencia de lectura, luego puede inactivarse o eliminarse la misma o su entidad de señalización unida, y puede aplicarse una segunda sonda de ácido nucleico secundaria que puede reconocer una segunda secuencia de lectura. Este procedimiento puede repetirse múltiples veces, cada una con una sonda de ácido nucleico secundaria diferente, por ejemplo, para determinar las secuencias de lectura que estaban presentes en las diversas sondas de ácido nucleico primarias. Por tanto, pueden determinarse ácidos nucleicos primarios dentro de la muestra basándose en el patrón de unión de las sondas de ácido nucleico secundarias.

Por ejemplo, una primera ubicación dentro de la célula u otra muestra puede mostrar la unión de una primera sonda secundaria y una tercera sonda secundaria, pero no la unión de una segunda o una cuarta sonda secundaria, mientras que una segunda ubicación puede mostrar un patrón de unión diferente de diversas sondas secundarias. La sonda de ácido nucleico primaria a la que son capaces de unirse o hibridarse las sondas secundarias puede determinarse considerando el patrón de unión de diversas sondas secundarias. Por ejemplo, haciendo referencia a la figura 3A, si una primera sonda secundaria es capaz de determinar la secuencia 27 de lectura, una segunda sonda secundaria es capaz de determinar la secuencia 28 de lectura, y una tercera sonda secundaria es capaz de determinar la secuencia 29 de lectura, entonces puede determinarse el ácido 25 nucleico primario a través de la unión de la primera y tercera sondas secundarias (pero no la segunda sonda secundaria), mientras que puede determinarse el ácido 26 nucleico primario a través de la unión de la primera y segunda sondas secundarias (pero no la tercera sonda secundaria). De manera similar, si se sabe que la primera sonda 21 contiene la secuencia 25 diana mientras que la segunda sonda 22 contiene la secuencia 26 diana, entonces también pueden determinarse los ácidos 11 y 12 nucleicos dentro de la muestra, por ejemplo, espacialmente, basándose en el patrón de unión de las diversas sondas de ácido nucleico secundarias. Además, debe observarse que debido a la presencia de más de una secuencia de lectura en las sondas de ácido nucleico primarias, aunque la primera sonda 21 y la segunda sonda 22 contienen una secuencia de lectura común (secuencia 27 de lectura), estas sondas pueden distinguirse en la

muestra debido a los patrones de unión diferentes de las diversas sondas de ácido nucleico secundarias.

En unas realizaciones, este patrón de unión o hibridación de las sondas de ácido nucleico secundarias se convierte en una "palabra de código". En este ejemplo, por ejemplo, las palabras de código son "101" y "110" para la primera sonda 21 y la segunda sonda 22, respectivamente, donde un valor de 1 representa unión y un valor de 0 representa no unión. Las palabras de código también pueden tener longitudes más largas en otras realizaciones; sólo se muestran tres sondas en este caso por motivos de claridad únicamente. Una palabra de código puede estar directamente relacionada con una secuencia de ácidos nucleicos diana específica de la sonda de ácido nucleico primaria. Por consiguiente, las diferentes sondas de ácido nucleico primarias pueden coincidir con determinadas palabras de código, que entonces pueden usarse para identificar las diferentes dianas de las sondas de ácido nucleico primarias basándose en los patrones de unión de las sondas secundarias, incluso si en algunos casos hay un solapamiento en las secuencias de lectura de las diferentes sondas secundarias, por ejemplo, tal como se muestra en la figura 3A. Sin embargo, si no es evidente la unión (por ejemplo, para el ácido 13 nucleico), entonces la palabra de código sería "000" en este ejemplo.

Los valores en cada palabra de código también pueden asignarse de diferentes maneras en algunas realizaciones. Por ejemplo, un valor de 0 puede representar unión mientras que un valor de 1 representa no unión. De manera similar, un valor de 1 puede representar unión de una sonda de ácido nucleico secundaria con un tipo de entidad de señalización mientras que un valor de 0 puede representar unión de una sonda de ácido nucleico secundaria con otro tipo de entidad de señalización distinguible. Estas entidades de señalización pueden distinguirse, por ejemplo, a través de diferentes colores de fluorescencia. En algunos casos, no es necesario que los valores en las palabras de código estén limitados a 0 y 1. Los valores también pueden extraerse de alfabetos más grandes, tales como sistemas ternarios (por ejemplo, 0, 1, y 2) o cuaternarios (por ejemplo, 0, 1, 2, y 3). Por ejemplo, cada valor diferente puede estar representado por una entidad de señalización distinguible diferente, incluyendo (en algunos casos) un valor que puede estar representado por la ausencia de señal.

Las palabras de código para cada diana pueden asignarse secuencialmente, o pueden asignarse al azar. Por ejemplo, haciendo referencia a la figura 3A, una primera diana de ácido nucleico puede asignarse a 101, mientras que una segunda diana de ácido nucleico puede asignarse a 110. Además, las palabras de código se asignan usando un sistema de detección de error o un sistema de corrección de error, tal como un sistema de Hamming, un código de Golay, o un sistema de Hamming extendido (o un sistema SECDED, es decir, corrección de error único y detección de error doble). En términos generales, tales sistemas pueden usarse para identificar dónde se han producido errores, y en algunos casos, tales sistemas también pueden usarse para corregir los errores y determinar cuál debería haber sido la palabra de código correcta. Por ejemplo, una palabra de código tal como 001 puede detectarse como inválida y corregirse usando un sistema de este tipo a 101, por ejemplo, si 001 no se asigna previamente a una secuencia diana diferente. Pueden usarse una variedad de códigos de corrección de error diferentes, muchos de los cuales se han desarrollado previamente para su uso dentro de la industria informática; sin embargo, tales sistemas de corrección de error no se han usado normalmente dentro de sistemas biológicos. A continuación se comentan con más detalle ejemplos adicionales de tales códigos de corrección de error.

También debe entenderse que no es necesario usar todas las palabras de código posibles en un código en algunos casos. Por ejemplo, en algunas realizaciones, las palabras de código que no se usan pueden servir como controles negativos. De manera similar, en algunas realizaciones, algunas palabras de código pueden excluirse porque son más propensas a errores en la medición que otras palabras de código. Por ejemplo, en algunas implementaciones, leer una palabra de código con más valores de "1" puede ser más propenso a errores que leer una palabra de código con menos valores de "1".

Debe entenderse que la descripción anterior es un ejemplo de una realización de la invención.

Por ejemplo, haciendo referencia ahora a la figura 3B, los ácidos 11, 12, y 13 nucleicos se exponen a diferentes rondas de sondas 21, 22, 23, 24, etc., en este ejemplo. Cada una de estas sondas puede contener una secuencia diana que puede reconocer un ácido nucleico (por ejemplo, una secuencia dentro del ácido 11 ó 12 nucleico). Cada una de estas sondas puede dirigirse al mismo ácido nucleico, pero a regiones diferentes del ácido nucleico. Además, parte o la totalidad de las sondas pueden contener una o más entidades de señalización, por ejemplo, entidad 29 de señalización en la sonda 21. Por ejemplo, la entidad de señalización puede ser una entidad fluorescente unida a la sonda, o una determinada secuencia de ácidos nucleicos que puede determinarse de cierta manera.

Puede aplicarse la primera ronda de sondas (por ejemplo, sonda 21 y sonda 22) a la célula u otra muestra. Puede permitirse que la sonda 21 se una al ácido 11 nucleico a través de la secuencia 25 diana. Tal unión puede determinarse determinando la entidad 29 de señalización. Por ejemplo, si la entidad de señalización es fluorescente, entonces puede usarse microscopía de fluorescencia para determinar la entidad de señalización, por ejemplo, espacialmente dentro de la célula u otra muestra. En parte pero no la totalidad de las realizaciones, puede usarse obtención de imágenes de una muestra para determinar la entidad de señalización a resoluciones relativamente altas, y en algunos casos, pueden usarse técnicas de obtención de imágenes de superresolución. También pueden estar presentes otras sondas diferentes; por ejemplo, la sonda 22 que contiene la secuencia 26 diana puede unirse al ácido 12 nucleico, y determinarse a través de la entidad 29 de señalización dentro de la sonda 22. Esto puede

producirse, por ejemplo, secuencialmente o simultáneamente. Opcionalmente, también pueden eliminarse o inactivarse las sondas 21 y 22, por ejemplo, entre la aplicación de diferentes rondas de sondas.

A continuación, se aplica una segunda ronda de sondas (por ejemplo, sonda 23) a la muestra. En este ejemplo, la sonda 23 es capaz de unirse al ácido 11 nucleico a través de una región de direccionamiento, aunque no hay ninguna sonda en la segunda ronda que es capaz de unirse al ácido 12 nucleico. Se permite que se produzca la unión de las sondas tal como se comentó anteriormente, y puede producirse la determinación de unión a través de las entidades de señalización. Estas entidades de señalización pueden ser iguales o diferentes a las de la primera ronda de sondas. Este procedimiento puede repetirse cualquier número de veces con sondas diferentes. Por ejemplo, tal como se muestra en la figura 3B, la ronda 2 contiene sondas capaces de unirse al ácido 11 nucleico, mientras que la ronda 3 contiene sondas capaces de unirse al ácido 12 nucleico.

En las realizaciones, cada ronda de unión o hibridación de sondas de ácido nucleico puede convertirse en una "palabra de código". En este ejemplo, usando las sondas 21, 22, 23, y 24, pueden formarse las palabras de código 101 o 110, donde 1 representa unión y 0 representa no unión y la primera posición corresponde a la unión de las sondas 21 ó 22 mientras que la segunda posición corresponde a la unión de las sondas 22, y la tercera posición corresponde a la unión de la sonda 24. Una palabra de código de 000 representaría no unión, por ejemplo, tal como se muestra con el ácido 13 nucleico en este ejemplo. Una palabra de código puede estar directamente relacionada con una secuencia de ácidos nucleicos diana específica de las sondas de ácido nucleico, mediante el diseño de sondas de ácido nucleico apropiadas. Por tanto, por ejemplo, 110 puede corresponder a un primer ácido 12 nucleico diana (por ejemplo, la primera y segunda rondas de sondas de ácido nucleico que contienen sondas capaces de dirigirse al ácido 11 nucleico, y estas sondas pueden dirigirse a regiones iguales o diferentes del ácido 11 nucleico) mientras que 101 puede corresponder a un segundo ácido nucleico diana (por ejemplo, la primera y tercera rondas de sondas de ácido nucleico que contienen sondas capaces de dirigirse al ácido 12 nucleico, y estas sondas pueden dirigirse a regiones iguales o diferentes del ácido 12 nucleico). Además, debe observarse que cada ronda de sondas puede contener entidades de señalización iguales o diferentes a otras sondas en la misma ronda, y/u otras sondas en rondas diferentes. Por ejemplo, en un conjunto de realizaciones, sólo se usa una entidad de señalización en la totalidad de las rondas de sondas.

De manera similar a lo anterior, las palabras de código para cada diana pueden asignarse secuencialmente, o pueden asignarse al azar. Las palabras de código pueden asignarse dentro de un espacio de código en algunas realizaciones usando un sistema de detección de error o corrección de error, tal como un sistema de Hamming, un código de Golay, o un sistema de Hamming extendido o un sistema SECDED (corrección de error único y detección de error doble). En términos generales, tales sistemas de corrección de error pueden usarse para identificar dónde se han producido errores, y en algunos casos, tales sistemas también pueden usarse para corregir los errores y determinar cuál debería haber sido la palabra de código correcta.

De manera similar a lo anterior, los valores en cada posición en la palabra de código pueden asignarse arbitrariamente en determinadas realizaciones a unión o no unión de sondas que contienen más de una entidad de señalización distinguible.

En algunos casos, las sondas de ácido nucleico pueden formarse en "agrupaciones" o grupos de ácidos nucleicos que comparten una característica común. Por ejemplo, las sondas para todas las dianas con palabras de código que contienen un 1 en la primera posición, por ejemplo 110 y 101 pero no 011, pueden comprender una agrupación mientras que las sondas para todas las dianas que contienen un 1 en la segunda posición, por ejemplo 110 y 011 pero no 101, pueden comprender otra agrupación. Véase también la figura 1C. En algunos casos, una sonda de ácido nucleico puede ser un miembro de más de un grupo o agrupación. Los miembros de una agrupación de ácidos nucleicos también pueden contener características además de las secuencias diana, secuencias de lectura, y/o entidades de señalización que les permiten distinguirse de otros grupos. Estas características pueden ser secuencias de ácido nucleico cortas que se usan para la amplificación, producción, o separación de estas secuencias. Las sondas de ácido nucleico de cada grupo pueden aplicarse a una muestra, por ejemplo, secuencialmente, tal como se comenta en el presente documento.

Por tanto, en algunos aspectos, la presente invención se refiere generalmente a sistemas y métodos para determinar ácidos nucleicos dentro de una célula u otra muestra. La muestra puede incluir un cultivo de células, una suspensión de células, un tejido biológico, una biopsia, un organismo, o similares. La muestra también puede estar libre de células pero, no obstante, contener ácidos nucleicos. Si la muestra contiene una célula, la célula puede ser una célula humana, o cualquier otra célula adecuada, por ejemplo, una célula de mamífero, una célula de pez, una célula de insecto, una célula de planta, o similares. Puede estar presente más de una célula en algunos casos.

Los ácidos nucleicos que van a determinarse pueden ser, por ejemplo, ADN, ARN, u otros ácidos nucleicos que están presentes dentro de una célula (u otra muestra). Los ácidos nucleicos pueden ser endógenos con respecto a la célula, o añadirse a la célula. Por ejemplo, el ácido nucleico puede ser viral, o crearse artificialmente. En algunos casos, el ácido nucleico que va a determinarse puede expresarse por la célula. El ácido nucleico es ARN en algunas realizaciones. El ARN puede ser ARN codificante y/o no codificante. Los ejemplos no limitativos de ARN que puede estudiarse dentro de la célula incluyen ARNm, ARNip, ARNr, miARN, ARNt, ARNInc, ARNnop, ARNnp, ARNex,

piARN, o similares.

En algunos casos, puede estudiarse una porción significativa del ácido nucleico dentro de la célula. Por ejemplo, en algunos casos, puede determinarse lo suficiente del ARN presente dentro de una célula para producir un transcriptoma parcial o completo de la célula. En algunos casos, se determinan al menos 4 tipos de ARNm dentro de una célula, y en algunos casos, pueden determinarse al menos 3, al menos 4, al menos 7, al menos 8, al menos 12, al menos 14, al menos 15, al menos 16, al menos 22, al menos 30, al menos 31, al menos 32, al menos 50, al menos 63, al menos 64, al menos 72, al menos 75, al menos 100, al menos 127, al menos 128, al menos 140, al menos 255, al menos 256, al menos 500, al menos 1.000, al menos 1.500, al menos 2.000, al menos 2.500, al menos 3.000, al menos 4.000, al menos 5.000, al menos 7.500, al menos 10.000, al menos 12.000, al menos 15.000, al menos 20.000, al menos 25.000, al menos 30.000, al menos 40.000, al menos 50.000, al menos 75.000, o al menos 100.000 tipos de ARNm dentro de una célula.

En algunos casos, puede determinarse el transcriptoma de una célula. Debe entenderse que el transcriptoma abarca generalmente todas las moléculas de ARN producidas dentro de una célula, no sólo el ARNm. Por tanto, por ejemplo, el transcriptoma también puede incluir ARNr, ARNt, ARNip, etc. En algunas realizaciones, puede determinarse al menos el 5 %, al menos el 10 %, al menos el 15 %, al menos el 20 %, al menos el 25 %, al menos el 30 %, al menos el 40 %, al menos el 50 %, al menos el 60 %, al menos el 70 %, al menos el 80 %, al menos el 90 %, o el 100 % del transcriptoma de una célula.

La determinación de uno o más ácidos nucleicos dentro de la célula u otra muestra puede ser cualitativa y/o cuantitativa. Además, la determinación también puede ser espacial, por ejemplo, puede determinarse la posición del ácido nucleico dentro de la célula u otra muestra en dos o tres dimensiones. En algunas realizaciones, pueden determinarse las posiciones, el número, y/o las concentraciones de los ácidos nucleicos dentro de la célula (u otra muestra).

En algunos casos, puede determinarse una porción significativa del genoma de una célula. Los segmentos genómicos determinados pueden ser continuos o estar intercalados en el genoma. Por ejemplo, en algunos casos, se determinan al menos 4 segmentos genómicos dentro de una célula, y en algunos casos, pueden determinarse al menos 3, al menos 4, al menos 7, al menos 8, al menos 12, al menos 14, al menos 15, al menos 16, al menos 22, al menos 30, al menos 31, al menos 32, al menos 50, al menos 63, al menos 64, al menos 72, al menos 75, al menos 100, al menos 127, al menos 128, al menos 140, al menos 255, al menos 256, al menos 500, al menos 1.000, al menos 1.500, al menos 2.000, al menos 2.500, al menos 3.000, al menos 4.000, al menos 5.000, al menos 7.500, al menos 10.000, al menos 12.000, al menos 15.000, al menos 20.000, al menos 25.000, al menos 30.000, al menos 40.000, al menos 50.000, al menos 75.000, o al menos 100.000 segmentos genómicos dentro de una célula.

En algunos casos, puede determinarse el genoma completo de una célula. Debe entenderse que el genoma abarca generalmente todas las moléculas de ADN producidas dentro de una célula, no sólo el ADN cromosómico. Por tanto, por ejemplo, el genoma también puede incluir, en algunos casos, ADN mitocondrial, ADN cloroplástico, ADN plasmídico, etc. En algunas realizaciones, puede determinarse al menos aproximadamente el 5 %, al menos aproximadamente el 10 %, al menos aproximadamente el 15 %, al menos aproximadamente el 20 %, al menos aproximadamente el 25 %, al menos aproximadamente el 30 %, al menos aproximadamente el 40 %, al menos aproximadamente el 50 %, al menos aproximadamente el 60 %, al menos aproximadamente el 70 %, al menos aproximadamente el 80 %, al menos aproximadamente el 90 %, o el 100 % del genoma de una célula.

Tal como se comenta en el presente documento, pueden usarse una variedad de sondas de ácido nucleico para determinar uno o más ácidos nucleicos dentro de una célula u otra muestra. Las sondas pueden comprender ácidos nucleicos (o entidades que pueden hibridarse con un ácido nucleico, por ejemplo, específicamente) tales como ADN, ARN, ANB (ácidos nucleicos bloqueados), ANP (ácidos nucleicos peptídicos), o combinaciones de los mismos. En algunos casos, también pueden estar presentes componentes adicionales dentro de las sondas de ácido nucleico, por ejemplo, tal como se comenta a continuación. Puede usarse cualquier método adecuado para introducir sondas de ácido nucleico en una célula.

Por ejemplo, en algunas realizaciones, la célula se fija antes de introducir las sondas de ácido nucleico, por ejemplo, para conservar las posiciones de los ácidos nucleicos dentro de la célula. Los expertos habituales en la técnica conocen técnicas para fijar las células. Como ejemplos no limitativos, una célula puede fijarse usando sustancias químicas tales como formaldehído, paraformaldehído, glutaraldehído, etanol, metanol, acetona, ácido acético, o similares. En una realización, una célula puede fijarse usando disolvente orgánico mediado por tampón Hepes-ácido glutámico (HOPE).

Las sondas de ácido nucleico pueden introducirse en la célula (u otra muestra) usando cualquier método adecuado. En algunos casos, la célula puede permeabilizarse suficientemente de manera que las sondas de ácido nucleico pueden introducirse en la célula haciendo fluir un fluido que contiene las sondas de ácido nucleico alrededor de las células. En algunos casos, las células pueden permeabilizarse suficientemente como parte de un procedimiento de fijación; en otras realizaciones, las células pueden permeabilizarse mediante exposición a determinadas sustancias químicas tales como etanol, metanol, Triton, o similares. Además, en algunas realizaciones, pueden usarse técnicas

tales como electroporación o microinyección para introducir sondas de ácido nucleico en una célula u otra muestra.

Determinados aspectos de la presente invención se refieren generalmente a sondas de ácido nucleico que se introducen en una célula (u otra muestra). Las sondas pueden comprender cualquiera de una variedad de entidades que pueden hibridarse con un ácido nucleico, normalmente mediante apareamiento de bases de Watson-Crick, tales como ADN, ARN, ANB, ANP, etc., dependiendo de la aplicación. La sonda de ácido nucleico contiene normalmente una secuencia diana que es capaz de unirse a al menos una porción de un ácido nucleico diana, en algunos casos específicamente. Cuando se introduce en una célula u otro sistema, el sistema diana puede ser capaz de unirse a un ácido nucleico diana específico (por ejemplo, un ARNm, u otros ácidos nucleicos tal como se comenta en el presente documento). En algunos casos, las sondas de ácido nucleico pueden determinarse usando entidades de señalización (por ejemplo, tal como se comenta a continuación), y/o usando sondas de ácido nucleico secundarias capaces de unirse a las sondas de ácido nucleico (es decir, a sondas de ácido nucleico primarias). A continuación se comenta con detalle la determinación de tales sondas de ácido nucleico.

En algunos casos, puede aplicarse más de un tipo de sonda de ácido nucleico (primaria) a una muestra, por ejemplo, simultáneamente. Por ejemplo, puede haber al menos 2, al menos 5, al menos 10, al menos 25, al menos 50, al menos 75, al menos 100, al menos 300, al menos 1.000, al menos 3.000, al menos 10.000, o al menos 30.000 sondas de ácido nucleico distinguibles que se aplican a una muestra, por ejemplo, simultáneamente o secuencialmente.

La secuencia diana puede estar posicionada en cualquier lugar dentro de la sonda de ácido nucleico (o sonda de ácido nucleico primaria o sonda de codificación de ácido nucleico). La secuencia diana puede contener una región que es sustancialmente complementaria a una porción de un ácido nucleico diana. En algunos casos, las porciones pueden ser al menos el 50 %, al menos el 60 %, al menos el 70 %, al menos el 75 %, al menos el 80 %, al menos el 85 %, al menos el 90 %, al menos el 92 %, al menos el 94 %, al menos el 95 %, al menos el 96 %, al menos el 97 %, al menos el 98 %, al menos el 99 %, o el 100 % complementarias. En algunos casos, la secuencia diana puede tener una longitud de al menos 5, al menos 10, al menos 15, al menos 20, al menos 25, al menos 30, al menos 35, al menos 40, al menos 50, al menos 60, al menos 65, al menos 75, al menos 100, al menos 125, al menos 150, al menos 175, al menos 200, al menos 250, al menos 300, al menos 350, al menos 400, o al menos 450 nucleótidos. En algunos casos, la secuencia diana puede tener una longitud de no más de 500, no más de 450, no más de 400, no más de 350, no más de 300, no más de 250, no más de 200, no más de 175, no más de 150, no más de 125, no más de 100, no más de 75, no más de 70, no más de 65, no más de 60, no más de 55, no más de 50, no más de 45, no más de 40, no más de 35, no más de 30, no más de 20, o no más de 10 nucleótidos. También son posibles combinaciones de cualesquiera de estos, por ejemplo, la secuencia diana puede tener una longitud de entre 10 y 30 nucleótidos, entre 20 y 40 nucleótidos, entre 5 y 50 nucleótidos, entre 10 y 200 nucleótidos, o entre 25 y 35 nucleótidos, entre 10 y 300 nucleótidos, etc. Normalmente, la complementariedad se determina basándose en el apareamiento de bases nucleotídicas de Watson-Crick.

La secuencia diana de una sonda de ácido nucleico (primaria) puede determinarse con referencia a un ácido nucleico diana que se sospecha que está presente dentro de una célula u otra muestra. Por ejemplo, un ácido nucleico diana para una proteína puede determinarse usando la secuencia de la proteína, determinando los ácidos nucleicos que se expresan para formar la proteína. En algunos casos, se usa sólo una porción de los ácidos nucleicos que codifican para la proteína, por ejemplo, que tiene las longitudes tal como se comentaron anteriormente. Además, en algunos casos, puede usarse más de una secuencia diana que puede usarse para identificar una diana particular. Por ejemplo, pueden usarse, secuencialmente y/o simultáneamente, múltiples sondas que pueden unirse a, o hibridarse con, regiones diferentes de la misma diana. Hibridación se refiere normalmente a un procedimiento de emparejamiento mediante el cual ácidos nucleicos monocatenarios complementarios se asocian a través de apareamiento de bases nucleotídicas de Watson-Crick (por ejemplo, enlaces de hidrógeno, guanina-citosina y adenina-timina) para formar un ácido nucleico bicatenario.

En las realizaciones, una sonda de ácido nucleico, tal como una sonda de ácido nucleico primaria, también puede comprender una o más secuencias de "lectura". Sin embargo, debe entenderse que las secuencias de lectura no son necesarias en todos los casos. En algunas realizaciones, la sonda de ácido nucleico puede comprender 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16 o más, 20 o más, 32 o más, 40 o más, 50 o más, 64 o más, 75 o más, 100 o más, 128 o más secuencias de lectura. Las secuencias de lectura pueden estar posicionadas en cualquier lugar dentro de la sonda de ácido nucleico. Si está presente más de una secuencia de lectura, las secuencias de lectura pueden estar posicionadas unas junto a otras, y/o intercaladas con otras secuencias.

Las secuencias de lectura, si están presentes, pueden tener cualquier longitud. Si se usa más de una secuencia de lectura, las secuencias de lectura pueden tener independientemente longitudes iguales o diferentes. Por ejemplo, la secuencia de lectura puede tener una longitud de al menos 5, al menos 10, al menos 15, al menos 20, al menos 25, al menos 30, al menos 35, al menos 40, al menos 50, al menos 60, al menos 65, al menos 75, al menos 100, al menos 125, al menos 150, al menos 175, al menos 200, al menos 250, al menos 300, al menos 350, al menos 400, o al menos 450 nucleótidos. En algunos casos, la secuencia de lectura puede tener una longitud de no más de 500, no más de 450, no más de 400, no más de 350, no más de 300, no más de 250, no más de 200, no más de 175, no más de 150, no más de 125, no más de 100, no más de 75, no más de 70, no más de 65, no más de 60, no más de

55, no más de 50, no más de 45, no más de 40, no más de 35, no más de 30, no más de 20, o no más de 10 nucleótidos. También son posibles combinaciones de cualesquiera de estos, por ejemplo, la secuencia de lectura puede tener una longitud de entre 10 y 30 nucleótidos, entre 20 y 40 nucleótidos, entre 5 y 50 nucleótidos, entre 10 y 200 nucleótidos, o entre 25 y 35 nucleótidos, entre 10 y 300 nucleótidos, etc.

La secuencia de lectura puede ser arbitraria o aleatoria en algunas realizaciones. En determinados casos, las secuencias de lectura se eligen para reducir o minimizar la homología con otros componentes de la célula u otra muestra, por ejemplo, de manera que las secuencias de lectura por sí mismas no se unan a, o se hibriden con, otros ácidos nucleicos que se sospecha que están dentro de la célula u otra muestra. En algunos casos, la homología puede ser menor del 10 %, menor del 8 %, menor del 7 %, menor del 6 %, menor del 5 %, menor del 4 %, menor del 3 %, menor del 2 %, o menor del 1 %. En algunos casos, puede haber una homología de menos de 20 pares de bases, menos de 18 pares de bases, menos de 15 pares de bases, menos de 14 pares de bases, menos de 13 pares de bases, menos de 12 pares de bases, menos de 11 pares de bases, o menos de 10 pares de bases. En algunos casos, los pares de bases son secuenciales.

En un conjunto de realizaciones, una población de sondas de ácido nucleico puede contener un determinado número de secuencias de lectura, que puede ser menor que el número de dianas de las sondas de ácido nucleico en algunos casos. Los expertos habituales en la técnica serán conscientes de que si hay una entidad de señalización y n secuencias de lectura, entonces en general pueden identificarse de manera singular $2^n - 1$ dianas de ácido nucleico diferentes. Sin embargo, no es necesario usar todas las combinaciones posibles. Por ejemplo, una población de sondas de ácido nucleico puede dirigirse a 12 secuencias de ácido nucleico diferentes, y aun así contener no más de 8 secuencias de lectura. Como otro ejemplo, una población de ácidos nucleicos puede dirigirse a 140 especies de ácido nucleico diferentes, y aun así contener no más de 16 secuencias de lectura. Pueden identificarse por separado diferentes dianas de secuencia de ácidos nucleicos usando diferentes combinaciones de secuencias de lectura dentro de cada sonda. Por ejemplo, cada sonda puede contener 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, etc., o más secuencias de lectura. En algunos casos, cada una de una población de sondas de ácido nucleico puede contener el mismo número de secuencias de lectura, aunque en otros casos puede haber números diferentes de secuencias de lectura presentes en las diversas sondas.

Como un ejemplo no limitativo, una primera sonda de ácido nucleico puede contener una primera secuencia diana, una primera secuencia de lectura, y una segunda secuencia de lectura, mientras que una segunda sonda de ácido nucleico diferente puede contener una segunda secuencia diana, la misma primera secuencia de lectura, pero una tercera secuencia de lectura en lugar de la segunda secuencia de lectura. De ese modo, pueden distinguirse tales sondas determinando las diversas secuencias de lectura presentes o asociadas con una sonda o ubicación dada, tal como se comenta en el presente documento.

Además, las sondas de ácido nucleico (y sus sitios complementarios correspondientes en las sondas de codificación), en determinadas realizaciones, pueden prepararse usando sólo 2 o sólo 3 de las 4 bases, tal como excluyendo la totalidad de las "G" o excluyendo la totalidad de las "C" dentro de la sonda. Las secuencias que carecen o bien de las "G" o bien de las "C" pueden formar muy poca estructura secundaria en determinadas realizaciones, y pueden contribuir a una hibridación más uniforme y más rápida.

En las realizaciones, la sonda de ácido nucleico puede contener una entidad de señalización. Sin embargo, debe entenderse que las entidades de señalización no son requeridas en todos los casos; por ejemplo, la sonda de ácido nucleico se determina usando sondas de ácido nucleico secundarias en las realizaciones, tal como se comenta con detalle adicional a continuación. A continuación también se comentan con más detalle ejemplos de entidades de señalización que pueden usarse.

También pueden estar presentes otros componentes dentro de una sonda de ácido nucleico. Por ejemplo, en un conjunto de realizaciones, pueden estar presentes una o más secuencias de cebador, por ejemplo, para permitir la amplificación enzimática de las sondas. Los expertos habituales en la técnica serán conscientes de secuencias de cebador adecuadas para aplicaciones tales como amplificación (por ejemplo, usando PCR u otras técnicas adecuadas). Muchas de tales secuencias de cebador están disponibles comercialmente. Otros ejemplos de secuencias que pueden estar presentes dentro de una sonda de ácido nucleico primaria incluyen, pero no se limitan a, secuencias de promotor, operones, secuencias de identificación, secuencias de terminación, o similares.

Normalmente, un cebador es un ácido nucleico monocatenario o parcialmente bicatenario (por ejemplo, ADN) que sirve como punto de partida para la síntesis de ácidos nucleicos, permitiendo que las enzimas polimerasa tales como polimerasa de ácidos nucleicos extiendan el cebador y repliquen la cadena complementaria. Un cebador es (por ejemplo, está diseñado para ser) complementario a, y para hibridarse con, un ácido nucleico diana. En algunas realizaciones, un cebador es un cebador sintético. En algunas realizaciones, un cebador es un cebador que no se produce de manera natural. Un cebador tiene normalmente una longitud de 10 a 50 nucleótidos. Por ejemplo, un cebador puede tener una longitud de 10 a 40, de 10 a 30, de 10 a 20, de 25 a 50, de 15 a 40, de 15 a 30, de 20 a 50, de 20 a 40, o de 20 a 30 nucleótidos. En algunas realizaciones, un cebador tiene una longitud de 18 a 24 nucleótidos.

Además, los componentes de la sonda de ácido nucleico pueden estar dispuestos en cualquier orden adecuado. Por ejemplo, en una realización, los componentes pueden estar dispuestos en una sonda de ácido nucleico como: cebador—secuencias de lectura—secuencia de direccionamiento—secuencias de lectura—cebador inverso. Las “secuencias de lectura” en esta estructura pueden contener, cada una, cualquier número (incluyendo 0) de secuencias de lectura, siempre que al menos una secuencia de lectura esté presente en la sonda. Las estructuras de ejemplo no limitativas incluyen cebador—secuencia de direccionamiento—secuencias de lectura—cebador inverso, cebador—secuencias de lectura—secuencia de direccionamiento—cebador inverso, secuencia de direccionamiento—cebador—secuencia de direccionamiento—secuencias de lectura—cebador inverso, secuencia de direccionamiento—cebador—secuencias de lectura—secuencia de direccionamiento—cebador inverso, cebador—secuencia diana—secuencias de lectura—secuencia de direccionamiento—cebador inverso, secuencia de direccionamiento—cebador—secuencia de lectura—cebador inverso, secuencia de direccionamiento—secuencia de lectura—cebador, secuencia de lectura—secuencia de direccionamiento—cebador, secuencia de lectura—cebador—secuencia de direccionamiento—cebador inverso, etc. Además, el cebador inverso es opcional en algunas realizaciones, incluyendo en la totalidad de los ejemplos descritos anteriormente.

Después de la introducción de las sondas de ácido nucleico en un célula u otra muestra, las sondas de ácido nucleico pueden determinarse directamente determinando las entidades de señalización (si están presentes), y/o las sondas de ácido nucleico pueden determinarse usando una o más sondas de ácido nucleico secundarias, según determinados aspectos de la invención. Tal como se mencionó, en algunos casos, la determinación puede ser espacial, por ejemplo, en dos o tres dimensiones. Además, en algunos casos, la determinación puede ser cuantitativa, por ejemplo, puede determinarse la cantidad o concentración de una sonda de ácido nucleico primaria (y de un ácido nucleico diana). Adicionalmente, las sondas secundarias pueden comprender cualquiera de una variedad de entidades capaces de hibridarse con un ácido nucleico, por ejemplo, ADN, ARN, ANB, y/o ANP, etc., dependiendo de la aplicación. A continuación se comentan con más detalle las entidades de señalización.

Una sonda de ácido nucleico secundaria puede contener una secuencia de reconocimiento capaz de unirse a, o hibridarse con, una secuencia de lectura de una sonda de ácido nucleico primaria. En algunos casos, la unión es específica, o la unión puede ser de manera que, preferentemente, una secuencia de reconocimiento se une a, o se hibrida con, sólo una de las secuencias de lectura que están presentes. La sonda de ácido nucleico secundaria también puede contener una o más entidades de señalización. Si se usa más de una sonda de ácido nucleico secundaria, las entidades de señalización pueden ser iguales o diferentes.

Las secuencias de reconocimiento pueden tener cualquier longitud, y múltiples secuencias de reconocimiento pueden tener longitudes iguales o diferentes. Si se usa más de una secuencia de reconocimiento, las secuencias de reconocimiento pueden tener independientemente longitudes iguales o diferentes. Por ejemplo, la secuencia de reconocimiento puede tener una longitud de al menos 5, al menos 10, al menos 15, al menos 20, al menos 25, al menos 30, al menos 35, al menos 40, o al menos 50 nucleótidos. En algunos casos, la secuencia de reconocimiento puede tener una longitud de no más de 75, no más de 70, no más de 65, no más de 60, no más de 55, no más de 50, no más de 45, no más de 40, no más de 35, no más de 30, no más de 20, o no más de 10 nucleótidos. También son posibles combinaciones de cualesquiera de estos, por ejemplo, la secuencia de reconocimiento puede tener una longitud de entre 10 y 30, entre 20 y 40, o entre 25 y 35 nucleótidos, etc. En una realización, la secuencia de reconocimiento tiene la misma longitud que la secuencia de lectura. Además, en algunos casos, la secuencia de reconocimiento puede ser al menos el 50 %, al menos el 60 %, al menos el 70 %, al menos el 75 %, al menos el 80 %, al menos el 85 %, al menos el 90 %, al menos el 92 %, al menos el 94 %, al menos el 95 %, al menos el 96 %, al menos el 97 %, al menos el 98 %, al menos el 99 %, o al menos el 100 % complementaria a una secuencia de lectura de la sonda de ácido nucleico primaria.

Tal como se mencionó, en algunos casos, la sonda de ácido nucleico secundaria puede comprender una o más entidades de señalización. A continuación se comentan con más detalle ejemplos de entidades de señalización.

Tal como se comentó, en determinados aspectos de la invención, se usan sondas de ácido nucleico que contienen diversas “secuencias de lectura”. Por ejemplo, una población de sondas de ácido nucleico primarias puede contener determinadas “secuencias de lectura” que pueden unirse a determinadas sondas de ácido nucleico secundarias, y las ubicaciones de las sondas de ácido nucleico primarias se determinan dentro de la muestra usando sondas de ácido nucleico secundarias, por ejemplo, que comprenden una entidad de señalización. Tal como se mencionó, en algunos casos, una población de secuencias de lectura puede combinarse en diversas combinaciones para producir sondas de ácido nucleico diferentes, por ejemplo, de manera que puede usarse un número relativamente pequeño de secuencias de lectura para producir un número relativamente grande de sondas de ácido nucleico diferentes.

Por tanto, en algunos casos, cada una de una población de sondas de ácido nucleico primarias (u otras sondas de ácido nucleico) puede contener un determinado número de secuencias de lectura, algunas de las cuales se comparten entre sondas de ácido nucleico primarias diferentes de manera que la población total de sondas de ácido nucleico primarias puede contener un determinado número de secuencias de lectura. Una población de sondas de ácido nucleico puede tener cualquier número adecuado de secuencias de lectura. Por ejemplo, una población de sondas de ácido nucleico primarias puede tener 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, etc., secuencias de lectura. También son posibles más de 20 en algunas realizaciones. Además, en algunos casos, una

población de sondas de ácido nucleico puede tener, en total, 1 o más, 2 o más, 3 o más, 4 o más, 5 o más, 6 o más, 7 o más, 8 o más, 9 o más, 10 o más, 11 o más, 12 o más, 13 o más, 14 o más, 15 o más, 16 o más, 20 o más, 24 o más, 32 o más, 40 o más, 50 o más, 60 o más, 64 o más, 100 o más, 128 o más, etc., secuencias de lectura posibles presentes, aunque parte o la totalidad de las sondas pueden contener, cada una, más de una secuencia de lectura, tal como se comenta en el presente documento. Además, en algunas realizaciones, la población de sondas de ácido nucleico puede tener no más de 100, no más de 80, no más de 64, no más de 60, no más de 50, no más de 40, no más de 32, no más de 24, no más de 20, no más de 16, no más de 15, no más de 14, no más de 13, no más de 12, no más de 11, no más de 10, no más de 9, no más de 8, no más de 7, no más de 6, no más de 5, no más de 4, no más de 3, o no más de dos secuencias de lectura presentes. También son posibles combinaciones de cualesquiera de estos, por ejemplo, una población de sondas de ácido nucleico puede comprender entre 10 y 15 secuencias de lectura en total.

Como un ejemplo no limitativo de un enfoque para producir combinatoriamente un número relativamente grande de sondas de ácido nucleico a partir de un número relativamente pequeño de secuencias de lectura, en una población de 6 tipos diferentes de sondas de ácido nucleico que comprenden, cada una, una o más secuencias de lectura, el número total de secuencias de lectura dentro de la población puede ser no mayor de 4. Debe entenderse que, aunque se usan 4 secuencias de lectura en este ejemplo para facilitar la explicación, en otras realizaciones, pueden efectuarse mayores números de sondas de ácido nucleico, por ejemplo, usando 5, 8, 10, 16, 32, etc., o más secuencias de lectura, o cualquier otro número adecuado de secuencias de lectura descrito en el presente documento, dependiendo de la aplicación. Haciendo referencia ahora a la figura 4A, si cada una de las sondas de ácido nucleico primarias contiene dos secuencias de lectura diferentes, entonces usando 4 de tales secuencias de lectura (A, B, C, y D), pueden identificarse por separado hasta 6 sondas. Debe observarse que, en este ejemplo, el orden de secuencias de lectura en una sonda de ácido nucleico no es esencial, es decir, "AB" y "BA" pueden tratarse como sinónimos (aunque, en otras realizaciones, el orden de secuencias de lectura puede ser esencial y "AB" y "BA" no tienen por qué ser sinónimos). De manera similar, si se usan 5 secuencias de lectura (A, B, C, D, y E) en la población de sondas de ácido nucleico primarias, pueden identificarse por separado hasta 10 sondas, tal como se muestra en la figura 4B. Por ejemplo, un experto habitual en la técnica entendería que, para k secuencias de lectura

en una población con n secuencias de lectura en cada sonda, pueden producirse hasta $\binom{n}{k}$ sondas diferentes, suponiendo que el orden de secuencias de lectura no es esencial; dado que no es necesario que la totalidad de las sondas tengan el mismo número de secuencias de lectura y no es necesario que se use la totalidad de las combinaciones de secuencias de lectura en cada realización, también puede usarse o bien más o bien menos que este número de sondas diferentes en determinadas realizaciones. Además, también debe entenderse que no es necesario que el número de secuencias de lectura en cada sonda sea idéntico en algunas realizaciones. Por ejemplo, algunas sondas pueden contener 2 secuencias de lectura mientras que otras sondas pueden contener 3 secuencias de lectura.

En algunos aspectos, pueden usarse las secuencias de lectura y/o el patrón de unión de sondas de ácido nucleico dentro de una muestra para definir un código de detección de error y/o corrección de error, por ejemplo, para reducir o prevenir la identificación errónea o errores de los ácidos nucleicos, por ejemplo, tal como se comentó con referencia a la figura 3. Por tanto, por ejemplo, si se indica unión (por ejemplo, tal como se determina usando una entidad de señalización), entonces la ubicación puede identificarse con un "1"; a la inversa, si se indica no unión, entonces la ubicación puede identificarse con un "0" (o viceversa, en algunos casos). Entonces pueden usarse múltiples rondas de determinaciones de unión, por ejemplo, usando sondas de ácido nucleico diferentes, para crear una "palabra de código", por ejemplo, para esa ubicación espacial. En las realizaciones, la palabra de código puede someterse a detección y/o corrección de error. Por ejemplo, las palabras de código pueden organizarse de manera que, si no se encuentra coincidencia para un conjunto dado de secuencias de lectura o patrón de unión de sondas de ácido nucleico, entonces la coincidencia puede identificarse como un error, y opcionalmente, puede aplicarse corrección de error a las secuencias para determinar la diana correcta para las sondas de ácido nucleico. En algunos casos, las palabras de código pueden tener menos "letras" o posiciones que el número total de ácidos nucleicos codificados por las palabras de código, por ejemplo, donde cada palabra de código codifica para un ácido nucleico diferente.

Tal código de detección de error y/o corrección de error puede adoptar una variedad de formas. Previamente se han desarrollado una variedad de tales códigos en otros contextos tales como la industria de telecomunicaciones, tales como códigos de Golay o códigos de Hamming. En un conjunto de realizaciones, se asignan las secuencias de lectura o los patrones de unión de las sondas de ácido nucleico de manera que no se asigna cada combinación posible.

Por ejemplo, si son posibles 4 secuencias de lectura y una sonda de ácido nucleico primaria contiene 2 secuencias de lectura, entonces pueden identificarse hasta 6 sondas de ácido nucleico primarias; pero el número de sondas de ácido nucleico primarias usadas puede ser menor de 6. De manera similar, para k secuencias de lectura en una población con n secuencias de lectura en cada sonda de ácido nucleico primaria, pueden producirse $\binom{n}{k}$ sondas diferentes, pero el número de sondas de ácido nucleico primarias que se usan puede ser cualquier número mayor o

menor que $\binom{n}{k}$. Además, estas pueden asignarse al azar, o asignarse de maneras específicas para aumentar la capacidad para detectar y/o corregir errores.

Como otro ejemplo, si se usan múltiples rondas de sondas de ácido nucleico, el número de rondas puede elegirse arbitrariamente. Si, en cada ronda, cada diana puede dar lugar a dos resultados posibles, tal como detectarse o no detectarse, pueden ser posibles hasta 2^n dianas diferentes para n rondas de sondas, pero el número de dianas de ácido nucleico que se usan realmente puede ser cualquier número menor que 2^n . Por ejemplo, si, en cada ronda, cada diana puede dar lugar a más de dos resultados posibles, tal como detectarse en canales de color diferentes, pueden ser posibles más de 2^n (por ejemplo 3^n , 4^n , ...) dianas diferentes para n rondas de sondas. En algunos casos, el número de dianas de ácido nucleico que se usan realmente puede ser cualquier número menor que este número. Además, estas pueden asignarse al azar, o asignarse de maneras específicas para aumentar la capacidad para detectar y/o corregir errores.

Por ejemplo, en un conjunto de realizaciones, las palabras de código o sondas de ácido nucleico pueden asignarse dentro de un espacio de código de manera que las asignaciones están separadas por una distancia de Hamming, que mide el número de "lecturas" incorrectas en un patrón dado que hace que se interprete erróneamente la sonda de ácido nucleico como una sonda de ácido nucleico válida diferente. En determinados casos, la distancia de Hamming puede ser de al menos 2, al menos 3, al menos 4, al menos 5, al menos 6, o similares. Además, en un conjunto de realizaciones, las asignaciones pueden formarse como un código de Hamming, por ejemplo, un código de Hamming(7, 4), un código de Hamming(15, 11), un código de Hamming(31, 26), un código de Hamming(63, 57), un código de Hamming(127, 120), etc. En otro conjunto de realizaciones, las asignaciones pueden formar un código SECDED, por ejemplo, un código SECDED(8,4), un código SECDED(16,4), un código SECDED(16, 11), un código SECDED(22, 16), un código SECDED(39, 32), un código SECDED(72, 64), etc. En aún otro conjunto de realizaciones, las asignaciones pueden formar un código de Golay binario extendido, un código de Golay binario perfecto, o un código de Golay ternario. En otro conjunto de realizaciones, las asignaciones pueden representar un subconjunto de los valores posibles tomados de cualquiera de los códigos descritos anteriormente.

Por ejemplo, puede formarse un código con las mismas propiedades de corrección de error del código SECDED usando sólo palabras binarias que contienen un número fijo de "1" bits, tal como 4, para codificar para las dianas. En otro conjunto de realizaciones, las asignaciones pueden representar un subconjunto de los valores posibles tomados de los códigos descritos anteriormente con el propósito de abordar errores de lectura asimétricos. Por ejemplo, en algunos casos, un código en el que puede fijarse el número de "1" bits para todas las palabras binarias usadas puede eliminar la medición sesgada de palabras con números diferentes de "1" cuando las tasas a las que "0" bits se miden como "1" o "1" bits se miden como "0" son diferentes.

Por consiguiente, en algunas realizaciones, una vez determinada la palabra de código (por ejemplo, tal como se comenta en el presente documento), puede compararse la palabra de código con las palabras de código de ácido nucleico conocidas. Si se encuentra una coincidencia, entonces puede identificarse o determinarse la diana de ácido nucleico. Si no se encuentra coincidencia, entonces puede identificarse un error en la lectura de la palabra de código. En algunos casos, también puede aplicarse corrección de error para determinar la palabra de código correcta y, por tanto, dar como resultado la identidad correcta de la diana de ácido nucleico. En algunos casos, las palabras de código pueden seleccionarse de manera que, suponiendo que sólo hay un error presente, sólo está disponible una palabra de código correcta posible y, por tanto, sólo es posible una identidad correcta de la diana de ácido nucleico. En algunos casos, esto también puede generalizarse para distancias de Hamming o separaciones de palabra de código más grandes; por ejemplo, las palabras de código pueden seleccionarse de manera que si están presentes dos, tres, o cuatro errores (o más en algunos casos), sólo está disponible una palabra de código correcta posible y, por tanto, sólo es posible una identidad correcta de las dianas de ácido nucleico.

El código de corrección de error puede ser un código de corrección de error binario, o puede estar basado en otros sistemas de numeración, por ejemplo, códigos de corrección de error ternarios o cuaternarios. Por ejemplo, en un conjunto de realizaciones, puede usarse más de un tipo de entidad de señalización y asignarse a números diferentes dentro del código de corrección de error. Por tanto, como un ejemplo no limitativo, una primera entidad de señalización (o más de una entidad de señalización, en algunos casos) puede asignarse como "1" y una segunda entidad de señalización (o más de una entidad de señalización, en algunos casos) puede asignarse como "2" (indicando "0" que no hay ninguna entidad de señalización presente), y pueden distribuirse las palabras de código para definir un código de corrección de error ternario. De manera similar, una tercera entidad de señalización puede asignarse adicionalmente como "3" para constituir un código de corrección de error cuaternario, etc.

Tal como se comentó anteriormente, en determinados aspectos, se determinan entidades de señalización, por ejemplo, para determinar sondas de ácido nucleico y/o para crear palabras de código. En algunos casos, pueden determinarse entidades de señalización dentro de una muestra, por ejemplo, espacialmente, usando una variedad de técnicas. En algunas realizaciones, las entidades de señalización pueden ser fluorescentes, y pueden usarse técnicas para determinar la fluorescencia dentro de una muestra, tales como microscopía de fluorescencia o microscopía confocal, para identificar espacialmente las posiciones de las entidades de señalización dentro de una

célula. En algunos casos, las posiciones de las entidades dentro de la muestra pueden determinarse en dos o incluso tres dimensiones. Además, en algunas realizaciones, puede determinarse más de una entidad de señalización a la vez (por ejemplo, entidades de señalización con colores o emisiones diferentes), y/o secuencialmente.

Además, en algunas realizaciones, puede determinarse un nivel de confianza para la diana de ácido nucleico identificada. Por ejemplo, el nivel de confianza puede determinarse usando una razón del número de coincidencias exactas con respecto al número de coincidencias que tienen uno o más errores de un bit. En algunos casos, pueden usarse sólo coincidencias que tienen una razón de confianza mayor que un valor determinado. Por ejemplo, en determinadas realizaciones, pueden aceptarse las coincidencias sólo si la razón de confianza para la coincidencia es mayor que aproximadamente 0,01, mayor que aproximadamente 0,03, mayor que aproximadamente 0,05, mayor que aproximadamente 0,1, mayor que aproximadamente 0,3, mayor que aproximadamente 0,5, mayor que aproximadamente 1, mayor que aproximadamente 3, mayor que aproximadamente 5, mayor que aproximadamente 10, mayor que aproximadamente 30, mayor que aproximadamente 50, mayor que aproximadamente 100, mayor que aproximadamente 300, mayor que aproximadamente 500, mayor que aproximadamente 1000, o cualquier otro valor adecuado. Además, en algunas realizaciones, pueden aceptarse las coincidencias sólo si la razón de confianza para la diana de ácido nucleico identificada es mayor que un patrón interno o control falso positivo en aproximadamente 0,01, aproximadamente 0,03, aproximadamente 0,05, aproximadamente 0,1, aproximadamente 0,3, aproximadamente 0,5, aproximadamente 1, aproximadamente 3, aproximadamente 5, aproximadamente 10, aproximadamente 30, aproximadamente 50, aproximadamente 100, aproximadamente 300, aproximadamente 500, aproximadamente 1000, o cualquier otro valor adecuado.

En algunas realizaciones, pueden determinarse las posiciones espaciales de las entidades (y, por tanto, las sondas de ácido nucleico con las que pueden estar asociadas las entidades) a resoluciones relativamente altas. Por ejemplo, las posiciones pueden determinarse a resoluciones espaciales mejores de aproximadamente 100 micrómetros, mejores de aproximadamente 30 micrómetros, mejores de aproximadamente 10 micrómetros, mejores de aproximadamente 3 micrómetros, mejores de aproximadamente 1 micrómetro, mejores de aproximadamente 800 nm, mejores de aproximadamente 600 nm, mejores de aproximadamente 500 nm, mejores de aproximadamente 400 nm, mejores de aproximadamente 300 nm, mejores de aproximadamente 200 nm, mejores de aproximadamente 100 nm, mejores de aproximadamente 90 nm, mejores de aproximadamente 80 nm, mejores de aproximadamente 70 nm, mejores de aproximadamente 60 nm, mejores de aproximadamente 50 nm, mejores de aproximadamente 40 nm, mejores de aproximadamente 30 nm, mejores de aproximadamente 20 nm, o mejores de aproximadamente 10 nm, etc.

Hay una variedad de técnicas capaces de determinar u obtener imágenes de las posiciones espaciales de las entidades ópticamente, por ejemplo, usando microscopía de fluorescencia. En algunos casos, las posiciones espaciales pueden determinarse a superresoluciones, o a resoluciones mejores que la longitud de onda de la luz o el límite de difracción. Los ejemplos no limitativos incluyen STORM (microscopía de reconstrucción óptica estocástica), STED (microscopía de agotamiento de emisión estimulada), NSOM (microscopía óptica de barrido de campo cercano), microscopía 4Pi, SIM (microscopía de iluminación estructurada), microscopía de SMI (iluminación espacialmente modulada), RESOLFT (microscopía de transición de fluorescencia ópticamente lineal saturable reversible), GSD (microscopía de agotamiento de estado fundamental), SSIM (microscopía de iluminación estructurada saturada), SPDM (microscopía de distancia de precisión espectral), microscopía de localización fotoactivada (PALM), microscopía de localización por fotoactivación de fluorescencia (FPALM), LIMON (microscopía de nanodimensionamiento microscópica óptica en 3D), obtención de imágenes por fluctuación óptica de superresolución (SOFI), o similares. Véanse, por ejemplo, la patente estadounidense n.º 7.838.302, concedida el 23 de noviembre de 2010, titulada "Sub-Diffraction Limit Image Resolution and Other Imaging Techniques", de Zhuang, *et al.*; la patente estadounidense n.º 8.564.792, concedida el 22 de octubre de 2013, titulada "Sub-diffraction Limit Image Resolution in Three Dimensions", de Zhuang, *et al.*; o la publicación de solicitud de patente internacional n.º WO 2013/090360, publicada el 20 de junio de 2013, titulada "High Resolution Dual-Objective Microscopy", de Zhuang, *et al.*

Como un ejemplo no limitativo ilustrativo, en un conjunto de realizaciones, pueden obtenerse imágenes de la muestra con una alta apertura numérica, un objetivo de inmersión en aceite con un aumento de 100X y luz recopilada en una cámara con CCD de multiplicación de electrones. En otro ejemplo, pueden obtenerse imágenes de la muestra con una alta apertura numérica, una lente de inmersión en aceite con un aumento de 40X y luz recopilada con una cámara con CMOS científico de campo amplio. Con diferentes combinaciones de objetivos y cámaras, un solo campo de visión puede corresponder a no menos de 40x40 micrómetros, 80x80 micrómetros, 120x120 micrómetros, 240x240 micrómetros, 340x340 micrómetros, o 500x500 micrómetros, etc., en diversas realizaciones no limitativas. De manera similar, un solo píxel de cámara puede corresponder, en algunas realizaciones, a regiones de la muestra de no menos de 80x80 nm, 120x120 nm, 160x160 nm, 240x240 nm, o 300x300 nm, etc. En otro ejemplo, pueden obtenerse imágenes de la muestra con una baja apertura numérica, una lente de aire con un aumento de 10X y luz recopilada con una cámara con sCMOS. En realizaciones adicionales, puede seccionarse ópticamente la muestra iluminándola a través de un único foco o unos múltiples focos limitado(s) por difracción sometido(s) a barrido generado(s) o bien mediante espejos de barrido o bien mediante un disco giratorio y lo recogido hacerse pasar a través de un único orificio pasante o unos múltiples orificios pasantes. En otra

realización, la muestra también puede iluminarse a través de una lámina delgada de luz generada a través de uno cualquiera de múltiples métodos conocidos por los expertos en la técnica.

En una realización, la muestra puede iluminarse mediante líneas láser en modo gaussiano individuales. En algunas realizaciones, el perfil de iluminación puede aplanarse haciendo pasar estas líneas láser a través de una fibra multimodal que se hace vibrar a través de un medio piezoeléctrico u otro medio mecánico. En algunas realizaciones, el perfil de iluminación puede aplanarse haciendo pasar haces gaussianos monomodales a través de una variedad de dispositivos de conformación de haces de refracción, tales como piShaper o una serie de lentes Powell apiladas. En aún otro conjunto de realizaciones, los haces gaussianos pueden hacerse pasar a través de una variedad de elementos de difusión diferentes, tales como difusores diseñados por ingeniería o de vidrio deslustrado, que pueden hacerse girar en algunos casos a altas velocidades para eliminar la granularidad láser residual. En aún otra realización, la iluminación láser puede hacerse pasar a través de una serie de matrices de pequeñas lentes para producir imágenes solapantes de la iluminación que se aproximan a un campo de iluminación plano.

En algunas realizaciones, pueden determinarse los centroides de las posiciones espaciales de las entidades. Por ejemplo, puede determinarse un centroide de una entidad de señalización dentro de una imagen o serie de imágenes usando algoritmos de análisis de imágenes conocidos por los expertos habituales en la técnica. En algunos casos, los algoritmos pueden seleccionarse para determinar emisores individuales no solapantes y/o emisores individuales parcialmente solapantes en una muestra. Los ejemplos no limitativos de técnicas adecuadas incluyen un algoritmo de probabilidad máxima, un algoritmo de mínimos cuadrados, un algoritmo bayesiano, un algoritmo de detección comprimida, o similares. También pueden usarse combinaciones de estas técnicas en algunos casos.

Además, la entidad de señalización puede inactivarse en algunos casos. Por ejemplo, en algunas realizaciones, puede aplicarse una primera sonda de ácido nucleico secundaria que contiene una entidad de señalización a una muestra que puede reconocer una primera secuencia de lectura, entonces puede inactivarse la primera sonda de ácido nucleico secundaria antes de aplicar una segunda sonda de ácido nucleico secundaria a la muestra. Si se usan múltiples entidades de señalización, pueden usarse técnicas iguales o diferentes para inactivar las entidades de señalización, y puede inactivarse parte o la totalidad de las múltiples entidades de señalización, por ejemplo, secuencialmente o simultáneamente.

La inactivación puede estar provocada por la eliminación de la entidad de señalización (por ejemplo, a partir de la muestra, o a partir de la sonda de ácido nucleico, etc.), y/o por la alteración química de la entidad de señalización de cierta manera, por ejemplo, mediante fotoblanqueo de la entidad de señalización, blanqueo o alteración química de la estructura de la entidad de señalización, por ejemplo, mediante reducción, etc.). Por ejemplo, en un conjunto de realizaciones, puede inactivarse una entidad de señalización fluorescente mediante técnicas químicas u ópticas tales como oxidación, fotoblanqueo, blanqueo químico, lavado riguroso o digestión o reacción enzimática mediante la exposición a una enzima, disociación de la entidad de señalización a partir de otros componentes (por ejemplo, una sonda), reacción química de la entidad de señalización (por ejemplo, con un reactante capaz de alterar la estructura de la entidad de señalización), o similares. Por ejemplo, el blanqueo puede producirse mediante la exposición a oxígeno, agentes reductores, o la entidad de señalización puede escindir-se químicamente a partir de la sonda de ácido nucleico y eliminarse por lavado a través de un flujo de fluido.

En algunas realizaciones, diversas sondas de ácido nucleico (incluyendo sondas de ácido nucleico primarias y/o secundarias) pueden incluir una o más entidades de señalización. Si se usa más de una sonda de ácido nucleico, cada una de las entidades de señalización pueden ser iguales o diferentes. En determinadas realizaciones, una entidad de señalización es cualquier entidad capaz de emitir luz. Por ejemplo, en una realización, la entidad de señalización es fluorescente. En otras realizaciones, la entidad de señalización puede ser fosforescente, radiactiva, absorbente, etc. En algunos casos, la entidad de señalización es cualquier entidad que puede determinarse dentro de una muestra a resoluciones relativamente altas, por ejemplo, a resoluciones mejores que la longitud de onda de la luz visible o el límite de difracción. La entidad de señalización puede ser, por ejemplo, un colorante, una molécula pequeña, un péptido o una proteína, o similares. La entidad de señalización puede ser una sola molécula en algunos casos. Si se usan múltiples sondas de ácido nucleico secundarias, las sondas de ácido nucleico pueden comprender entidades de señalización iguales o diferentes.

Los ejemplos no limitativos de entidades de señalización incluyen entidades fluorescentes (fluoróforos) o entidades fosforescentes, por ejemplo, colorantes de cianina (por ejemplo, Cy2, Cy3, Cy3B, Cy5, Cy5.5, Cy7, etc.), colorantes Alexa Fluor, colorantes Atto, colorantes fotoconmutables, colorantes fotoactivables, colorantes fluorescentes, nanopartículas metálicas, nanopartículas semiconductoras o "puntos cuánticos", proteínas fluorescentes tales como GFP (proteína fluorescente verde), o proteínas fluorescentes fotoactivables, tales como PAGFP, PSCFP, PSCFP2, Dendra, Dendra2, EosFP, tdEos, mEos2, mEos3, PAmCherry, PAtagRFP, mMaple, mMaple2, y mMaple3. Los expertos habituales en la técnica conocen otras entidades de señalización adecuadas. Véanse, por ejemplo, la patente estadounidense n.º 7.838.302 o la solicitud de patente estadounidense con n.º de serie 61/979.436, cada una de ellas incorporada en el presente documento mediante referencia en su totalidad.

En un conjunto de realizaciones, la entidad de señalización puede unirse a una secuencia de oligonucleótidos a

través de un enlace que puede escindirse para liberar la entidad de señalización. En un conjunto de realizaciones, puede conjugarse un fluoróforo a un oligonucleótido a través de un enlace escindible, tal como un enlace fotoescindible. Los ejemplos no limitativos de enlaces fotoescindibles incluyen, pero no se limitan a, 1-(2-nitrofenil)etilo, 2-nitrobencilo, fosforamida de biotina, fosforamida acrílica, dietilaminocumarina, 1-(4,5-dimetoxi-2-nitrofenil)etilo, ciclododecil-(dimetoxi-2-nitrofenil)etilo, 4-aminometil-3-nitrobencilo, éster de ácido (4-nitro-3-(1-clorocarboniloxietil)fenil)metil-S-acetiltoico, éster de ácido (4-nitro-3-(1-clorocarboniloxietil)fenil)metil-3-(2-piridilditiopropiónico), 3-(4,4'-dimetoxitritil)-1-(2-nitrofenil)-propano-1,3-diol-[2-cianoetil-(N,N-diisopropil)]-fosforamida, 1-[2-nitro-5-(6-trifluoroacetilcaproamidometil)fenil]-etil-[2-cianoetil-(N,N-diisopropil)]-fosforamida, 1-[2-nitro-5-(6-(4,4'-dimetoxitritiloxi)butiramidometil)fenil]-etil-[2-cianoetil-(N,N-diisopropil)]-fosforamida, 1-[2-nitro-5-(6-(N-(4,4'-dimetoxitritil))-biotinamidocaproamidometil)fenil]-etil-[2-cianoetil-(N,N-diisopropil)]-fosforamida, o grupos de unión similares. En otro conjunto de realizaciones, puede conjugarse el fluoróforo a un oligonucleótido a través de un enlace disulfuro. El enlace disulfuro puede escindirse mediante una variedad de agentes reductores tales como, pero sin limitarse a, ditiotretol, ditioeritrol, beta-mercaptoetanol, borohidruro de sodio, tiorredoxina, glutarredoxina, tripsinógeno, hidrazina, hidruro de diisobutilaluminio, ácido oxálico, ácido fórmico, ácido ascórbico, ácido fosforoso, cloruro de estaño, glutatión, tioglicolato, 2,3-dimercaptopropanol, 2-mercaptoetilamina, 2-aminoetanol, tris(2-carboxietil)fosfina, bis(2-mercaptoetil)sulfona, N,N'-dimetil-N,N'-bis(mercaptoacetil)hidrazina, 3-mercaptopropionato, dimetilformamida, tiopropil-agarosa, tri-*n*-butilfosfina, cisteína, sulfato de hierro, sulfito de sodio, fosfito, hipofosfito, fosforotioato, o similares, y/o combinaciones de cualquiera de estos. En otra realización, puede conjugarse el fluoróforo a un oligonucleótido a través de uno o más nucleótidos modificados con fosforotioato en los que la modificación con azufre sustituye al oxígeno que forma puente y/o que no forma puente. El fluoróforo puede escindirse a partir del oligonucleótido, en determinadas realizaciones, a través de la adición de compuestos tales como, pero sin limitarse a, yodoetanol, yodo mezclado en etanol, nitrato de plata, o cloruro de mercurio. En aún otro conjunto de realizaciones, la entidad de señalización puede inactivarse químicamente a través de reducción u oxidación. Por ejemplo, en una realización, puede reducirse un cromóforo tal como Cy5 o Cy7 usando borohidruro de sodio para dar un estado estable sin fluorescencia. En todavía otro conjunto de realizaciones, puede conjugarse un fluoróforo a un oligonucleótido a través de un enlace azo, y el enlace azo puede escindirse con 2-[(2-N-arilamino)fenilazo]piridina. En aún otro conjunto de realizaciones, puede conjugarse un fluoróforo a un oligonucleótido a través de un segmento de ácido nucleico adecuado que puede escindirse tras la exposición adecuada a ADNasa, por ejemplo, una exodesoxirribonucleasa o una endodesoxirribonucleasa. Los ejemplos incluyen, pero no se limitan a, desoxirribonucleasa I o desoxirribonucleasa II. En un conjunto de realizaciones, la escisión puede producirse a través de una endonucleasa de restricción. Los ejemplos no limitativos de endonucleasas de restricción posiblemente adecuadas incluyen BamHI, BsrI, NotI, XmaI, PspAI, DpnI, MboI, MnlI, Eco57I, Ksp632I, DraII, AhaI, SmaI, MluI, HpaI, ApaI, BclI, BstEII, TaqI, EcoRI, SacI, HindII, HaeII, DraII, Tsp509I, Sau3AI, PaeI, etc. Se han estudiado con detalle más de 3000 enzimas de restricción, y están disponibles comercialmente más de 600 de estas. En aún otro conjunto de realizaciones, puede conjugarse un fluoróforo a biotina, y conjugarse el oligonucleótido a avidina o estreptavidina. Una interacción entre biotina y avidina o estreptavidina permite jugar el fluoróforo al oligonucleótido, mientras que con una exposición suficiente a un exceso de adición, la biotina libre puede "vencer" a la unión y provocar de ese modo que se produzca la escisión. Además, en otro conjunto de realizaciones, las sondas pueden eliminarse usando "sondas de sujeción" correspondientes, que comprenden la misma secuencia que la sonda, así como un número adicional de bases de homología con respecto a las sondas de codificación (por ejemplo, 1-20 bases adicionales, por ejemplo, 5 bases adicionales). Estas sondas pueden eliminar la sonda de lectura marcada a través de una interacción de desplazamiento de cadena.

Tal como se usa en el presente documento, el término "luz" se refiere generalmente a radiación electromagnética, que tiene cualquier longitud de onda (o, de manera equivalente, frecuencia) adecuada. Por ejemplo, en algunas realizaciones, la luz puede incluir longitudes de onda en el rango óptico o visual (por ejemplo, que tienen una longitud de onda de entre aproximadamente 400 nm y aproximadamente 700 nm, es decir, "luz visible"), longitudes de onda infrarrojas (por ejemplo, que tienen una longitud de onda de entre aproximadamente 300 micrómetros y 700 nm), longitudes de onda ultravioletas (por ejemplo, que tienen una longitud de onda de entre aproximadamente 400 nm y aproximadamente 10 nm), o similares. En determinados casos, tal como se comenta con detalle a continuación, puede usarse más de una entidad, es decir, entidades que son químicamente diferentes o distintas, por ejemplo, desde el punto de vista estructural. Sin embargo, en otros casos, las entidades pueden ser idénticas químicamente o al menos sustancialmente idénticas químicamente.

En un conjunto de realizaciones, la entidad de señalización es "conmutable", es decir, la entidad puede conmutarse entre dos o más estados, al menos uno de los cuales emite luz que tiene una longitud de onda deseada. En el/los otro(s) estado(s), la entidad puede no emitir luz, o emitir luz a una longitud de onda diferente. Por ejemplo, una entidad puede "activarse" a un primer estado capaz de producir luz que tiene una longitud de onda deseada, y "desactivarse" a un segundo estado que no es capaz de emitir luz de la misma longitud de onda. Una entidad es "fotoactivable" si puede activarse mediante luz incidente de una longitud de onda adecuada. Como un ejemplo no limitativo, Cy5 puede conmutarse entre un estado fluorescente y un estado oscuro de manera controlada y reversible mediante luz de diferentes longitudes de onda, es decir, la luz roja de 633 nm (o 642 nm, 647 nm, 656 nm) puede conmutar o desactivar Cy5 a un estado oscuro estable, mientras que la luz verde de 405 nm puede conmutar o activar Cy5 de vuelta al estado fluorescente. En algunos casos, la entidad puede conmutarse reversiblemente entre los dos o más estados, por ejemplo, tras la exposición a los estímulos apropiados. Por ejemplo, puede usarse un

primer estímulo (por ejemplo, una primera longitud de onda de luz) para activar la entidad conmutable, mientras que puede usarse un segundo estímulo (por ejemplo, una segunda longitud de onda de luz) para desactivar la entidad conmutable, por ejemplo, a un estado de no emisión. Puede usarse cualquier método adecuado para activar la entidad. Por ejemplo, en una realización, puede usarse luz incidente de una longitud de onda adecuada para activar la entidad para emitir luz, es decir, la entidad es "fotoconmutable". Por tanto, la entidad fotoconmutable puede conmutarse entre diferentes estados de emisión y no emisión de luz mediante luz incidente, por ejemplo, de diferentes longitudes de onda. La luz puede ser monocromática (por ejemplo, producida usando un láser) o policromática. En otra realización, la entidad puede activarse tras la estimulación mediante campo eléctrico y/o campo magnético. En otras realizaciones, la entidad puede activarse tras la exposición a un entorno químico adecuado, por ejemplo, ajustando el pH, o induciendo una reacción química reversible que involucra a la entidad, etc. De manera similar, puede usarse cualquier método adecuado para desactivar la entidad, y no es necesario que los métodos de activación y desactivación de la entidad sean iguales. Por ejemplo, la entidad puede desactivarse tras la exposición a luz incidente de una longitud de onda adecuada, o la entidad puede desactivarse esperando un tiempo suficiente.

Normalmente, un experto habitual en la técnica puede identificar una entidad "conmutable" determinando las condiciones en las que una entidad en un primer estado puede emitir luz cuando se expone a una longitud de onda de excitación, conmutando la entidad desde el primer estado hasta el segundo estado, por ejemplo, tras la exposición a luz de una longitud de onda de conmutación, mostrando entonces que la entidad, mientras está en el segundo estado, deja de emitir luz (o emite luz a una intensidad muy reducida) cuando se expone a la longitud de onda de excitación.

En un conjunto de realizaciones, tal como se comentó, una entidad conmutable puede conmutarse tras la exposición a luz. En algunos casos, la luz usada para activar la entidad conmutable puede proceder de una fuente externa, por ejemplo, una fuente de luz tal como una fuente de luz láser, otra entidad emisora de luz próxima a la entidad conmutable, etc. En algunos casos, la segunda entidad emisora de luz puede ser una entidad fluorescente, y en determinadas realizaciones, la segunda entidad emisora de luz también puede ser por sí misma una entidad conmutable.

En algunas realizaciones, la entidad conmutable incluye una primera porción emisora de luz (por ejemplo, un fluoróforo), y una segunda porción que activa o "conmuta" la primera porción. Por ejemplo, tras la exposición a luz, la segunda porción de la entidad conmutable puede activar la primera porción, provocando que la primera porción emita luz. Los ejemplos de porciones activadoras incluyen, pero no se limitan a, Alexa Fluor 405 (Invitrogen), Alexa Fluor 488 (Invitrogen), Cy2 (GE Healthcare), Cy3 (GE Healthcare), Cy3B (GE Healthcare), Cy3.5 (GE Healthcare), u otros colorantes adecuados. Los ejemplos de porciones emisoras de luz incluyen, pero no se limitan a, Cy5, Cy5.5 (GE Healthcare), Cy7 (GE Healthcare), Alexa Fluor 647 (Invitrogen), Alexa Fluor 680 (Invitrogen), Alexa Fluor 700 (Invitrogen), Alexa Fluor 750 (Invitrogen), Alexa Fluor 790 (Invitrogen), DiD, DiR, YOYO-3 (Invitrogen), YO-PRO-3 (Invitrogen), TOT-3 (Invitrogen), TO-PRO-3 (Invitrogen), u otros colorantes adecuados. Estos pueden unirse juntos, por ejemplo, covalentemente, por ejemplo, directamente, o a través de un grupo de unión, por ejemplo, formando compuestos tales como, pero sin limitarse a, Cy5-Alexa Fluor 405, Cy5-Alexa Fluor 488, Cy5-Cy2, Cy5-Cy3, Cy5-Cy3.5, Cy5.5-Alexa Fluor 405, Cy5.5-Alexa Fluor 488, Cy5.5-Cy2, Cy5.5-Cy3, Cy5.5-Cy3.5, Cy7-Alexa Fluor 405, Cy7-Alexa Fluor 488, Cy7-Cy2, Cy7-Cy3, Cy7-Cy3.5, Alexa Fluor 647-Alexa Fluor 405, Alexa Fluor 647-Alexa Fluor 488, Alexa Fluor 647-Cy2, Alexa Fluor 647-Cy3, Alexa Fluor 647-Cy3.5, Alexa Fluor 750-Alexa Fluor 405, Alexa Fluor 750-Alexa Fluor 488, Alexa Fluor 750-Cy2, Alexa Fluor 750-Cy3, o Alexa Fluor 750-Cy3.5. Los expertos habituales en la técnica serán conscientes de las estructuras de estos y otros compuestos, muchos de los cuales están disponibles comercialmente. Las porciones pueden unirse a través de un enlace covalente, o mediante un grupo de unión, tal como los descritos con detalle a continuación. Otras porciones emisoras de luz o activadoras pueden incluir porciones que tienen dos átomos de nitrógeno cuaternizados unidos mediante una cadena de polimetina, donde cada nitrógeno forma parte independientemente de un resto heteroaromático, tal como pirrol, imidazol, tiazol, piridina, quinolina, indol, benzotiazol, etc., o parte de una amina no aromática. En algunos casos, puede haber 5, 6, 7, 8, 9, o más átomos de carbono entre los dos átomos de nitrógeno.

En determinados casos, cada una de la porción emisora de luz y las porciones activadoras, cuando están aisladas unas de otras, pueden ser fluoróforos, es decir, entidades que pueden emitir luz de una determinada longitud de onda de emisión cuando se exponen a un estímulo, por ejemplo, una longitud de onda de excitación. Sin embargo, cuando se forma una entidad conmutable que comprende el primer fluoróforo y el segundo fluoróforo, el primer fluoróforo forma una primera porción emisora de luz y el segundo fluoróforo forma una porción activadora que activa o "conmuta" la primera porción en respuesta a un estímulo. Por ejemplo, la entidad conmutable puede comprender un primer fluoróforo directamente unido al segundo fluoróforo, o la primera y segunda entidades pueden estar conectadas a través de un grupo de unión o una entidad común. Puede someterse a prueba si un par de porción emisora de luz y porción activadora produce una entidad conmutable adecuada mediante métodos conocidos por los expertos habituales en la técnica. Por ejemplo, puede usarse luz de diversas longitudes de onda para estimular el par y puede medirse la luz de emisión procedente de la porción emisora de luz para determinar si el par realiza una conmutación adecuada.

Como un ejemplo no limitativo, Cy3 y Cy5 pueden unirse juntos para formar una entidad de este tipo. En este

ejemplo, Cy3 es una porción activadora que es capaz de activar Cy5, la porción emisora de luz. Por tanto, luz en o cerca del máximo de absorción (por ejemplo, luz cerca de 532 nm para Cy3) de la activación de la segunda porción de la entidad puede provocar que la porción active la primera porción emisora de luz, provocando de ese modo que la primera porción emita luz (por ejemplo, cerca de 647 nm para Cy5). Véase, por ejemplo, la patente estadounidense n.º 7.838.302. En algunos casos, la primera porción emisora de luz puede desactivarse posteriormente mediante cualquier técnica adecuada (por ejemplo, dirigiendo luz roja de 647 nm a la porción Cy5 de la molécula).

Otros ejemplos no limitativos de porciones activadoras posiblemente adecuadas incluyen 1,5-IAEDANS, 1,8-ANS, 4-metilumbeliferona, 5-carboxi-2,7-diclorofluoresceína, 5-carboxifluoresceína (5-FAM), 5-carboxinaftofluoresceína, 5-carboxitetrametilrodamina (5-TAMRA), 5-FAM (5-carboxifluoresceína), 5-HAT (hidroxi-triptamina), 5-hidroxi-triptamina (HAT), 5-ROX (carboxi-X-rodamina), 5-TAMRA (5-carboxitetrametilrodamina), 6-carboxirrodamina 6G, 6-CR 6G, 6-JOE, 7-amino-4-metilcumarina, 7-aminoactinomicina D (7-AAD), 7-hidroxi-4-metilcumarina, 9-amino-6-cloro-2-metoxiacridina, ABQ, fucsina ácida, ACMA (9-amino-6-cloro-2-metoxiacridina), naranja de acridina, rojo de acridina, amarillo de acridina, acriflavina, acriflavina-Feulgen SITSA, Alexa Fluor 350, Alexa Fluor 405, Alexa Fluor 430, Alexa Fluor 488, Alexa Fluor 500, Alexa Fluor 514, Alexa Fluor 532, Alexa Fluor 546, Alexa Fluor 555, Alexa Fluor 568, Alexa Fluor 594, Alexa Fluor 610, Alexa Fluor 633, Alexa Fluor 635, complejo de alizarina, rojo de alizarina, AMC, AMCA-S, AMCA (aminometilcumarina), AMCA-X, aminoactinomicina D, aminocumarina, aminometilcumarina (AMCA), azul de anilina, estearato de antrociolo, APTRA-BTC, APTS, rojo brillante de astrazón 4G, naranja de astrazón R, rojo de astrazón 6B, amarillo de astrazón 7 GLL, atabrina, ATTO 390, ATTO 425, ATTO 465, ATTO 488, ATTO 495, ATTO 520, ATTO 532, ATTO 550, ATTO 565, ATTO 590, ATTO 594, ATTO 610, ATTO 611X, ATTO 620, ATTO 633, ATTO 635, ATTO 647, ATTO 647N, ATTO 655, ATTO 680, ATTO 700, ATTO 725, ATTO 740, ATTO-TAG CBQCA, ATTO-TAG FQ, auramina, aurofosfina G, aurofosfina, BAO 9 (bisaminofeniloxadiazol), BCECF (pH alto), BCECF (pH bajo), sulfato de berberina, bimano, bisbenzamida, bisbencimida (Hoechst), bis-BTC, Blancophor FFG, Blancophor SV, BOBO-1, BOBO-3, Bodipy 492/515, Bodipy 493/503, Bodipy 500/510, Bodipy 505/515, Bodipy 530/550, Bodipy 542/563, Bodipy 558/568, Bodipy 564/570, Bodipy 576/589, Bodipy 581/591, Bodipy 630/650-X, Bodipy 650/665-X, Bodipy 665/676, Bodipy FI, Bodipy FL ATP, Bodipy FI-ceramida, Bodipy R6G, Bodipy TMR, conjugado de Bodipy TMR-X, Bodipy TMR-X, SE, Bodipy TR, Bodipy TR ATP, Bodipy TR-X SE, BO-PRO-1, BO-PRO-3, sulfoflavina brillante FF, BTC, BTC-5N, calceína, azul de calceína, Calcium Crimson, colorante Calcium Green, Calcium Green-1 Ca^{2+} , Calcium Green-2 Ca^{2+} , Calcium Green-5N Ca^{2+} , Calcium Green-C18 Ca^{2+} , Calcium Orange, blanco de calcoflúor, carboxi-X-rodamina (5-ROX), azul cascada, amarillo cascada, catecolamina, CCF2 (GeneBlazer), CFDA, cromomicina A, cromomicina A, CL-NERF, CMFDA, cumarina-faloidina, CPM-metilcumarina, CTC, CTC-formazano, Cy2, Cy3.1 8, Cy3.5, Cy3, Cy5.1 8, fluorosensor de AMP cíclico (FiCRhR), dabcilo, dansilo, dansil-amina, dansil-cadaverina, cloruro de dansilo, dansil-DHPE, fluoruro de dansilo, DAPI, dapoxilo, dapoxilo 2, dapoxilo 3'-DCFDA, DCFH (diacetato de diclorodihidrofluoresceína), DDAO, DHR (dihidrorrodamina 123), di-4-ANEPPS, di-8-ANEPPS (sin porción), DiA (4-di-16-ASP), diacetato de diclorodihidrofluoresceína (DCFH), DiD-trazador lipófilo, DiD (DiIC18(5)), DIDS, dihidrorrodamina 123 (DHR), Dil (DiIC18(3)), dinitrofenol, DiO (DiOC18(3)), DiR, DiR (DiIC18(7)), DM-NERF (pH alto), DNP, dopamina, DTAF, DY-630-NHS, DY-635-NHS, DyLight 405, DyLight 488, DyLight 549, DyLight 633, DyLight 649, DyLight 680, DyLight 800, ELF 97, eosina, eritrosina, eritrosina ITC, bromuro de etidio, homodímero de etidio 1 (EtD-1), eucrisina, EukoLight, cloruro de europio (III), Fast Blue, FDA, Feulgen (pararosanilina), FIF (fluorescencia inducida por formaldehído), FITC, Flazo Orange, Fluo-3, Fluo-4, fluoresceína (FITC), diacetato de fluoresceína, Fluoro-Emerald, Fluoro-Gold (hidroxiestilbamidina), Fluor-Ruby, FluorX, FM 1-43, FM 4-46, Fura Red (pH alto), Fura Red/Fluo-3, Fura-2, Fura-2/BCECF, rojo brillante de genacril B, amarillo brillante de genacril 10GF, rosa de genacril 3G, amarillo de genacril 5GF, GeneBlazer (CCF2), ácido gloxálico, azul granular, hematoporfirina, Hoechst 33258, Hoechst 33342, Hoechst 34580, HPTS, hidroxycumarina, hidroxiestilbamidina (FluoroGold), hidroxitriptamina, Indo-1, calcio alto, Indo-1, calcio bajo, indodicarbocianina (DiD), indotricarbocianina (DiR), Intrawhite Cf, JC-1, JO-JO-1, JO-PRO-1, LaserPro, Laurodan, LDS 751 (ADN), LDS 751 (ARN), Leucophor PAF, Leucophor SF, Leucophor WS, lisamina-rodamina, lisamina-rodamina B, homodímero de calceína/etidio, LOLO-1, LO-PRO-1, amarillo Lucifer, Lyso Tracker Blue, Lyso Tracker Blue-White, Lyso Tracker Green, Lyso Tracker Red, Lyso Tracker Yellow, LysoSensor Blue, LysoSensor Green, LysoSensor Yellow/Blue, Mag Green, rojo de Magdala (floxina B), Mag-Fura Red, Mag-Fura-2, Mag-Fura-5, Mag-Indo-1, verde de magnesio, naranja de magnesio, verde de malaquita, Marina Blue, Maxilon Brilliant Flavin 10 GFF, Maxilon Brilliant Flavin 8 GFF, merocianina, metoxicumarina, Mitotracker Green FM, Mitotracker Orange, Mitotracker Red, mitramicina, monobromobimano, monobromobimano (mBBR-GSH), monoclorobimano, MPS (Methyl Green Pyronine Stilbene), NBD, NBD-amina, rojo Nilo, nitrobenzoxadiazol, noradrenalina, Nuclear Fast Red, Nuclear Yellow, Nylosan Brilliant Iavin E8G, Oregon Green, verde Oregon 488-X, Oregon Green, verde Oregon 488, verde Oregon 500, verde Oregon 514, azul Pacífico, pararosanilina (Feulgen), PBF1, floxina B (rojo de Magdala), Phorwite AR, Phorwite BKL, Phorwite Rev, Phorwite RPA, fosfina 3R, PKH26 (Sigma), PKH67, PMIA, Pontochrome Blue Black, POPO-1, POPO-3, PO-PRO-1, PO-PRO-3, primulina, Procion Yellow, yoduro de propidio (PI), PyMPO, pireno, pironina, pironina B, Pyrozal Brilliant Flavin 7GF, QSY 7, mostaza de quinacrina, resorufina, RH 414, Rhod-2, rodamina, rodamina 110, rodamina 123, rodamina 5 GLD, rodamina 6G, rodamina B, rodamina B 200, rodamina B extra, rodamina BB, rodamina BG, verde de rodamina, rodamina-falocidina, rodamina-faloidina, rojo de rodamina, rodamina WT, rosa de Bengala, S65A, S65C, S65L, S65T, SBFI, serotonina, Sevron Brilliant Red 2B, Sevron Brilliant Red 4G, Sevron Brilliant Red B, Sevron Orange, Sevron Yellow L, SITS, SITS (primulina), SITS (ácido estilben-isotiosulfónico), SNAFL-calceína, SNAFL-1, SNAFL-2, SNARF-calceína, SNARF1, verde de sodio, SpectrumAqua, SpectrumGreen, SpectrumOrange,

SpectrumRed, SPQ (6-metoxi-N-(3-sulfopropil)quinolinio), estilbeno, sulforrodamina B can C, sulforrodamina Extra, SYTO 11, SYTO 12, SYTO 13, SYTO 14, SYTO 15, SYTO 16, SYTO 17, SYTO 18, SYTO 20, SYTO 21, SYTO 22, SYTO 23, SYTO 24, SYTO 25, SYTO 40, SYTO 41, SYTO 42, SYTO 43, SYTO 44, SYTO 45, SYTO 59, SYTO 60, SYTO 61, SYTO 62, SYTO 63, SYTO 64, SYTO 80, SYTO 81, SYTO 82, SYTO 83, SYTO 84, SYTO 85, SYTOX Blue, SYTOX Green, SYTOX Orange, tetraciclina, tetrametilrodamina (TAMRA), rojo Texas, conjugado de rojo Texas-X, tiadicarbocianina (DiSC3), rojo de tiazina R, naranja de tiazol, tioflavina 5, tioflavina S, tioflavina TCN, Thiolyte, naranja de tiazol, Tinopol CBS (blanco de calcoflúor), TMR, TO-PRO-1, TO-PRO-3, TO-PRO-5, TOTO-1, TOTO-3, TRITC (isotiocianato de tetrametilrodamina), TrueBlue, TruRed, Ultralite, Uranine B, Uvitex SFC, WW 781, X-rodamina, XRITC, naranja de xileno, Y66F, Y66H, Y66W, YO-PRO-1, YO-PRO-3, YOYO-1, YOYO-3, SYBR Green, naranja de tiazol (colorantes interquelantes), o combinaciones de los mismos.

Otro aspecto de la invención se refiere a un método implementado por ordenador. Por ejemplo, puede proporcionarse un ordenador y/o un sistema automatizado que es capaz de realizar de manera automática y/o repetitiva cualquiera de los métodos descritos en el presente documento. Tal como se usan en el presente documento, los dispositivos "automatizados" se refieren a dispositivos que son capaces de funcionar sin instrucciones humanas, es decir, un dispositivo automatizado puede realizar una función durante un periodo de tiempo después de que cualquier ser humano haya terminado de realizar cualquier acción para fomentar la función, por ejemplo, introduciendo instrucciones en un ordenador para iniciar el procedimiento. Normalmente, los equipos automatizados pueden realizar funciones repetitivas después de este momento. Las etapas de procesamiento también pueden grabarse en un medio legible por máquina en algunos casos.

Por ejemplo, en algunos casos, puede usarse un ordenador para controlar la obtención de imágenes de la muestra, por ejemplo, usando microscopía de fluorescencia, STORM u otras técnicas de superresolución tales como las descritas en el presente documento. En algunos casos, el ordenador también puede controlar operaciones tales como corrección de deriva, registro físico, hibridación y alineación de agrupamientos en análisis de imágenes, decodificación de agrupamientos (por ejemplo, decodificación de agrupamientos fluorescentes), detección o corrección de error (por ejemplo, tal como se comenta en el presente documento), reducción de ruido, identificación de características de primer plano con respecto a características de fondo (tales como ruido o residuos en imágenes), o similares. Como un ejemplo, el ordenador puede usarse para controlar la activación y/o excitación de entidades de señalización dentro de la muestra, y/o la adquisición de imágenes de las entidades de señalización. En un conjunto de realizaciones, puede excitarse una muestra usando luz que tiene diversas longitudes de onda y/o intensidades, y la secuencia de las longitudes de onda de luz usada para excitar la muestra puede correlacionarse, usando un ordenador, con las imágenes adquiridas de la muestra que contiene las entidades de señalización. Por ejemplo, el ordenador puede aplicar luz que tiene diversas longitudes de onda y/o intensidades a una muestra para producir diferentes números promedio de entidades de señalización en cada región de interés (por ejemplo, una entidad activada por cada ubicación, dos entidades activadas por cada ubicación, etc.). En algunos casos, esta información puede usarse para construir una imagen y/o determinar las ubicaciones de las entidades de señalización, en algunos casos a altas resoluciones, tal como se observó anteriormente.

En algunos aspectos, la muestra se coloca bajo un microscopio. En algunos casos, el microscopio puede contener uno o más canales, tales como canales microfluídicos, para dirigir o controlar el fluido hacia o desde la muestra. Por ejemplo, en una realización, las sondas de ácido nucleico tales como las comentadas en el presente documento pueden introducirse en y/o eliminarse a partir de la muestra haciendo fluir fluido a través de uno o más canales hacia o desde la muestra. En algunos casos, también puede haber una o más cámaras o depósitos para contener fluido, por ejemplo, en comunicación fluídica con el canal, y/o con la muestra. Los expertos habituales en la técnica estarán familiarizados con los canales, incluyendo canales microfluídicos, para mover fluido hacia o desde una muestra.

Tal como se usa en el presente documento, "microfluídico", "microscópico", "microescala", el prefijo "micro" (por ejemplo, como en "microcanal"), y similares se refieren generalmente a elementos o artículos que tienen anchuras o diámetros de menos de aproximadamente 1 mm, y menos de aproximadamente 100 micrómetros en algunos casos. En algunas realizaciones, pueden usarse canales más grandes en lugar de, o junto con, canales microfluídicos para cualquiera de las realizaciones comentadas en el presente documento. Por ejemplos, pueden usarse canales que tienen anchuras o diámetros de menos de aproximadamente 10 mm, menos de aproximadamente 9 mm, menos de aproximadamente 8 mm, menos de aproximadamente 7 mm, menos de aproximadamente 6 mm, menos de aproximadamente 5 mm, menos de aproximadamente 4 mm, menos de aproximadamente 3 mm, o menos de aproximadamente 2 mm en determinados casos. En algunos casos, el elemento o artículo incluye un canal a través del cual puede fluir un fluido. En todas las realizaciones, las anchuras especificadas pueden ser la anchura más pequeña (es decir, una anchura tal como se especifica donde, en esa ubicación, el artículo puede tener una mayor anchura en una dimensión diferente), o la anchura más grande (es decir, donde, en esa ubicación, el artículo tiene una anchura que no es más ancha que la anchura especificada, pero puede tener una longitud que es superior). Por tanto, por ejemplo, el canal microfluídico puede tener una dimensión de sección transversal promedio (por ejemplo, perpendicular a la dirección de flujo de fluido en el canal microfluídico) de menos de aproximadamente 1 mm, menos de aproximadamente 500 micrómetros, menos de aproximadamente 300 micrómetros, o menos de aproximadamente 100 micrómetros. En algunos casos, el canal microfluídico puede tener un diámetro promedio de menos de aproximadamente 60 micrómetros, menos de aproximadamente 50 micrómetros, menos de aproximadamente 40 micrómetros, menos de aproximadamente 30 micrómetros, menos de aproximadamente 25

micrómetros, menos de aproximadamente 10 micrómetros, menos de aproximadamente 5 micrómetros, menos de aproximadamente 3 micrómetros, o menos de aproximadamente 1 micrómetro.

Un "canal", tal como se usa en el presente documento, significa una característica sobre o en un artículo (por ejemplo, un sustrato) que dirige al menos parcialmente el flujo de un fluido. En algunos casos, el canal puede estar formado, al menos en parte, por un solo componente, por ejemplo, un sustrato grabado químicamente o una unidad moldeada. El canal puede tener cualquier forma de sección transversal, por ejemplo, circular, ovalada, triangular, irregular, cuadrada o rectangular (que tiene cualquier relación de aspecto), o similares, y puede estar cubierto o no cubierto (es decir, abierto al entorno externo que rodea el canal). En realizaciones en las que el canal está completamente cubierto, al menos una porción del canal puede tener una sección transversal que está completamente encerrada, y/o todo el canal puede estar completamente encerrado a lo largo de toda su longitud con la excepción de su entrada y salida.

Un canal puede tener cualquier relación de aspecto, por ejemplo, una relación de aspecto (longitud con respecto a dimensión de sección transversal promedio) de al menos aproximadamente 2:1, más normalmente de al menos aproximadamente 3:1, al menos aproximadamente 5:1, al menos aproximadamente 10:1, etc. Tal como se usa en el presente documento, una "dimensión de sección transversal", en referencia a un canal fluido o microfluido, se mide en una dirección generalmente perpendicular al flujo de fluido dentro del canal. Un canal incluirá generalmente características que facilitan el control sobre el transporte de fluido, por ejemplo, características estructurales y/o características físicas o químicas (hidrofobia frente a hidrofilia) y/u otras características que pueden ejercer una fuerza (por ejemplo, una fuerza de contención) sobre un fluido. El fluido dentro del canal puede llenar el canal de manera parcial o completa. En algunos casos, el fluido puede mantenerse o confinarse dentro del canal o una porción del canal de cierta manera, por ejemplo, usando tensión superficial (por ejemplo, de manera que el fluido se mantiene dentro del canal dentro de un menisco, tal como un menisco cóncavo o convexo). En un artículo o sustrato, parte (o la totalidad) de los canales pueden tener un tamaño particular o menor, por ejemplo, que tienen una dimensión más grande perpendicular al flujo de fluido de menos de aproximadamente 5 mm, menos de aproximadamente 2 mm, menos de aproximadamente 1 mm, menos de aproximadamente 500 micrómetros, menos de aproximadamente 200 micrómetros, menos de aproximadamente 100 micrómetros, menos de aproximadamente 60 micrómetros, menos de aproximadamente 50 micrómetros, menos de aproximadamente 40 micrómetros, menos de aproximadamente 30 micrómetros, menos de aproximadamente 25 micrómetros, menos de aproximadamente 10 micrómetros, menos de aproximadamente 3 micrómetros, menos de aproximadamente 1 micrómetro, menos de aproximadamente 300 nm, menos de aproximadamente 100 nm, menos de aproximadamente 30 nm, o menos de aproximadamente 10 nm o menor en algunos casos. En una realización, el canal es un capilar.

Pueden usarse una variedad de materiales y métodos, según determinados aspectos de la invención, para formar dispositivos o componentes que contienen canales microfluidicos, cámaras, etc. Por ejemplo, diversos dispositivos o componentes pueden estar formados por materiales sólidos, en los que los canales pueden formarse a través de micromecanizado, procedimientos de deposición de película tales como recubrimiento por centrifugación y deposición química en fase de vapor, deposición física en fase de vapor, fabricación por láser, técnicas fotolitográficas, métodos de grabado químico incluyendo procedimientos con plasma o químicos en húmedo, electrodeposición, y similares. Véase, por ejemplo, *Scientific American*, 248:44-55, 1983 (Angell, *et al.*).

En un conjunto de realizaciones, diversas estructuras o componentes pueden estar formados por un polímero, por ejemplo, un polímero elastomérico tal como polidimetilsiloxano ("PDMS"), politetrafluoroetileno ("PTFE" o Teflon®), o similares. Por ejemplo, según una realización, un canal tal como un canal microfluido puede implementarse fabricando el sistema fluido por separado usando PDMS u otras técnicas de litografía blanda (los detalles de las técnicas de litografía blanda adecuadas para esta realización se comentan en las referencias tituladas "Soft Lithography", de Younan Xia y George M. Whitesides, publicada en *Annual Review of Material Science*, 1998, vol. 28, páginas 153-184, y "Soft Lithography in Biology and Biochemistry", de George M. Whitesides, Emanuele Ostuni, Shuichi Takayama, Xingyu Jiang y Donald E. Ingber, publicada en *Annual Review of Biomedical Engineering*, 2001, vol. 3, páginas 335-373).

Otros ejemplos de polímeros posiblemente adecuados incluyen, pero no se limitan a, poli(tereftalato de etileno) (PET), poliacrilato, polimetacrilato, policarbonato, poliestireno, polietileno, polipropileno, poli(cloruro de vinilo), copolímero olefínico cíclico (COC), politetrafluoroetileno, un polímero fluorado, una silicona tal como polidimetilsiloxano, poli(cloruro de vinilideno), bis-benzociclobuteno ("BCB"), una poliimida, un derivado fluorado de una poliimida, o similares. También se prevén combinaciones, copolímeros, o mezclas que involucran polímeros incluyendo los descritos anteriormente. El dispositivo también puede estar formado por materiales compuestos, por ejemplo, un material compuesto de un polímero y un material semiconductor.

En algunas realizaciones, diversas estructuras o componentes microfluidicos del dispositivo están fabricados a partir de materiales poliméricos y/o flexibles y/o elastoméricos, y pueden estar formados convenientemente por un fluido endurecible, que facilita la fabricación a través de moldeo (por ejemplo, moldeo por réplica, moldeo por inyección, moldeo por colada, etc.). El fluido endurecible puede ser esencialmente cualquier fluido que pueda inducirse su solidificación, o que solidifique espontáneamente, para dar un sólido capaz de contener y/o transportar fluidos contemplados para su uso en y con la red fluidica. En una realización, el fluido endurecible comprende un líquido

polimérico o un precursor polimérico líquido (es decir, un “prepolímero”). Los líquidos poliméricos adecuados pueden incluir, por ejemplo, polímeros termoplásticos, polímeros termoestables, ceras, metales, o mezclas o materiales compuestos de los mismos calentados por encima de su punto de fusión. Como otro ejemplo, un líquido polimérico adecuado puede incluir una disolución de uno o más polímeros en un disolvente adecuado, disolución que forma un material polimérico sólido tras la eliminación del disolvente, por ejemplo, mediante evaporación. Los expertos habituales en la técnica también conocen bien tales materiales poliméricos que pueden solidificarse a partir de, por ejemplo, un estado fundido o mediante evaporación del disolvente. Son adecuados una variedad de materiales poliméricos, muchos de los cuales son elastoméricos, y también son adecuados para formar moldes o moldes matriz, para realizaciones en las que uno o ambos de los moldes matriz están compuestos por un material elastomérico. Una lista no limitativa de ejemplos de tales polímeros incluye polímeros de las clases generales de polímeros de silicona, polímeros epoxídicos, y polímeros de acrilato. Los polímeros epoxídicos se caracterizan por la presencia de un grupo éter cíclico de tres miembros habitualmente denominado grupo epoxi, 1,2-epóxido, u oxirano. Por ejemplo, pueden usarse diglicidil éteres de bisfenol A, además de los compuestos basados en amina aromática, triazina, y estructuras principales cicloalifáticas. Otro ejemplo incluye los polímeros novolacos bien conocidos. Los ejemplos no limitativos de elastómeros de silicona adecuados para su uso según la invención incluyen aquellos formados a partir de precursores incluyendo los clorosilanos tales como metilclorosilanos, etilclorosilanos, fenilclorosilanos, etc.

En determinadas realizaciones se usan polímeros de silicona, por ejemplo, el elastómero de silicona polidimetilsiloxano. Los ejemplos no limitativos de polímeros de PDMS incluyen aquellos vendidos con el nombre comercial Sylgard por Dow Chemical Co., Midland, MI, y particularmente Sylgard 182, Sylgard 184, y Sylgard 186. Los polímeros de silicona incluyendo el PDMS tienen varias propiedades beneficiosas que simplifican la fabricación de diversas estructuras de la invención. Por ejemplo, tales materiales son económicos, están disponibles fácilmente, y pueden solidificarse a partir de un líquido prepolimérico a través de curado con calor. Por ejemplo, los PDMS son normalmente curables por exposición del líquido prepolimérico a temperaturas de aproximadamente, por ejemplo, de aproximadamente 65 °C a aproximadamente 75 °C durante tiempos de exposición de, por ejemplo, al menos aproximadamente una hora. Además, los polímeros de silicona, tales como el PDMS, pueden ser elastoméricos y, por tanto, pueden ser útiles para formar características muy pequeñas con relaciones de aspecto relativamente altas, necesarias en determinadas realizaciones de la invención. A este respecto pueden ser ventajosos moldes matriz o moldes flexibles (por ejemplo, elastoméricos).

Una ventaja de formar estructuras tales como estructuras o canales microfluídicos a partir de polímeros de silicona, tales como PDMS, es la capacidad de oxidación de tales polímeros, por ejemplo, mediante exposición a un plasma que contiene oxígeno tal como un plasma de aire, de modo que las estructuras oxidadas contienen, en su superficie, grupos químicos capaces de reticularse con otras superficies de polímero de silicona oxidadas o con las superficies oxidadas de una variedad de otros materiales poliméricos y no poliméricos. Por tanto, las estructuras pueden fabricarse y luego oxidarse y sellarse de manera esencialmente irreversible a otras superficies de polímero de silicona, o a las superficies de otros sustratos reactivos con las superficies de polímero de silicona oxidadas, sin la necesidad de adhesivos u otros medios de sellado independientes. En la mayoría de los casos, el sellado puede completarse simplemente poniendo en contacto una superficie de silicona oxidada con otra superficie sin la necesidad de aplicar presión auxiliar para formar el sello. Es decir, la superficie de silicona previamente oxidada actúa como adhesivo de contacto frente a superficies de acoplamiento adecuadas. Específicamente, además de poder sellarse irreversiblemente a sí misma, la silicona oxidada tal como el PDMS oxidado también puede sellarse irreversiblemente a una gama de materiales oxidados distintos del mismo incluyendo, por ejemplo, vidrio, silicio, óxido de silicio, cuarzo, nitrato de silicio, polietileno, poliestireno, carbono vítreo, y polímeros epoxídicos, que se han oxidado de manera similar a la superficie de PDMS (por ejemplo, a través de la exposición a un plasma que contiene oxígeno). En la técnica se describen métodos de oxidación y sellado útiles en el contexto de la presente invención, así como técnicas de moldeo generales, por ejemplo, en un artículo titulado “Rapid Prototyping of Microfluidic Systems and Polydimethylsiloxane”, Anal. Chem., 70:474-480, 1998 (Duffy *et al.*).

Se tienen en cuenta los siguientes documentos: la patente estadounidense n.º 7.838.302, concedida el 23 de noviembre de 2010, titulada “Sub-Diffraction Limit Image Resolution and Other Imaging Techniques”, de Zhuang, *et al.*; la patente estadounidense n.º 8.564.792, concedida el 22 de octubre de 2013, titulada “Sub-diffraction Limit Image Resolution in Three Dimensions”, de Zhuang, *et al.*; y la publicación de solicitud de patente internacional n.º WO 2013/090360, publicada el 20 de junio de 2013, titulada “High Resolution Dual-Objective Microscopy”, de Zhuang, *et al.*

Además, se citan la solicitud de patente provisional estadounidense con n.º de serie 62/031.062, presentada el 30 de julio de 2014, titulada “Systems and Methods for Determining Nucleic Acids”, de Zhuang, *et al.*; la solicitud de patente provisional estadounidense con n.º de serie 62/050.636, presentada el 15 de septiembre de 2014, titulada “Probe Library Construction”, de Zhuang, *et al.*; la solicitud de patente provisional estadounidense con n.º de serie 62/142.653, presentada el 3 de abril de 2015, titulada “Systems and Methods for Determining Nucleic Acids”, de Zhuang, *et al.*; y una solicitud PCT presentada en una misma fecha con la misma, titulada “Probe Library Construction”, de Zhuang, *et al.*

EJEMPLO 1

El ejemplo presenta una plataforma para permitir la detección simultánea del número y la organización espacial de miles de ARNm distintos dentro de células individuales con alta eficiencia y baja tasa de error usando una forma nueva de hibridación fluorescente *in situ* altamente multiplexada (FISH). Este ejemplo logra estas mediciones mediante la integración y la innovación de métodos para la síntesis de sondas masivamente paralelas, la obtención de imágenes con superresolución, y códigos de comprobación de errores autocorregibles.

En este caso, estos ejemplos presentan métodos para la detección simultánea de parte o la totalidad de los miles de ARN únicos expresados en una célula. Este enfoque no sólo promete revolucionar el rendimiento del ya eficaz enfoque de FISH de una sola molécula (smFISH), sino que también permite a los investigadores beneficiarse del enfoque de descubrimiento sin hipótesis que ha hecho tan eficaces otros enfoques de la biología basados en sistemas de genoma completo. Por ejemplo, este enfoque basado en genoma completo puede permitir a los investigadores descubrir ARN cuyos niveles de expresión y/o patrones de localización subcelular cambian bajo determinadas condiciones de interés, tales como estados patológicos, sin saber, *a priori*, qué ARNm cambiará en cuanto a abundancia o localización. En algunos casos, las mediciones simultáneas de cientos de genes dentro de una única célula también permiten la identificación de correlaciones entre genes en cuanto a patrones de localización y expresión.

Esto puede lograrse usando métodos para smFISH altamente multiplexada a través de la hibridación secuencial de sondas de detección ortogonales y obtención de imágenes de superresolución, reduciendo el coste de síntesis de sonda, y el desarrollo de un sistema altamente automatizado para minimizar las demandas sobre el usuario, tal como se comenta en el presente documento. Esto proporciona una plataforma integrada para manejar la bioinformática de diseño de sondas, las matemáticas de códigos de corrección de error, la complejidad de registro y análisis de imágenes, y el engorroso manejo de fluidos a través de una sencilla suite de interfaces fáciles de usar. Esta integración permite un manejo sencillo con una formación limitada del usuario y facilita la rápida recopilación de datos.

Este ejemplo ilustra: (1) diseño computacional de "palabras de código" para su asociación con todas las dianas de ARN en la célula que permitirán la identificación singular de cada ARN con cierto grado de tolerancia de error experimental, (2) traducción de estas palabras de código a secuencias de nucleótidos y síntesis de las sondas de oligonucleótidos monocatenarios (mc) requeridas (por ejemplo, ADNmc), (3) fijación de muestra e hibridación de estas sondas con las dianas de ARN *in situ*, (4) lectura de estas palabras de código mediante rondas sucesivas de hibridación de distintas sondas fluorescentes cuyas imágenes se obtienen con microscopía de fluorescencia convencional o microscopía de fluorescencia de superresolución, y (5) decodificación automatizada de las palabras de código medidas en combinación con corrección computacional de errores para identificar de forma singular y robusta ARNm individuales.

En la primera etapa, se asigna una "palabra de código" a cada ARN que va a marcarse. En un diseño típico estas pueden ser cadenas de N posiciones o letras binarias. Las palabras de código pueden elegirse de la misma amplia gama de esquemas de codificación de corrección de errores o con tolerancia de errores existentes desarrollados para el almacenamiento y la comunicación digital, por ejemplo, usando códigos de Hamming o similares. Por ejemplo, a ARN actina se le puede asignar la palabra de código binaria 11001010. Cada palabra de código puede ser singular y estar separada de las demás palabras de código por una distancia de Hamming, h , que mide el número de letras o posiciones que deben leerse incorrectamente para que una palabra de código sea malinterpretada como otra diferente. Una distancia de Hamming mayor de 1 entre todas las palabras de código permite detectar algunos errores de medición, ya que los errores simples producirían palabras de código que no se usan para codificar ARN. Para una distancia de Hamming mayor de 2, también es posible corregir algunos errores, ya que las palabras de código con un error serán las más cercanas en distancia de Hamming a una sola palabra de código singular. La longitud de las palabras de código viene determinada por el número total de ARN diferentes que va a detectarse a partir del transcriptoma y la cantidad de corrección de errores deseada. La teoría de la información proporciona varios algoritmos eficientes para ensamblar libros de códigos binarios de corrección de errores.

En la segunda etapa, este esquema de codificación se traduce en un conjunto de secuencias de sonda de oligonucleótido (por ejemplo, ADN), que pueden denominarse sondas primarias o sondas de codificación, cada una de las cuales no sólo dirige una sonda al ARN de interés, sino que también codifica para la palabra de código binaria singular dentro de un conjunto de sitios de unión secundarios (figura 1). Por ejemplo, pueden diseñarse en primer lugar secuencias de unión primarias para cada ARNm diana. Estas secuencias son "secuencias diana" que están compuestas por secuencias de nucleótidos complementarias a sus ARNm diana seleccionadas computacionalmente para satisfacer un conjunto estricto de condiciones de hibridación, incluyendo la singularidad en el genoma diana. Para mejorar la eficiencia de hibridación con ARNm individuales, se diseñan múltiples secuencias diana primarias para cada ARN individual. A continuación, a cada posición dentro del conjunto de palabras de código se le asigna una secuencia de oligonucleótidos (por ejemplo, ADN) singular, que se denomina secuencia de lectura. Estas etiquetas están diseñadas para no interactuar con las secuencias de ARNm endógenas o entre sí. Por ejemplo, para todos los valores "1" en una palabra de código de un ARNm individual, la secuencia de lectura correspondiente está asociada a las secuencias de direccionamiento primarias contra ese ARNm. En general, cada sonda contendrá una secuencia diana y una o más secuencias de lectura. Si la longitud total de las secuencias de lectura necesarias y la

secuencia diana primaria excede las capacidades de síntesis, entonces pueden añadirse subconjuntos de las secuencias de lectura a secuencias diana distintas. Por ejemplo, considerar la posible palabra de código 11001010 para actina. Las secuencias de sonda para este ARN podrían contener las secuencias de lectura correspondientes a las posiciones 1, 2, 5, y 7 en la palabra de código asociada a una variedad de secuencias diana específicas de actina. Después de que todas las secuencias se hayan diseñado, el complejo conjunto resultante de secuencias de oligonucleótidos (por ejemplo, ADN) personalizadas singulares se fabrica y amplifica usando los métodos tal como se describen a continuación.

En la tercera etapa, la agrupación de ADN resultante se hibrida, por ejemplo, con células fijadas y permeabilizadas. En este procedimiento, pueden unirse sondas individuales a cada ARN en la célula mediante hibridación de sus secuencias diana correspondientes con el ARN mientras que las secuencias de lectura permanecen libres para unirse a las sondas secundarias apropiadas tal como se comenta a continuación.

En la cuarta etapa (la etapa de lectura) se hibridan sucesivamente sondas de ácido nucleico secundarias marcadas con fluorescencia (también denominadas sondas de lectura) con las secuencias de lectura unidas a las secuencias diana que se unen a las dianas de ARNm en la etapa anterior. Cuando se obtienen simultáneamente imágenes de un gran número de diferentes especies de ARN en las células, la densidad de los ARN marcados puede exceder aquella a la que cada ARN puede resolverse mediante métodos de obtención de imágenes convencionales. Por tanto, esto puede realizarse usando un método de obtención de imágenes de superresolución, por ejemplo, STORM (microscopía de reconstrucción óptica estocástica), para resolver las moléculas marcadas. Después de cada ronda de hibridación y obtención de imágenes con las sondas secundarias, los fluoróforos se extinguen o se inactivan de otro modo mediante técnicas o bien químicas o bien ópticas tales como oxidación, blanqueo químico, fotoblanqueo, lavado riguroso o digestión enzimática, etc. A continuación, la muestra se tiñe con la siguiente sonda secundaria, y el ciclo continúa hasta que se hayan leído todas las posiciones de las palabras de código. En la realización más simple, habrá una etapa de hibridación para cada posición dentro de la palabra de código, por ejemplo, 8 etapas de hibridación para una palabra de código de 8 letras (figura 1).

La figura 1 muestra diagramas esquemáticos de este ejemplo. La figura 1A muestra que a cada posición de las palabras de código se le asigna una secuencia de oligonucleótidos singular cuando esta posición tiene un valor "1". Todas las palabras de código de ARNm se traducen a continuación en combinaciones de secuencias de lectura que se unen a la secuencia de direccionamiento. La figura 1B muestra diversas etapas del esquema de marcaje de este ejemplo. En la primera etapa, todos los ARNm (I - III) se etiquetan con múltiples sondas de oligonucleótidos (por ejemplo, ADNmc) que comprenden una secuencia de direccionamiento primaria que se hibrida con el ARN de interés, y una "cola" (es decir, que contiene secuencias de lectura) que porta la palabra de código traducida, que no interactúa con secuencias de nucleótidos endógenas. En la siguiente etapa, se añade la primera sonda secundaria, que puede unirse a todas las sondas cuyas colas tengan una secuencia de lectura correspondiente al valor de "1" en la primera posición. Se obtienen imágenes de y se blanquean los colorantes de estas sondas secundarias, y a continuación se añade la siguiente sonda secundaria para unirse a las sondas unidas a ARNm que tengan un valor de "1" en la segunda posición de su palabra de código asignada, y así sucesivamente.

En la etapa final, las imágenes de microscopía de cada ronda de tinción y obtención de imágenes se alinean, por ejemplo, computacionalmente (por ejemplo, usando perlas fiduciales u otros marcadores rastreados durante la adquisición de imágenes), y se identifican los agrupamientos de localizaciones resueltos por microscopía convencional de fluorescencia u obtención de imágenes de superresolución (por ejemplo, STORM) de las diferentes rondas. Estos agrupamientos de localizaciones surgen de moléculas de ARNm diana individuales, y las rondas de hibridación en las que se detectó un punto en un agrupamiento dado corresponden al "1" en la palabra de código para ese ARNm. Si no hay acontecimientos de detección fallida o señales de falso positivo en las imágenes, esta palabra de código coincidirá perfectamente con una de las palabras de código esperadas. La figura 1 describe un ejemplo en el que la palabra de código tiene tres letras, es decir, tres posiciones, y los tres ARNm diana tienen palabras de código 110, 101, y 011 asignadas a los mismos. En ejemplos experimentales reales, la palabra de código podría contener más dígitos. Por ejemplo, al ARNm para actina se le puede asignar la palabra de código 11001010. En ese caso, los agrupamientos detectados que contengan señales de localización solapantes en las etapas de hibridación 1ª, 2ª, 5ª y 7ª (es decir, las sondas 1ª, 2ª, 5ª y 7ª secundarias unidas a este sitio) pueden identificarse como moléculas de ARNm de actina individuales, ya que el patrón de uniones positivas coincide con la palabra de código de actina (11001010). Además, si hay acontecimientos de detección fallida o señales de falso positivo en los datos de imagen, estas aberraciones pueden corregirse mediante el esquema de corrección de errores implementado. Por ejemplo, los agrupamientos de localizaciones con una palabra de código detectada que tiene una discrepancia de sólo un dígito de 11001010 (tal como 11000010 ó 11101010) también pueden identificarse como ARNm de actina ya que todas las demás palabras de código válidas en este ejemplo difieren del patrón detectado en dos posiciones más.

EJEMPLO 2

Este ejemplo describe otro enfoque alternativo que difiere en varias de las etapas descritas anteriormente. Este enfoque comienza con la primera etapa, la construcción de las palabras de código para las dianas de ARNm deseadas, tal como se describió anteriormente.

En la segunda etapa de este enfoque, se diseñan sondas de ácido nucleico que se unen de manera singular a las dianas de ARNm de interés, tal como se describió anteriormente. Sin embargo, en lugar de añadir secuencias de lectura singulares a estas secuencias de direccionamiento, se construyen agrupaciones o grupos de sondas singulares a partir de estas secuencias diana. Cada agrupación comprende la totalidad o un subconjunto de las secuencias que se dirigen a todos los ARNm que contienen el mismo valor en una posición dada en su palabra de código. Por ejemplo, la primera agrupación tendría la totalidad o un subconjunto de las secuencias diana diseñadas para todos los ARNm que contienen un 1 en la primera posición de sus palabras de código, por ejemplo, 110 y 101 pero no 011; la segunda agrupación tendría la totalidad o un subconjunto de las secuencias diana diseñadas para todos los ARNm que contienen un 1 en la segunda posición de sus palabras de código, por ejemplo, 110 y 011 pero no 101; la tercera agrupación tendría la totalidad o un subconjunto de las secuencias diana diseñadas para todos los ARNm que contienen un 1 en la tercera posición de sus palabras de código, por ejemplo, 011 y 101 pero no 110 (figura 1C). Como otro ejemplo, considerar la posible palabra de código 11001010 para actina. Las sondas que se dirigen a este ARNm se incluirían en las agrupaciones 1, 2, 5, y 7 pero no en las agrupaciones 3, 4, 6, y 8. La misma diana para un ARNm dado puede o no estar incluida en las agrupaciones. Por ejemplo, una sonda que se dirige a la misma región de actina puede estar incluida en las agrupaciones 1, 2, 5, y 7 o en cualquier subconjunto de estas agrupaciones. Después de que se hayan diseñado todas las agrupaciones, cada complejo conjunto de secuencias de oligonucleótidos personalizadas singulares se fabrica y amplifica usando los métodos tal como se describen a continuación.

En la tercera etapa de este enfoque, la primera agrupación de sondas se hibrida, por ejemplo, con células fijadas y permeabilizadas. En este procedimiento, los fluoróforos unidos a cada una de las sondas en esta agrupación se unen a cada una de las dianas de esa agrupación. La unión de estas sondas se determina entonces mediante microscopía de fluorescencia. Tal como se describió anteriormente, estas imágenes pueden obtenerse mediante una serie de métodos que incluyen tanto métodos de obtención de imágenes por fluorescencia convencionales como métodos de obtención de imágenes de superresolución tales como STORM. Después de una ronda de obtención de imágenes, las sondas de la primera agrupación se inactivan o se retiran de la muestra mediante los métodos descritos anteriormente. Luego se repite este procedimiento para cada agrupación sucesiva de sondas hasta que parte o la totalidad de las agrupaciones se hayan aplicado a la muestra y se hayan obtenido imágenes de manera que todas las posiciones en las palabras de código se hayan leído. En la realización más simple, habrá una etapa de hibridación y obtención de imágenes para cada posición en la palabra de código, por ejemplo, 3 rondas de hibridación y obtención de imágenes para una palabra de código con 3 posiciones (figura 1C) u 8 rondas de hibridación y obtención de imágenes para una palabra de código con 8 posiciones.

La etapa final de este enfoque es idéntica a la descrita anteriormente.

EJEMPLO 3

En este ejemplo, se codificaron 14 genes (PGK1, H3F3B, PKM, ENO1, GPI, EEF2, GNAS, HSPA8, GAPDH, CALM1, RHOA, PPIA, UBA52, y VCP) usando un subconjunto del código SECDED(8,4) (figuras 2A-2E). Para determinar la precisión de estas medidas, las abundancias medidas de estos 14 ARNm se compararon con las abundancias medidas a partir de secuenciación masiva de ARN (RNA-Seq) de células A549 (datos publicados de ENCODE). Sorprendentemente, se encontró que existía una excelente concordancia entre estas dos mediciones, ya que el recuento de transcritos medido usando el enfoque de hibridación secuencial se correlacionaba con la expresión génica medida usando RNA-seq con un coeficiente de correlación r de Pearson de 0,75 (figura 2F). También se midió la expresión génica de otras 3 células, y se observó que la expresión génica de estos 14 genes estaba altamente correlacionada entre las células con un r de 0,96 (figura 2G).

Diseño del libro de códigos. A cada ARNm del conjunto diana se le asignó una palabra de código binaria usando un código de corrección de error único y detección de error doble (SECDED). SECDED es un libro de códigos de Hamming extendido con un bit de paridad adicional. En resumen, se usó la caja de herramientas del sistema de comunicaciones de Matlab para generar códigos SECDED de o bien 8 o bien 16 letras o posiciones. En ambos casos, sólo se usaron las palabras de código que contenían cuatro 1. Estas palabras se asignaron aleatoriamente a ARNm en el conjunto diana. [0 1 0 1 1 1 0 0] es un ejemplo de las palabras de código de 8 letras usadas (es decir, estas palabras de código contenían cada una cuatro 1 y cuatro 0.) [0 1 0 1 1 1 0 0 0 0 0 0 0 0 0] es un ejemplo de las palabras de código de 16 letras usadas (es decir, cada palabra de código contenía cuatro 1 y doce 0). No todas las palabras de código se asignaron necesariamente a un ARNm.

Ensamblaje computacional de secuencia de sonda primaria de ADNmc. El número de sondas de ácido nucleico primarias usadas para la hibridación con dianas de ARNm osciló desde 200 hasta 2000 oligonucleótidos singulares, dependiendo del experimento. Por ejemplo, para marcar 14 ARNm con 28 oligonucleótidos que se dirigen a cada gen, se usaron 392 secuencias singulares. Se adquirió un gran número de oligonucleótidos con secuencias singulares en una agrupación de LC Sciences o CustomArray. Sin embargo, los oligonucleótidos sintetizados en matriz estaban en una cantidad mínima que era insuficiente para la hibridación *in situ*. El protocolo para su amplificación se describe a continuación.

Cada sonda primaria contenía tres componentes: secuencias de cebador flanqueantes para permitir la amplificación enzimática de sondas, secuencia de direccionamiento para hibridación *in situ* con ARNm, y secuencia de etiqueta secundaria que contenía una o más secuencias de lectura para la lectura secuencial de palabras de código.

5 Lo siguiente es un ejemplo de una sonda primaria:

GTGGCGACGAAAGCACTGCGATTGGAACCGTCCCAAGCGTTGCG
CTTAATGGATCATCAATTTTGTCTCACTACGACGGTCAATCGCGCTGCATACTTG
CGTCGGTCGGACAAACGAGG

(SEQ ID NO: 1)

- 10 Los componentes están dispuestos en el siguiente orden: cebador directo (no subrayado), secuencia de lectura secundaria 1 (subrayada), secuencia de direccionamiento a ARNm (no subrayada), secuencia de lectura secundaria 2 (subrayada), y cebador inverso (no subrayado). Las secuencias de lectura secundarias son el complemento inverso de las sondas secundarias correspondientes. Dado que sólo se usaron palabras de código que contenían cuatro "1", las sondas primarias para cada ARNm debían contener 4 secuencias de lectura secundarias diferentes en este ejemplo. Sin embargo, con el fin de reducir la longitud total de las sondas primarias, la agrupación de
- 15 secuencias de direccionamiento para cada diana de ARNm se dividió aleatoriamente en dos agrupaciones. Dos secuencias de lectura secundarias se unen a cada sonda en una de las dos agrupaciones y las otras dos secuencias de lectura secundarias se unen a las sondas en la otra agrupación. A continuación, se describen los criterios de diseño de cada componente.
- 20 Diseño de cebadores. Se generaron los cebadores de índice específicos a partir de una colección de 240.000 secuencias publicadas de secuencias ortogonales de 25 pb de longitud. Estas secuencias se recortaron a 20 pb, se seleccionaron para una temperatura de fusión estrecha de 70 a 80 °C, la ausencia de repeticiones consecutivas de 3 o más bases, y la presencia de una pinza GC, es decir, una de las dos bases en el extremo terminal 3' debe ser G o C. Para mejorar adicionalmente la especificidad, estas secuencias se cribaron a continuación con respecto al
- 25 genoma humano usando BLAST+ (Camacho *et al.* 2009), y se eliminaron los cebadores con 14 o más bases contiguas de homología. En un cribado posterior mediante BLAST+, también se eliminaron los cebadores que compartían 11 o más bases contiguas o más de 5 bases en el extremo 3' con cualquier otro cebador o con el promotor T7.
- 30 Diseño de sondas secundarias. Se crearon sondas secundarias de 30 pb de longitud concatenando fragmentos del conjunto de cebadores ortogonales descrito anteriormente. A continuación, se cribaron estas sondas secundarias para determinar su ortogonalidad con otras sondas secundarias (no más de 11 pares de bases de homología) y la presencia de posibles sitios de unión inespecíficos en el genoma humano (no más de 14 pares de bases de homología). En la tabla 1 se proporcionan las secuencias secundarias usadas en este ejemplo.
- 35

Tabla 1

<u>Bit</u>	<u>Secuencias secundarias</u>	<u>Número de secuencia</u>
B1	CGCAACGCTTGGGACGGTCCAATCGGATC	SEQ ID NO: 2
B2	CGAATGCTCTGGCCTCGAACGAACGATAGC	SEQ ID NO: 3
B3	ACAAATCCGACCAGATCGGACGATCATGGG	SEQ ID NO: 4
B4	CAAGTATGCAGCGCGATTGACCGTCTCGTT	SEQ ID NO: 5
B5	TGCGTCGTCTGGCTAGCACGGCACGCAAAT	SEQ ID NO: 6
B6	AAGTCGTACGCCGATGCGCAGCAATTCCT	SEQ ID NO: 7
B7	CGAAACATCGGCCACGGTCCCGTTGAACTT	SEQ ID NO: 8
B8	ACGAATCCACCGTCCAGCGCGTCAAACAGA	SEQ ID NO: 9
B9	CGCGAAATCCCCGTAACGAGCGTCCCTTGC	SEQ ID NO: 10
B10	GCATGAGTTGCCTGGCGTTGCGACGACTAA	SEQ ID NO: 11
B11	CCGTCTGTCTCCGGTCCACCGTTGCGCTTAC	SEQ ID NO: 12
B12	GGCCAATGGCCCAGGTCCGTCACGCAATTT	SEQ ID NO: 13
B13	TTGATCGAATCGGAGCGTAGCGGAATCTGC	SEQ ID NO: 14
B14	CGCGCGGATCCGCTTGTCGGGAACGGATAC	SEQ ID NO: 15
B15	GCCTCGATTACGACGGATGTAATTCGGCCG	SEQ ID NO: 16
B16	GCCCGTATTCCCGCTTGCGAGTAGGGCAAT	SEQ ID NO: 17

- 5 Diseño de secuencia de direccionamiento a ARNm. Para determinar la abundancia relativa de todas las isoformas de todos los genes expresados en estas líneas celulares, se procesaron los datos de perfil transcriptoma del proyecto ENCODE para el ARN total de las células A549 e IMR90 usando el software disponible públicamente *cufflinks*, junto con anotaciones del genoma humano de gencode v18. Se usaron modelos de genes correspondientes a la isoforma con la expresión más alta para construir una biblioteca de secuencias en formato FASTA que registrara la isoforma dominante de cada gen. Se seleccionaron genes de interés a partir de esta biblioteca. Estos genes se dividieron en segmentos de 1 kb, luego se usó el software OligoArray 2.1 para generar secuencias de sondas primarias para el transcriptoma humano con las siguientes restricciones: longitud de 30 pb o 40 pb, dependiendo del experimento; temperaturas de fusión de la sonda-diana superiores a 70 °C (parámetro variable); sin dianas de hibridación cruzada con temperaturas de fusión superiores a 72 °C (parámetro variable); sin estructuras secundarias internas predichas con temperaturas de fusión superiores a 76 °C (parámetro variable); y sin repeticiones contiguas de un solo nucleótido de 6 o más bases. Después de la selección de la sonda por OligoArray, se rechazaron todas las posibles sondas que se mapearon a un gen diferente, mientras que se retuvieron todas las posibles sondas con alineamientos múltiples con el mismo gen. Se creó una base de datos BLAST a partir de la biblioteca FASTA de todos los genes expresados para examinar la singularidad de las sondas. Para cada gen, se seleccionaron de 14 a 28 secuencias de direccionamiento producidas durante el procesamiento por OligoArray.

- Síntesis de sondas - PCR de índice. Se seleccionó el molde para conjuntos de sondas específicos de la agrupación de oligonucleótidos compleja mediante PCR de ciclo limitado. En resumen, se combinaron de 0,5 a 1 ng de la agrupación de oligonucleótidos compleja con 0,5 micromolar de cada cebador. El cebador directo correspondía a la secuencia de cebado para el subconjunto deseado, mientras que el cebador inverso era una concatenación 5' de esta secuencia con un promotor T7. Para evitar la generación de cuádruples de G, que pueden ser difíciles de sintetizar, las G terminales requeridas en el promotor T7 se generaron a partir de G situadas en el extremo 5' de la región de cebado cuando era apropiado. Todos los cebadores se sintetizaron mediante IDT. Se amplificó un volumen de reacción de 50 microlitros usando o bien el kit de amplificación de biblioteca en tiempo real KAPA (KAPA Biosystems; KK2701) o bien mediante una mezcla de qPCR casera que incluía EvaGreen 0,8X (Biotum; 31000-T) y

la polimerasa Phusion de arranque en caliente (New England Biolabs; M0535S). La amplificación se siguió en tiempo real usando el software MX300P de Aligent o CFX Connect de Biorad. Las muestras individuales se retiraron inmediatamente antes de la meseta de amplificación para minimizar la distorsión de la abundancia del molde debido a la sobreamplificación. Los moldes individuales se purificaron con columnas según las instrucciones del fabricante (Zymo DNA Clean and Concentrator; D4003) y se eluyeron en agua desionizada libre de ARNasa.

Amplificación mediante transcripción *in vitro*. Luego el molde se amplificó mediante transcripción *in vitro*. Brevemente, se amplificaron de 0,5 a 1 microgramo de ADN molde a 100-200 microgramos de ARN en una única reacción de 20-30 microlitros con una polimerasa ARN de alto rendimiento (New England Biolabs; E2040S). Las reacciones se complementaron con 1X inhibidor de ARNasa (Promega RNasin; N2611). La amplificación se realizó normalmente durante de 4 a 16 horas a 37 °C para maximizar el rendimiento. El ARN no se purificó después de la reacción y se almacenó a -80 °C o se convirtió inmediatamente en ADN tal como se describe a continuación.

Transcripción inversa. Se crearon de 1 a 2 nmol de sonda de ADNmc marcada con fluorescencia a partir de las reacciones de transcripción *in vitro* anteriores usando la transcriptasa inversa Maxima H- (Thermo Scientific; EP0751). Se usó esta enzima debido a su mayor capacidad de procedimiento y resistencia a la temperatura, lo que permitió la conversión de grandes cantidades de ARN en ADN dentro de pequeños volúmenes a temperaturas que desfavorecen la formación de estructura secundaria. El ARN no purificado creado anteriormente se complementó con 1,6 mM de cada dNTP, 1-2 nmol de cebador directo marcado con fluorescencia, 300 unidades de Maxima H-, 60 unidades de RNasin, y una concentración final 1X del tampón Maxima RT. El volumen final de 75 microlitros se incubó a 50 °C durante 60 minutos.

Selección y purificación de cadena. El ARN molde en la reacción anterior se retiró luego del ADN mediante hidrólisis alcalina. Se añadieron 75 microlitros de EDTA 0,25 M y NaOH 0,5 N a cada reacción de transcripción inversa, y la muestra se incubó a 95 °C durante 10 minutos. La reacción se neutralizó inmediatamente purificando la sonda de ADNmc con una versión modificada del protocolo Zymo Oligo Clean and Concentrator. Específicamente, la columna de 5 microgramos de capacidad se sustituyó por una columna de ADN de 25 microgramos o 100 microgramos de capacidad, según fuera apropiado. El resto del protocolo se ejecutó según las instrucciones del fabricante. La sonda se eluyó en 100 microlitros de agua desionizada libre de ARNasa y se evaporó en un concentrador de vacío. El sedimento final se resuspendió en 10 microlitros de agua libre de ARNasa y se almacenó a -20 °C. La electroforesis en gel de poliacrilimida desnaturalizante y la espectroscopia de absorción revelaron que este protocolo producía normalmente un 90-100 % de incorporación del cebador fluorescente en la sonda de longitud completa y un 75-90 % de recuperación de la sonda fluorescente total. Por tanto, sin exceder un volumen de reacción de 150 microlitros, este protocolo puede usarse para crear ~2 nmol de sonda fluorescente.

Cultivo y fijación celular. Las células A549 e IMR90 (Colección Americana de Cultivos Tipo) se cultivaron con medio de Eagle modificado por Dulbecco y medio esencial mínimo de Eagle, respectivamente. Las células se incubaron a 37 °C con el 5 % de CO₂ durante 36-48 horas. Las células se fijaron en paraformaldehído al 3 % (Electron Microscopy Sciences) en PBS durante 15 minutos, se lavaron con PBS, y se permeabilizaron en etanol al 70 % durante la noche a 4 °C.

Hibridación de fluorescencia *in situ* (FISH) - sondas primarias (de codificación). Las células se hidrataron en tampones de lavado (2xSSC, formamida al 50 %) durante 10 minutos, se marcaron con oligonucleótidos primarios (0,5 nM por secuencia) en tampones de hibridación (2xSSC, formamida al 50 %, ARNt de levadura 1 mg/ml, y sulfato de dextrano al 10 %) durante la noche a 37 °C, se lavaron con tampones de lavado a 47 °C durante 10 minutos dos veces, y se lavaron con 2xSSC dos veces. Se añadieron perlas fiduciales fluorescentes (Molecular Probes, F-8809) a una dilución 1:10.000 en 2xSSC antes de la obtención de imágenes.

Sondas secundarias. Las sondas secundarias (de lectura) (10 nM) se hibridaron en tampones de hibridación secundaria (2xSSC, formamida al 20 % y sulfato de dextrano al 10 %) con sus dianas primarias durante 30 minutos a 37 °C. Las células permanecieron en la platina del microscopio durante la hibridación. Se usó un calentador por objetivos para mantener la temperatura a 37 °C. Las células se lavaron con tampones de lavado secundarios (2xSSC, formamida al 20 %).

Fluídica y obtención de imágenes mediante STORM. Se realizaron múltiples rondas de marcaje secuencial, lavado, obtención de imágenes y blanqueo en una plataforma automatizada que consistía en un sistema fluídico y un microscopio de STORM (microscopía de reconstrucción óptica estocástica). El sistema fluídico incluía una cámara de flujo (Biopetech FCS2), una bomba peristáltica (Rainin Dynamax RP-1) y tres válvulas de 8 vías controladas por ordenador (Hamilton MVP y Hamilton HVXM 8-5). Este sistema permitió la integración automatizada de la colección de películas de STORM e hibridación secundaria.

El tampón de obtención de imágenes incluía Tris 50 mM (pH 8), glucosa al 10 % (p/v), β ME (2-mercaptoetanol) al 1 % o MEA 25 mM, con o sin 1,5-ciclooctadieno 2 mM, y un sistema eliminador de oxígeno (glucosa oxidasa 0,5 mg/ml (Sigma-Aldrich) y catalasa 40 microgramos/ml (Sigma-Aldrich)). Se usó una capa de aceite mineral para sellar el tampón de obtención de imágenes, evitando su acidificación a lo largo de múltiples hibridaciones.

El sistema de STORM incluía un microscopio invertido Olympus IX-71 configurado para excitación por incidencia oblicua. Las muestras se iluminaron continuamente con un láser de estado sólido bombeado por diodo de 642 nm (VFL-P500-642; MPB communications). Se usó para la activación de los colorantes un láser de estado sólido de 405 nm (Cube 405-100C; Coherent). La fluorescencia se recogió usando una lente objetiva Olympus (UPlanSApo 100x, 1,4 NA) y se hizo pasar a través de un filtro dicróico personalizado, así como de un divisor de haz de vista cuádruple. Todas las películas se grabaron usando una cámara EMCCD (Andor iAxiON 897), obteniendo las imágenes a 60 Hz. El campo de vista de 512x256 de la cámara se dividió en películas separadas de 256x256 píxeles antes de guardarlo. La mitad izquierda de este campo de vista contenía los datos de STORM y la mitad derecha las imágenes de las perlas fiduciales fluorescentes. Estas últimas películas se redujeron a 1 Hz antes de guardarse. Durante la adquisición de datos, se usó un bloqueo de enfoque casero para mantener un plano focal constante. Las películas de STORM incluían de 20.000 a 30.000 fotogramas en el tampón de STORM, mientras que las películas de blanqueo incluían 10.000 fotogramas en el tampón de lavado.

Análisis de imágenes - análisis de localizaciones de una sola molécula. Se procesaron por separado películas de localizaciones de una sola molécula y perlas fiduciales fluorescentes usando un software de localización de un solo emisor publicado previamente.

Registro de imágenes. La posición inicial de las perlas de cada ronda de hibridación se usó para alinear las películas de cada ronda. La autocorrelación 2D entre imágenes de perlas de hibridaciones consecutivas, seguida de la coincidencia del vecino más próximo, se usó para emparejar perlas entre imágenes. El par de perlas con el vector de desplazamiento más similar se usó para calcular una deformación rígida de traslación-rotación para alinear las perlas. Este método de alineación es robusto para muestras en las que se desplazan múltiples fiduciales o se sueltan y vuelven a unirse durante la obtención de imágenes.

Corrección de la deriva. La deriva durante la adquisición de imágenes se corrigió usando la trayectoria de las perlas fiduciales (registrada a 1 Hz). Las posiciones de las perlas se vincularon en cada trama. La trayectoria de las dos perlas que se movían de manera más correlacionada se tomó como la trayectoria de la deriva.

Identificación correcta de agrupamiento de ARNm. Las localizaciones se cribaron en primer lugar para estar por encima de un umbral de número de fotones (generalmente 2000) y debían estar dentro de 32 nm de otras 5 localizaciones (los parámetros pueden ajustarse). El resto de las localizaciones de moléculas se agruparon en un histograma 2D de 10x10 nm (el tamaño de grupo es un parámetro variable). Todos los grupos conectados se consideraron parte de un agrupamiento (los contactos diagonales se clasifican como conectados). Los agrupamientos debían tener más de 80 localizaciones totales en todas las hibridaciones (parámetro variable) para denominarse agrupamiento de ARNm. Los centroides ponderados de estos agrupamientos del histograma 2D se registraron como las posiciones de ARNm.

Un agrupamiento dado se registra para representarse en una ronda de hibridación individual si más de 9 localizaciones (parámetro variable) se encuentran dentro de un radio de 48 nm (parámetro variable) del centroide para ese ARNm en cada ronda de hibridación.

Decodificación de agrupamientos. Para cada agrupamiento de ARNm, se lee una palabra de código, que incluye "0" para todas las rondas de hibridación en las que se encuentra cerca del centroide un número de localizaciones inferior al umbral y "1" para las rondas en las que se cuenta un número de localizaciones superior al umbral. El libro de códigos SECEDED los descodificó como coincidencias perfectas con los códigos de palabra de ARNm diana, errores corregibles que pueden mapearse de vuelta inequívocamente al ARNm diana, o errores no corregibles que difieren en dos o más letras de las palabras del libro de códigos.

La figura 2A muestra una imagen de STORM de una célula. La figura 2B muestra un aumento en la región recuadrada en la figura 2A. Cada punto indica una localización. Las localizaciones de diferentes rondas de obtención de imágenes se muestran de manera diferente. La figura 2C muestra un agrupamiento de localizaciones representativo de la región recuadrada en la figura 2B. El agrupamiento muestra señales de localización de 4 hibridaciones diferentes. Este agrupamiento es un ARNm supuesto codificado con la palabra de código [0 1 0 1 1 1 0 0]. La figura 2D muestra una imagen de célula reconstruida de 14 genes después de la decodificación y corrección de errores. Cada gen se muestra de manera diferente. La figura 2E muestra expresión génica medida para los 14 genes de la célula. La figura 2F muestra una comparación de recuento de transcritos con datos de secuenciación ARN de Ensembl. La figura 2G muestra correlación de nivel de expresión de transcritos entre dos células detectadas usando el enfoque descrito.

EJEMPLO 4

Los siguientes ejemplos se refieren generalmente a obtención de imágenes de una sola molécula multiplexada con codificación robusta frente a errores que permite las mediciones simultáneas de miles de especies de ARN en células individuales. En general, el conocimiento del perfil de expresión y del paisaje espacial de los ARN en células individuales es esencial para comprender el rico repertorio de comportamientos celulares. Los siguientes ejemplos informan de diversas técnicas dirigidas a enfoques de obtención de imágenes de una sola molécula que permiten

determinar los números de copias y las localizaciones espaciales de miles de especies de ARN en células individuales. Algunas de estas técnicas se denominan hibridación fluorescente *in situ* robusta frente a errores multiplexada o “MERFISH”.

Usando esquemas de codificación robustos frente a errores para combatir los errores de marcaje y detección de una sola molécula, estos ejemplos demostraron la obtención de imágenes de cientos a miles de especies de ARN singulares en cientos de células individuales. El análisis de correlación de los $\sim 10^4$ a $\sim 10^6$ pares de genes permitió restringir las redes de regulación génica, predecir nuevas funciones para muchos genes no anotados e identificar distintos patrones de distribución espacial de ARN que se correlacionan con las propiedades de las proteínas codificadas.

Los análisis en todo el sistema de la abundancia y organización espacial de los ARN en células individuales prometen transformar la comprensión de muchas áreas de la biología celular y del desarrollo, tales como los mecanismos de regulación génica, el comportamiento heterogéneo de las células y el desarrollo y mantenimiento del destino celular. La hibridación fluorescente *in situ* de una sola molécula (smFISH) se ha convertido en una potente herramienta para estudiar el número de copias y la organización espacial de los ARN en células individuales, o bien en aislamiento o bien en su contexto tisular nativo. Aprovechando su capacidad para mapear las distribuciones espaciales de ARN específicos con alta resolución, smFISH ha revelado la importancia de la localización subcelular de ARN en diversos procedimientos tales como migración, desarrollo, y polarización celulares. Paralelamente, la capacidad de smFISH para medir con precisión los números de copias de ARN específicos sin sesgo de amplificación ha permitido la medición cuantitativa de las fluctuaciones naturales en la expresión génica, lo que a su vez ha dilucidado los mecanismos reguladores que dan forma a tales fluctuaciones y su papel en una variedad de procedimientos biológicos.

Sin embargo, la aplicación del enfoque de smFISH a muchas cuestiones a nivel de sistemas sigue estando limitada por el número de especies de ARN que pueden medirse simultáneamente en células individuales. Los esfuerzos más avanzados que usan el marcaje combinatorio o bien mediante códigos de barras basados en colores o bien mediante hibridación secuencial han permitido realizar mediciones simultáneas de 10-30 especies de ARN diferentes en células individuales, pero muchas cuestiones biológicas interesantes se beneficiarían de la medición de cientos a miles de ARN dentro de una sola célula, lo que no es posible de lograr usando tales técnicas. Por ejemplo, el análisis de cómo el perfil de expresión de un número tan grande de ARN varía entre células y cómo estas variaciones se correlacionan entre diferentes genes podría usarse para identificar sistemáticamente genes regulados conjuntamente y mapear redes reguladoras; el conocimiento de las organizaciones subcelulares de numerosos ARN y sus correlaciones podría ayudar a dilucidar los mecanismos moleculares que subyacen al establecimiento y mantenimiento de muchas estructuras celulares locales; y la obtención del perfil de ARN de células individuales en tejidos nativos podría permitir la identificación *in situ* del tipo de célula.

Los siguientes ejemplos generalmente comentan determinadas técnicas denominadas MERFISH, que son métodos de obtención de imágenes mediante smFISH altamente multiplexada que aumentan sustancialmente el número de especies de ARN de las que pueden obtenerse imágenes simultáneamente en células individuales usando marcaje combinatorio y obtención de imágenes secuencial con esquemas de codificación robustos frente a errores. Estos ejemplos demuestran este enfoque de obtención de imágenes multiplexada midiendo simultáneamente 140 especies de ARN usando un esquema de codificación que puede tanto detectar como corregir errores y 1001 especies de ARN usando un esquema de codificación que puede detectar, pero no corregir errores. Debe entenderse que estos números son sólo a modo de ejemplificación, no de limitación. Los análisis de correlación de las variaciones del número de copias y las distribuciones espaciales de estos genes permitieron identificar grupos de genes que están regulados conjuntamente y grupos de genes que comparten patrones similares de distribución espacial dentro de la célula.

Marcaje combinatorio con esquemas de codificación robustos frente a errores. El marcaje combinatorio que identifica cada especie de ARN mediante múltiples (N) señales distintas ofrece una vía para aumentar rápidamente el número de especies de ARN que pueden investigarse simultáneamente en células individuales (figura 5A). Sin embargo, este enfoque para aumentar el rendimiento de smFISH a escala de los sistemas se enfrenta a un desafío importante porque no sólo el número de especies de ARN abordables aumenta exponencialmente con N , sino que las tasas de errores de detección también crecen exponencialmente con N (figuras 5B-5D). Considérese un esquema conceptualmente simple para implementar marcaje combinatorio donde cada especie de ARN se codifica con una palabra binaria de N bits y la muestra se investiga con N rondas de hibridación correspondientes, dirigiendo cada ronda sólo al subconjunto de ARN que deberían ser “1” en el bit correspondiente (figura 11). N rondas de hibridación permitirían investigar $2^N - 1$ especies de ARN. Con sólo 16 hibridaciones, podrían identificarse más de 64.000 especies de ARN, que deberían cubrir todo el transcriptoma humano incluyendo tanto ARN mensajeros (ARNm) como ARN no codificantes (figura 5B; símbolos superiores). Sin embargo, a medida que N aumenta, la fracción de ARN detectados correctamente (la tasa de identificaciones correctas) disminuiría rápidamente y, lo que es más preocupante, la fracción de ARN que se identifican como especies incorrectas (la tasa de identificaciones erróneas) aumentaría rápidamente (figura 5C, símbolos inferiores; figura 5D, símbolos superiores). Con tasas de errores por hibridación realistas (medidas a continuación), la mayoría de las moléculas de ARN se identificarían erróneamente después de 16 rondas de hibridación.

Para abordar este desafío, se diseñaron esquemas de codificación robustos frente a errores, en los que sólo un subconjunto de las 2^N-1 palabras separadas por una cierta distancia de Hamming se usaban para codificar ARN. En un libro de códigos donde la distancia de Hamming mínima es 4 (código HD4), deben leerse incorrectamente al menos cuatro bits para cambiar una palabra de código por otra (figura 12A). Como resultado, cada error de un solo bit produce una palabra que se aproxima de manera singular a una sola palabra de código, lo que permite detectar y corregir tales errores (figura 12B). Los errores de doble bit producen palabras con una distancia de Hamming igual de 2 de múltiples palabras de código y, por tanto, pueden detectarse, pero no corregirse (figura 12C). Un código de este tipo debería aumentar sustancialmente la tasa de identificaciones correctas y reducir la tasa de identificaciones erróneas (figura 5C y 5D, símbolos centrales). Para tener en cuenta adicionalmente el hecho de que es más probable pasar por alto un acontecimiento de hibridación (un error $1 \rightarrow 0$) que identificar erróneamente un punto de fondo como un ARN (un error $0 \rightarrow 1$) en mediciones de smFISH, se diseñó un código HD4 modificado (MHD4), en el que el número de bits "1" se mantuvo constante y relativamente bajo, sólo cuatro por palabra, para reducir el error y evitar una detección sesgada. Este código MHD4 debería aumentar adicionalmente la tasa de identificaciones correctas y reducir la tasa de identificaciones erróneas (figura 5C, símbolos superiores; figura 5D, símbolos inferiores).

Además de las consideraciones de error, varios desafíos prácticos también han dificultado investigar un gran número de especies de ARN, tales como el alto coste del enorme número de sondas de FISH marcadas con fluorescencia necesarias y el largo tiempo requerido para completar muchas rondas de hibridación. Para superar estos desafíos, en este ejemplo, se diseñó un esquema de marcaje de dos etapas para codificar y leer ARN celulares (figura 5E). En primer lugar, los ARN celulares se marcaron con un conjunto de sondas de codificación (también denominadas sondas primarias), comprendiendo cada sonda una secuencia de direccionamiento a ARN y dos secuencias de lectura flanqueantes. Se asignaron cuatro de las N secuencias de lectura singulares a cada especie de ARN basándose en la palabra de código MHD4 del ARN. En segundo lugar, estas N secuencias de lectura se identificaron con sondas de FISH complementarias, las sondas de lectura (también denominadas sondas secundarias) mediante N rondas de hibridación y obtención de imágenes, usando cada ronda una sonda de lectura singular. Para aumentar la relación señal-fondo, cada ARN celular se marcó con ~ 192 sondas de codificación. Dado que cada sonda de codificación contenía dos de las cuatro secuencias de lectura asociadas con ese ARN (figura 5E), un máximo de ~ 96 sondas de lectura pueden unirse a cada ARN celular por ronda de hibridación. Para generar el enorme número de sondas de codificación necesarias, se amplificaron a partir de agrupaciones de oligonucleótidos derivadas de matrices que contenían decenas de miles de secuencias personalizadas usando un procedimiento de amplificación enzimática que comprende la transcripción *in vitro* seguida de la transcripción inversa (figura 13, véase a continuación en relación con la síntesis de sondas). Este enfoque de marcaje de dos etapas disminuyó significativamente el tiempo de hibridación total para un experimento: se encontró que una hibridación eficaz con las secuencias de lectura sólo llevó 15 minutos mientras que una hibridación eficaz directa con ARN celular requirió más de 10 horas.

La figura 5 describe MERFISH, un enfoque de smFISH altamente multiplexada que usa marcaje combinatorio y codificación robusta frente a errores. La figura 5A muestra una representación esquemática de la identificación de múltiples especies de ARN en N rondas de obtención de imágenes. Cada especie de ARN se codifica con una palabra binaria de N bits y durante cada ronda de obtención de imágenes, sólo el subconjunto de ARN que debe leerse "1" en el bit correspondiente emite señal. Las figuras 5B-5D muestran el número de especies de ARN abordables (figura 5B), la tasa a la que estos ARN se identifican correctamente (tasa de identificaciones correctas) (figura 5C) y la tasa a la que los ARN se identifican incorrectamente como especies de ARN diferentes (tasa de identificaciones erróneas) (figura 5D) representadas gráficamente en función del número de bits (N) en las palabras binarias que codifican para el ARN. En las figuras 5B y 5D, los puntos superiores son un código binario simple que incluye todas las 2^N-1 palabras binarias posibles; los puntos centrales son el código HD4 donde la distancia de Hamming que separa las palabras es 4; y los puntos inferiores son el código HD4 modificado (MHD4) donde el número de bits "1" se mantiene en cuatro. Estos se invierten en la figura 5C.

Las tasas de identificaciones correctas y de identificaciones erróneas se calculan con tasas de error por bit del 10 % para el error $1 \rightarrow 0$ y del 4 % para el error $0 \rightarrow 1$. La figura 5E es un diagrama esquemático de la implementación de un código MHD4 para la identificación de ARN. Cada especie de ARN se marca en primer lugar con ~ 192 sondas de codificación que convierten el ARN en una combinación singular de secuencias de lectura (hib. de codificación). Estas sondas de codificación contienen, cada una, una región central de direccionamiento a ARN flanqueada por dos secuencias de lectura, extraídas de una agrupación de N secuencias diferentes, cada una asociada a una ronda de hibridación específica. Las sondas de codificación para una especie específica de ARN contienen una combinación singular de cuatro de las N secuencias de lectura, que corresponden a las cuatro rondas de hibridación en las que este ARN debería leerse "1". Se usaron N rondas posteriores de hibridación con las sondas fluorescentes de lectura para investigar las secuencias de lectura (hib. 1, hib. 2, ..., hib. N). Las sondas unidas se inactivaron mediante fotoblanqueo entre rondas sucesivas de hibridación. Por claridad, sólo se representa en este caso un posible apareamiento de las secuencias de lectura para las sondas de codificación; sin embargo, todos los posibles pares de las cuatro secuencias de lectura se usan a la misma frecuencia y se distribuyen de manera aleatoria a lo largo de cada ARN celular en los experimentos reales.

La figura 11 muestra una descripción esquemática de un enfoque de marcaje combinatorio basado en un código binario simple. En un enfoque de marcaje conceptualmente simple, pueden codificarse de manera singular 2^N-1 especies de ARN diferentes con todas las palabras binarias de N bits (excluyendo la palabra con todos "0"). En cada ronda de hibridación, se incluyen sondas de FISH que se dirigen a todas las especies de ARN que tienen un "1" en el bit correspondiente. Para aumentar la capacidad de discriminar los puntos de ARN del fondo, se aborda cada ARN con múltiples sondas de FISH por ronda de hibridación. La señal de las sondas unidas se extingue antes de la siguiente ronda de hibridación. Este procedimiento continúa para todas las N rondas de hibridación (hib. 1, hib. 2,...), y todas las 2^N-1 especies de ARN pueden identificarse mediante el patrón de encendido-apagado único de señales de fluorescencia en cada ronda de hibridación.

La figura 12 muestra descripciones esquemáticas de distancia de Hamming y su uso en la identificación y corrección de errores. La figura 12A es una representación esquemática de una distancia de Hamming de 4. Las figuras 12B y 12C son esquemas que muestran la capacidad de un esquema de codificación con distancia de Hamming de 4 para corregir errores de un solo bit (figura 12B) o detectar, pero no corregir, errores de doble bit (figura 12C). Las flechas resaltan los bits en los que difieren las palabras indicadas. Dos palabras de código están separadas por una distancia de Hamming de 4 si una de las palabras tiene que cambiar cuatro bits de "1" a "0" o de "0" a "1" para convertirse en la otra palabra. La corrección de error de un solo bit es posible porque si una palabra medida difiere de una palabra de código legítima en un solo bit, lo más probable es que se trate de un error derivado de una lectura errónea de esta palabra de código, ya que las palabras de código de todas las demás especies de ARN diferirán de la palabra medida en al menos tres bits. En este caso, la palabra medida puede corregirse a una palabra de código que difiera sólo en un bit. Si una palabra medida difiere de una palabra de código legítima en dos bits, esta palabra medida todavía puede identificarse como un error, pero la corrección ya no es posible puesto que más de una palabra de código legítima difiere de esta palabra medida en dos bits.

La figura 13 muestra la producción de la biblioteca de sondas de codificación. Una agrupación de oligonucleótidos compleja sintetizada por matriz que contiene ~100k secuencias se usa como molde para la amplificación enzimática de las sondas de codificación para diferentes experimentos. Cada secuencia molde en la agrupación de oligonucleótidos contiene una región diana central que puede unirse a un ARN celular, dos secuencias de lectura flanqueantes y dos cebadores de índice flanqueantes. En la primera etapa, se seleccionan las moléculas molde necesarias para un experimento específico y se amplifican con una reacción PCR indexada. Para permitir la amplificación mediante transcripción *in vitro*, se añade un promotor T7 a los productos de la PCR durante esta etapa. En la segunda etapa, el ARN se amplifica a partir de estas moléculas molde mediante transcripción *in vitro*. En la tercera etapa, este ARN se somete a transcripción inversa a ADN. En la última etapa, el ARN molde se retira mediante hidrólisis alcalina, dejando sólo las sondas de ADNmc deseadas. Este protocolo produce ~2 nmol de agrupaciones complejas de sondas de codificación que contienen ~20.000 secuencias diferentes para los experimentos de 140 genes o ~100.000 secuencias diferentes para los experimentos de 1001 genes.

EJEMPLO 5

Este ejemplo ilustra la medición de 140 genes con MERFISH usando un código MHD4 de 16 bits. Para someter a prueba la viabilidad de este enfoque de obtención de imágenes multiplexada robusto frente a errores, este ejemplo usa una medición de 140 genes en células de fibroblastos humanos (IMR90) usando un código MHD4 de 16 bits para codificar 130 especies de ARN y dejando 10 palabras de código como controles de identificaciones erróneas (figura 20). Después de cada ronda de hibridación con las sondas de lectura fluorescentes, se obtuvieron las imágenes de las células mediante la obtención de imágenes de campo amplio convencional con una geometría de iluminación de incidencia oblicua. Los puntos fluorescentes correspondientes a los ARN individuales se detectaron claramente y luego se extinguieron eficientemente mediante una breve etapa de fotoblanqueo (figura 6A). La muestra se mantuvo estable a lo largo de las 16 rondas de marcaje y obtención de imágenes iterativos. El cambio en el número de puntos fluorescentes de una ronda a otra coincidió con el cambio previsto basado en las abundancias relativas de las especies de ARN diana en cada ronda derivadas de la secuenciación masiva, y no se observó una tendencia sistemática a la disminución con el aumento del número de rondas de hibridación (figura 14A). El brillo promedio de los puntos varió de una ronda a otra con una desviación estándar del 40 %, probablemente debido a las diferentes eficacias de unión de las sondas de lectura a las diferentes secuencias de lectura en las sondas de codificación (figura 14B). Sólo se observó una pequeña tendencia sistemática a la disminución del brillo de los puntos con el aumento de las rondas de hibridación, que fue en promedio del 4 % por ronda (figura 14B).

A continuación, se construyeron palabras binarias a partir de los puntos fluorescentes observados basándose en sus patrones de encendido-apagado a lo largo de las 16 rondas de hibridación (figuras 6B-6D). Si la palabra coincidía exactamente con una de las 140 palabras del código MHD4 (coincidencias exactas) o difería sólo en un bit (coincidencias con errores corregibles), se asignaba a la especie de ARN correspondiente (figura 6D). Dentro de la única célula representada en las figuras 6A y 6B, más de 1500 moléculas de ARN correspondientes al 87 % de las 130 especies de ARN codificadas se detectaron después de la corrección de error (figura 6E). Se hicieron observaciones similares en ~400 células a partir de 7 experimentos independientes. En promedio, se detectaron ~4 veces más moléculas de ARN y ~2 veces más especies de ARN por célula después de la corrección de error en comparación con los valores obtenidos antes de la corrección de error (figura 15).

Pueden producirse dos tipos de errores en la medición del número de copias de cada especie de ARN:

1) Algunas moléculas de esta especie de ARN no se detectan, lo que conduce a un descenso en la tasa de identificaciones correctas, y

2) algunas moléculas de otras especies de ARN se identifican erróneamente como esta especie de ARN. Para evaluar el grado de identificaciones erróneas, se utilizaron las 10 palabras de control de identificaciones erróneas, es decir, palabras de código que no estaban asociadas con ningún ARN celular. Aunque se observaron coincidencias con estas palabras de control, se produjeron con mucha menor frecuencia que con las palabras de codificación de ARN reales: el 95 % de las 130 palabras de codificación de ARN se contaron con más frecuencia que la mediana del recuento de estas palabras de control. Además, normalmente se observó que la razón del número de coincidencias exactas con respecto al número de coincidencias con errores de un bit para una palabra de codificación de ARN real era sustancialmente mayor que las mismas razones observadas para los controles de identificaciones erróneas, tal como era de esperar (figuras 16A y 16B). Usando esta razón como medida de la confianza en la identificación de ARN, se encontró que el 91 % de las 130 especies después de la corrección de error, es decir, se decodificaron correctamente el ~80 % de los puntos fluorescentes correspondientes a una especie de ARN (figura 16E). Cabe señalar que, aunque el 20 % restante de puntos contribuyó a una pérdida de eficacia en la detección, la mayoría de ellos no provocaron una identificación errónea de las especies porque se decodificaron como palabras de error de doble bit y se descartaron.

Para estimar la tasa de identificaciones correctas, se utilizó la capacidad de corrección de error del código MHD4 para determinar las tasas de error $1- >0$ (10 % en promedio) y tasas de error $0- >1$ (4 % en promedio) para cada ronda de hibridación (figuras 16C y 16D). Usando estas tasas de error, se estimó una tasa de identificaciones correctas de ~80 % para especies de ARN individuales después de la corrección de error, es decir, se decodificaron correctamente el ~80 % de los puntos fluorescentes correspondientes a una especie de ARN (figura 16E). Cabe señalar que, aunque el 20 % restante de puntos contribuyó a una pérdida de eficacia en la detección, la mayoría de ellos no provocaron una identificación errónea de las especies porque se decodificaron como palabras de error de doble bit y se descartaron.

Para someter a prueba el posible sesgo técnico de estas mediciones, se investigaron las mismas 130 especies de ARN con un libro de códigos MHD4 diferente permutando al azar las palabras de código entre las diferentes especies de ARN (figura 20) y cambiando las secuencias de sonda de codificación. Las mediciones con este código alternativo dieron tasas de identificaciones correctas y de identificaciones erróneas similares (figura 17). Los números de copias de especies de ARN individuales por célula medidos con estos dos libros de código mostraron una excelente concordancia con un coeficiente de correlación de Pearson de 0,94 (figura 6G), lo que indica que la elección del esquema de codificación no sesgó los recuentos medidos.

Con el fin de validar los números de copias derivados de los experimentos de MERFISH, se realizaron mediciones de smFISH convencionales en 15 de los 130 genes, seleccionados del intervalo de abundancia completo medido de tres órdenes de magnitud. Para cada uno de estos genes, tanto el número de copias promedio como la distribución del número de copias en muchas células coincidieron cuantitativamente entre las mediciones de MERFISH y smFISH convencional (figuras 18A y 18B). La razón de los números de copias determinada por estos dos enfoques fue de $0,82 \pm 0,06$ (media $\pm$ EEM en las 15 especies de ARN medidas, figura 18B), que concordaba con la tasa de identificaciones correctas estimada del 80 % para el enfoque de obtención de imágenes multiplexada. La concordancia cuantitativa entre esta razón y la tasa de identificaciones correctas estimada a lo largo del intervalo de abundancia completo medido apoya además la evaluación de que el error de identificación errónea fue bajo. Dado que la concordancia entre los resultados de MERFISH y smFISH convencional se extendió a los genes con la menor abundancia medida (<1 copia por célula, figura 18B), se estimó que la sensibilidad de medición era de al menos 1 copia por célula.

Como validación final, la abundancia de cada especie de ARN promediada en cientos de células se comparó con las obtenidas a partir de una medición de secuenciación de ARN masiva realizada en la misma línea celular. Los resultados de obtención de imágenes se correlacionaron notablemente bien con los resultados de secuenciación masiva con un coeficiente de correlación de Pearson de 0,89 (figura 6H).

La figura 6 muestra la medición simultánea de 140 especies de ARN en células individuales usando MERFISH con un código MHD4 de 16 bits. La figura 6A muestra imágenes de moléculas de ARN en una célula IMR90 después de cada ronda de hibridación (hib. 1 - hib. 16). La imagen después del fotoblanqueo (blanqueo 1) demostró la retirada eficaz de las señales fluorescentes entre hibridaciones. La figura 6B muestra las localizaciones de todas las moléculas individuales detectadas en esta célula coloreadas basándose en sus palabras binarias medidas. Parte insertada: la imagen fluorescente compuesta de las 16 rondas de hibridación para la subregión recuadrada con círculos numerados que indican moléculas potenciales de ARN. Un círculo indica una molécula no identificable, cuya palabra binaria no coincide con ninguna de las palabras de código MHD4 de 16 bits, incluso después de la corrección de error. La figura 6C muestra imágenes fluorescentes de cada ronda de hibridación para la subregión recuadrada de la figura 6B, indicando círculos las moléculas potenciales de ARN. La figura 6D muestra las palabras correspondientes a los puntos identificados en la figura 6C. Las cruces representan los bits corregidos. La figura 6E muestra el número de copias de ARN para cada gen observado sin (menor) o con (mayor) corrección de error en esta célula. La figura 6F muestra la razón de confianza medida para las 130 especies de ARN (izquierda) y las 10

palabras de control de identificaciones erróneas (derecha) normalizada al valor máximo observado de los controles de identificaciones erróneas (línea discontinua). La figura 6G es un gráfico de dispersión del número de copias promedio de cada especie de ARN por célula medido con dos libros de códigos permutados al azar del código MHD4. El coeficiente de correlación de Pearson es de 0,94 con un valor de p de 1×10^{-53} . La línea discontinua corresponde a la línea $y = x$. La figura 6H es un gráfico de dispersión del número de copias promedio de cada especie de ARN por célula frente a la abundancia determinada por secuenciación masiva en fragmentos por kilobase por millón de lecturas (FPKM). El coeficiente de correlación de Pearson entre las abundancias logarítmicas de las dos mediciones fue de 0,89 con un valor de p de 3×10^{-39} .

La figura 14 muestra el número y brillo promedio de los puntos fluorescentes detectados en las 16 rondas de hibridación antes y después del fotoblanqueo. La figura 14A muestra el número de puntos fluorescentes observados por célula antes (mayor) y después (menor) del fotoblanqueo como función de la ronda de hibridación promediada en todas las mediciones con el primer código MHD4 de 16 bits. El fotoblanqueo reduce el número de puntos fluorescentes en dos o más órdenes de magnitud. Las rondas de hibridación sin barras inferiores representan rondas en las que no se observaron moléculas tras el blanqueo. También se representa el cambio esperado en el número de puntos fluorescentes de una ronda a otra (círculos), predicho basándose en las abundancias relativas de las especies de ARN diana en cada ronda de hibridación derivadas de la secuenciación de ARN masiva. La discrepancia promedio entre el número de puntos observado y previsto para cada hibridación es sólo del 15 % del número medio de puntos. Esta discrepancia no aumenta sistemáticamente con el número de rondas de hibridación. La figura 14B muestra el brillo promedio de los puntos fluorescentes identificados en cada ronda de hibridación promediado en todas las mediciones con el primer código MHD4 de 16 bits antes (superior) y después (inferior) del fotoblanqueo. El brillo varía en un 40 % (desviación estándar) en diferentes rondas de hibridación. El patrón de variación es reproducible entre experimentos con el mismo código, probablemente debido a diferencias en la eficacia de unión de las sondas de lectura a las diferentes secuencias de lectura. Existe una pequeña tendencia sistemática de disminución del brillo con el aumento de rondas de hibridación, que es en promedio un 4 % por ronda. El potoblanqueo extingue la fluorescencia a un nivel similar al de la autofluorescencia de la célula.

La figura 15 muestra que la corrección de error aumenta sustancialmente los números de moléculas de ARN y especies de ARN detectadas en células individuales. La figura 15A muestra un histograma de la razón del número total de moléculas detectadas por célula con corrección de error con respecto al número medido sin corrección de error. La figura 15B es un histograma del número total de especies de ARN detectadas en cada célula con corrección de error con respecto al número sin corrección de error. Ambas razones se determinan para ~200 células y los histogramas se construyen a partir de estas razones.

La figura 16 muestra la caracterización de las tasas de identificaciones erróneas y de identificaciones correctas de especies de ARN para los experimentos de 140 genes usando un código MHD4 específico de 16 bits. La figura 16A muestra el número de palabras medidas que coinciden exactamente con la palabra de código correspondiente a FLNA, representado por la barra en el centro del círculo, y el número de palabras medidas con error de un bit en comparación con la palabra de código de FLNA, representado por las 16 barras en el círculo. La figura 16B es igual que la figura 16A, pero para una palabra de código que no se asignó a ningún ARN, es decir, una palabra de control de identificaciones erróneas. Las líneas continuas conectan la coincidencia exacta con las palabras con error de un bit que se generan por errores $1 \rightarrow 0$. Basándose en la observación de que la razón del número de coincidencias exactas con respecto al número de coincidencias con errores corregibles para una palabra de codificación de ARN real era normalmente sustancialmente mayor que las mismas razones observadas para los controles de identificaciones erróneas, esta razón se definió como una razón de confianza para la identificación de ARN. La razón de confianza medida para todas las 130 especies de ARN (barra del centro) y 10 palabras de control de identificaciones erróneas no asignadas a ningún ARN (barras exteriores) usando este código MHD4 de 16 bits se muestra en la figura 6F. Las figuras 16C y 16D muestran las tasas de error promedio para el error $1 \rightarrow 0$ (figura 16C) y el error $0 \rightarrow 1$ (figura 16D) para cada ronda de hibridación. La figura 16E muestra la tasa de identificaciones correctas para cada especie de ARN estimadas a partir de las tasas de errores $1 \rightarrow 0$ y $0 \rightarrow 1$. Los genes se ordenan de izquierda a derecha basándose en la abundancia medida, que abarca tres órdenes de magnitud. Las tasas de identificaciones correctas son en gran medida independientes de la abundancia del gen.

La figura 17 muestra la caracterización de las tasas de identificaciones erróneas y de identificaciones correctas para un segundo código MHD4 de 16 bits. En este segundo esquema de codificación, las 140 palabras de código se permutaron al azar entre diferentes especies de ARN y se cambiaron las secuencias de sonda de codificación. La figura 17A muestra la tasa de confianza normalizada medida para las 130 especies de ARN (izquierda) y las 10 palabras de control de identificaciones erróneas no asignadas a ningún ARN (derecha). La tasa de confianza normalizada se determina del mismo modo que en la figura 6F. Las figuras 17B y 17C muestran las tasas de error promedio determinadas para el error $1 \rightarrow 0$ (figura 17B) y el error $0 \rightarrow 1$ (figura 17C) para cada ronda de hibridación. La figura 17D muestra la tasa de identificaciones correctas determinada para cada especie de ARN estimada a partir de las tasas de errores $1 \rightarrow 0$ y $0 \rightarrow 1$. Los genes se ordenan de izquierda a derecha basándose en la abundancia medida.

La figura 18 muestra una comparación de las mediciones de MERFISH con resultados de smFISH convencional para un subconjunto de genes. La figura 18A muestra las distribuciones de números de copias ARN en células

individuales para tres genes de ejemplo KIAA1199, DYNC1H1, y LMTK2 en los intervalos de abundancia alto, medio, y bajo, respectivamente. Barras más claras: distribuciones construidas a partir de ~400 células en las mediciones de 140 genes usando los códigos MHD4. Barras más oscuras: distribuciones construidas a partir de ~100 células en mediciones de smFISH convencional. La figura 18B muestra una comparación de los números de copias de ARN promedio por célula medidos en los experimentos de 140 genes usando los códigos MHD4 con respecto a los determinados mediante smFISH convencional para 15 genes. La razón promedio del número de copias medido usando las mediciones de MHD4 con respecto al número medido usando smFISH convencional fue de 0,82 +/- 0,06 (media +/- EEM en los 15 genes). La línea discontinua corresponde a la línea $y = x$.

La figura 20 muestra dos libros de códigos diferentes para los experimentos de 140 genes. Las palabras de código específicas del código MHD4 de 16 bits asignadas a cada especie de ARN estudiada en las dos permutaciones al azar del experimento de 140 genes. Las columnas "Genes" contienen el nombre del gen. Las columnas "Palabras de código" contienen la palabra binaria específica asignada a cada gen.

EJEMPLO 6

Este ejemplo se refiere generalmente al análisis de alto rendimiento de variación entre células en expresión génica. El enfoque de MERFISH permite la paralelización de mediciones de muchas especies de ARN individuales y análisis de covariación entre diferentes especies de ARN. En este ejemplo, el aspecto de la paralelización se ilustró en primer lugar examinando la variación entre células en el nivel de expresión de cada uno de los genes medidos (figura 7A). Para cuantificar la variación medida, los factores de Fano, definidos como la razón de la varianza con respecto al número de copias de ARN medio se calcularon para todas las especies de ARN medidas. Los factores de Fano se desviaron sustancialmente de 1, el valor esperado para un procedimiento de Poisson simple, para muchos genes y mostraron una tendencia en aumento con la abundancia de ARN media (figura 7B). Esta tendencia de aumento de factores de Fano con la abundancia de ARN media puede explicarse por los cambios en la tasa de transcripción y/o las tasas de desactivación del promotor, pero no por los cambios en la tasa de activación del promotor.

Además, se identificaron varias especies de ARN con factores de Fano sustancialmente mayores que esta tendencia media. Por ejemplo, se encontró que SLC5A3, CENPF, MKI67, TNC y KIAA1199 mostraban valores de factor de Fano sustancialmente superiores a los de los otros genes expresados a niveles de abundancia similares. La alta variabilidad de algunos de estos genes puede explicarse por su asociación con el ciclo celular. Por ejemplo, dos de estos genes particularmente "ruidosos", MKI67 y CENPF, se anotaron ambos como genes relacionados con el ciclo celular, y basándose en su expresión bimodal (figura 7C), se propone que su transcripción está fuertemente regulada por el ciclo celular. Otros genes de alta variabilidad no mostraron los mismos patrones de expresión bimodal y no se sabe que estén asociados con el ciclo celular.

El análisis de covariaciones en los niveles de expresión de diferentes genes puede revelar qué genes están regulados conjuntamente y elucidar rutas de regulación génica. A nivel de población, tal análisis a menudo requiere la aplicación de estímulos externos para impulsar la variación de la expresión génica; por tanto, pueden observarse cambios correlacionados de expresión entre genes que comparten elementos reguladores comunes influidos por los estímulos. A nivel de una sola célula, pueden aprovecharse las fluctuaciones estocásticas naturales de la expresión génica para este tipo de análisis y, por tanto, pueden estudiarse múltiples redes reguladoras sin tener que estimular cada una de ellas individualmente. Análisis de covariación de este tipo puede restringir las redes reguladoras, sugerir nuevas rutas reguladoras y predecir la función de genes no anotados basándose en asociaciones con genes covariantes.

Este enfoque se aplicó a las mediciones de 140 genes y se examinaron los ~10.000 coeficientes de correlación por pares que describen cómo variaron conjuntamente los niveles de expresión de cada par de genes entre células. Muchos de los genes altamente variables mostraron variaciones estrechamente correlacionadas o anticorrelacionadas (figura 7C). Para comprender mejor las correlaciones de todos los pares de genes, se adoptó un enfoque de agrupamiento jerárquico para organizar estos genes basándose en sus coeficientes de correlación (figura 7D). A partir de la estructura del árbol de agrupamientos, se identificaron siete grupos de genes con patrones de expresión sustancialmente correlacionados (figura 7D). Dentro de cada uno de los siete grupos, cada gen mostró una correlación promedio significativamente más fuerte con otros miembros del grupo que con genes fuera del grupo. Para validar y comprender adicionalmente estos grupos, se identificaron los términos de la ontología génica (GO) enriquecidos en cada uno de estos siete grupos. Notablemente, los términos de GO enriquecidos dentro de cada grupo compartían funciones similares y eran en gran medida únicos de cada grupo (figura 7E), validando la noción de que la covariación observada en la expresión refleja algunos puntos en común en la regulación de estos genes.

Este ejemplo describe dos de estos grupos como ejemplos ilustrativos. Los términos de GO predominantes asociados con el grupo 1 fueron términos asociados con la matriz extracelular (ECM) (figuras 7D y 7E). Los miembros destacados de este grupo incluyeron componentes de ECM, tales como FBN1, FBN2, COL5A, COL7A y TNC, y glucoproteínas que unen la ECM y las membranas celulares, tales como VCAN y THBS1. El grupo también incluía un gen no anotado, KIAA1199, que se prevé que desempeñe un papel en el metabolismo de la ECM.

basándose en su asociación con este agrupamiento. De hecho, este gen se ha identificado recientemente como una enzima implicada en la regulación del hialuronano, un importante componente de azúcar de la ECM.

El grupo 6 contenía muchos genes que codifican para proteínas de transporte de vesículas y proteínas asociadas con la motilidad celular (figuras 7D y 7E). Los genes de transporte de vesículas incluyen motores microtubulares y genes relacionados DYNC1H, CKAP1, y factores asociados con la formación y el tráfico de vesículas, tales como DNAJC13 y RAB3B. De nuevo, se encontró un gen no anotado, KIAA1462, dentro de este agrupamiento. Basándose en su fuerte correlación con DYNC1H1 y DNAJC13, se predice que este gen puede estar implicado en el transporte de vesículas. Los genes de motilidad celular de este grupo incluyen proteínas de unión a actina tales como AFAP1, SPTAN1, SPTBN1, y MYH10, y genes implicados en la formación de complejos de adhesión, tales como FLNA y FLNC. Varios factores asociados a GTPasa implicados en la regulación de la motilidad, adhesión y contracción celulares también se encuentran en este grupo, tales como DOCK7, ROCK2, IQGAP1, PRKCA, y AMOTL1. La observación de que algunos genes de motilidad celular se correlacionan con genes de transporte de vesículas es coherente con el papel del transporte de vesículas en la migración celular. Una característica adicional interesante del grupo 6 es que un subconjunto de estos genes, en particular los relacionados con la motilidad celular, estaban anticorrelacionados con miembros del grupo de ECM comentado anteriormente (figura 7D). Esta anticorrelación puede reflejar interacciones reguladoras que median en la activación de células entre estados adherentes y migratorios.

La figura 7 muestra las variaciones entre células y las correlaciones por pares para las especies de ARN determinadas a partir de las mediciones de 140 genes. La figura 7A muestra una comparación de los niveles de expresión génica en dos células individuales. La figura 7B muestra los factores de Fano para genes individuales. Las barras de error representan el error estándar de la media determinada a partir de 7 conjuntos de datos independientes. La figura 7C muestra las puntuaciones Z de las variaciones de expresión de cuatro pares de genes de ejemplo que muestran una variación correlacionada (los dos primeros) o anticorrelacionada (los dos últimos) para 100 células seleccionadas al azar. La puntuación Z se define como la diferencia de la media normalizada por la desviación estándar. La figura 7D es una matriz de los coeficientes de correlación por pares de la variación de expresión entre células para los genes medidos, mostrada junto con el árbol de agrupamiento jerárquico. Los siete grupos identificados por un umbral específico en el árbol de agrupamiento (línea discontinua) se indican mediante los recuadros negros de la matriz y las líneas en el árbol, indicando las líneas grises en el árbol los genes no agrupados. Pueden elegirse distintos umbrales en el árbol de agrupamientos para seleccionar subgrupos más pequeños con correlaciones más estrechas o supergrupos más grandes que contengan subgrupos más débilmente acoplados. Dos de los siete grupos se amplían a la derecha. La figura 7E muestra el enriquecimiento de 30 términos de GO seleccionados y estadísticamente significativamente enriquecidos en los siete grupos. El enriquecimiento se refiere a la razón de la fracción de genes dentro de un grupo que tienen el término de GO específico con respecto a la fracción de todos los genes medidos que tienen ese término. No todos los términos de GO presentados en este caso están en la lista de los 10 primeros.

EJEMPLO 7

Este ejemplo ilustra las distribuciones espaciales de mapeo de los ARN. Como enfoque basado en la obtención de imágenes, MERFISH también permitió la investigación de las distribuciones espaciales de muchas especies de ARN simultáneamente. De la inspección visual de genes individuales surgieron varios patrones, con algunos transcritos de ARN enriquecidos en la región perinuclear, otros enriquecidos en la periferia celular y otros dispersos por toda la célula (figura 8A). Para identificar genes con distribuciones espaciales similares, se determinaron los coeficientes de correlación para los perfiles de densidad espacial de todos los pares de especies de ARN, y estos ARN se organizaron basándose en las correlaciones por pares usando de nuevo el enfoque de agrupamiento jerárquico. La matriz de coeficientes de correlación mostró grupos de genes con organizaciones espaciales correlacionadas, y los dos grupos más notables con las correlaciones más fuertes se indican en la figura 8B. Los ARN del grupo I parecían estar enriquecidos en la región perinuclear, mientras que los ARN del grupo II parecían estar enriquecidos cerca de la región periférica de la célula (figura 8C). El análisis cuantitativo de las distancias entre cada molécula de ARN y el núcleo celular o la periferia celular confirmó efectivamente esta impresión visual (figura 8D).

El grupo I contenía genes que codificaban para proteínas extracelulares tales como FBN1, FBN2 y THSBI, proteínas secretadas tales como PAPPA, y proteínas integrales de la membrana tales como LRPI y GPR107. Estas proteínas no presentan similitudes evidentes en cuanto a su función. En cambio, un análisis de GO mostró un enriquecimiento significativo para términos de localización, tales como región extracelular, membrana basal, o espacio perivitelino (figura 8E). Para alcanzar estas localizaciones, las proteínas deben pasar por la ruta de secreción, que a menudo requiere la traducción del ARNm en el retículo endoplásmico (RE). Por tanto, se propone que el patrón espacial que se observó para estos ARNm refleja su enriquecimiento cotraduccional en el RE. El enriquecimiento de estos ARNm en la región perinuclear (figuras 8C y 8D, sombreado más claro), donde reside el RE rugoso, apoya esta conclusión.

El grupo II contenía genes que codificaban para proteínas de unión a actina, incluyendo filaminas FLNA y FLNC, talina TLN1, y espectrinas SPTAN1 y SPTBN1; la proteína de unión a microtúbulos CKAP5; y las proteínas motoras MYH10 y DYNC1H1. Este grupo estaba enriquecido con términos de GO tales como citoesqueleto cortical de actina, unión de filamentos de actina y unión adherente célula-célula (figura 8E). El ARNm de beta-actina puede estar

enriquecido cerca de la periferia celular en fibroblastos, al igual que los ARNm que codifican para miembros del complejo Arp2/3 de unión a actina. El enriquecimiento de los ARNm del grupo II en la región periférica de las células (figuras 8C y 8D) sugiere que la distribución espacial de los genes del grupo II podría estar relacionada con la distribución de los ARNm del citoesqueleto de actina.

La figura 8 muestra distintas distribuciones espaciales de ARN observadas en las mediciones de 140 genes. La figura 8A muestra ejemplos de las distribuciones espaciales observadas para cuatro especies diferentes de ARN en una célula. La figura 8B es una matriz de los coeficientes de correlación por pares que describe el grado de correlación de las distribuciones espaciales de cada par de genes, mostrada junto con el árbol de agrupamiento jerárquico. Los recuadros negros de la matriz y el sombreado del árbol indican dos grupos fuertemente correlacionados. La figura 8C muestra las distribuciones espaciales de todos los ARN de los dos grupos en dos células de ejemplo. Símbolos más claros: genes del grupo I; símbolos más oscuros: genes del grupo II. La figura 8D muestra las distancias promedio de los genes del grupo I y los genes del grupo II al borde de la célula o al núcleo normalizadas con respecto a las distancias promedio de todos los genes. Las barras de error representan el EEM en 7 conjuntos de datos. La figura 8E muestra el enriquecimiento de términos de GO en cada uno de los dos grupos.

EJEMPLO 8

Este ejemplo ilustra la medición de 1001 genes con un código MHD2 de 14 bits. Este ejemplo aumenta adicionalmente el rendimiento de las mediciones de MERFISH obteniendo simultáneamente imágenes de ~1000 especies de ARN. Este aumento podría lograrse con el código MHD4 aumentando el número de bits por palabra de código a 32 mientras se mantiene el número de bits "1" por palabra en cuatro (figura 5B). Mientras que la estabilidad de las muestras a través de muchas rondas de hibridación (figura 14) sugiere que tal extensión es potencialmente factible, en este caso se muestra un enfoque alternativo que no requiere un aumento en el número de hibridaciones relajando el requisito de corrección de error, pero manteniendo la capacidad de detección de error. Por ejemplo, reduciendo la distancia de Hamming desde 4 hasta 2, podrían usarse todas las palabras de 14 bits que contienen cuatro bits "1" para codificar 1001 genes y estos ARN se investigaron con sólo 14 rondas de hibridación. Sin embargo, debido a que un solo error puede producir una palabra igualmente parecida a dos palabras de código diferentes, la corrección de errores ya no es posible para este código de distancia de Hamming-2 modificada (MHD2). Por tanto, se esperaba que la tasa de identificaciones correctas fuera más baja y la tasa de identificaciones erróneas más alta con este esquema de codificación.

Para evaluar el rendimiento de este código MHD2 de 14 bits, se apartaron 16 de las 1001 palabras de código posibles como controles de identificaciones erróneas y se usaron las 985 palabras restantes para codificar para ARN celulares. Entre estos 985 ARN se incluyeron 107 especies de ARN investigadas en los experimentos de 140 genes como control adicional. Los experimentos de 1001 genes se realizaron en células IMR90 usando un procedimiento similar al descrito anteriormente. Para permitir la síntesis de todas las sondas de codificación a partir de una sola agrupación de oligonucleótidos de 100.000 miembros, el número de sondas de codificación por especie de ARN se redujo hasta ~94. Volvieron a detectarse puntos fluorescentes correspondientes a moléculas individuales de ARN en cada ronda de hibridación con las sondas de lectura y, basándose en sus patrones de encendido-apagado, estos puntos se decodificaron en ARN (figuras 9A, 19A y 19B). Se detectaron 430 especies de ARN en la célula mostrada en la figura 9A, y se obtuvieron resultados similares en ~200 células de las cuales se obtuvieron imágenes en 3 experimentos independientes.

Tal como se esperaba, la tasa de identificaciones erróneas de este esquema fue superior a la del código MHD4. El 77 % de todas las palabras reales de ARN se detectaron con mayor frecuencia que el recuento mediano para los controles de identificaciones erróneas, en lugar del valor del 95 % observado en las mediciones de MHD4. Usando el mismo análisis de razón de confianza descrito anteriormente, se descubrió que el 73 % (en lugar del 91 % de las mediciones de MHD4) de las 985 especies de ARN se midieron con una razón de confianza mayor que el valor máximo observado para los controles de identificaciones erróneas (figura 19C). Los números de copias de ARN medidos a partir de este 73 % de especies de ARN mostraron una excelente correlación con los resultados de la secuenciación de ARN masiva (coeficiente de correlación r de Pearson = 0,76; figura 9B, negro). Cabe señalar que el 27 % restante de los genes todavía muestran una buena, aunque menor, correlación con los datos de secuenciación de ARN masiva (r = 0,65; figura 9B, rojo), pero se tomó la medida conservadora de excluirlos de análisis posteriores.

La falta de una capacidad de corrección de errores también disminuyó la tasa de identificaciones correctas de cada especie de ARN: Al comparar las 107 especies de ARN comunes en las mediciones de 1001 genes y 140 genes, se observó que el número de copias por célula de estas especies de ARN era menor en las mediciones de 1001 genes (figuras 9C y 19D). El recuento total de estos ARN por célula fue ~1/3 del observado en las mediciones de 140 genes. Por tanto, la falta de corrección de errores en el código MHD2 produjo una disminución de ~3 veces en la tasa de identificaciones correctas, que es coherente con la disminución de ~4 veces en la tasa de identificaciones correctas observada para el código MHD4 cuando no se aplicó la corrección de errores. Tal como se esperaba a partir de la concordancia cuantitativa entre las mediciones de 140 genes y los resultados de smFISH convencional, la comparación de las mediciones de 1001 genes con los resultados de smFISH convencional para 10 especies de ARN también indicó una disminución de ~3 veces en la tasa de identificaciones correctas (figura 18C). A pesar de la

reducción esperada en la tasa de identificaciones correctas, las buenas correlaciones encontradas entre los números de copias observados en las mediciones de 1001 genes y los observados en las mediciones de 140 genes, así como en las mediciones de smFISH convencional y de secuenciación de ARN masiva, indican que la abundancia relativa de estos ARN puede cuantificarse con el esquema de codificación de MHD2.

La obtención de imágenes simultánea de ~1000 genes en células individuales amplió sustancialmente la capacidad de detectar genes coregulados. La figura 10A muestra la matriz de coeficientes de correlación por pares determinada a partir de las variaciones entre células en los niveles de expresión de estos genes. Usando el mismo análisis de agrupamiento jerárquico descrito anteriormente, se identificaron ~100 grupos de genes con expresión correlacionada. Notablemente, casi todos estos ~100 grupos mostraron un enriquecimiento estadísticamente significativo de términos de GO relacionados funcionalmente (figura 10B). Éstos incluían algunos de los grupos identificados en las mediciones de 140 genes, tales como el grupo asociado con genes de replicación celular y el grupo asociado con genes de motilidad celular (figuras 10A y 10B, grupos 7 y 102), así como muchos grupos nuevos. Los grupos identificados en este caso incluían 46 especies de ARN que carecían de cualquier anotación de GO previa, para las cuales puede hipotetizarse una función basada en su asociación a grupos. Por ejemplo, KIAA1462 es parte del grupo de motilidad celular, tal como también se muestra en los experimentos de 140 genes, lo que sugiere un papel potencial de este gen en la motilidad celular (figura 10A, grupo 102). Asimismo, KIAA0355 forma parte de un nuevo grupo enriquecido en genes asociados al desarrollo cardíaco (figura 10A, grupo 79), y C17orf70 forma parte de un grupo asociado al procesamiento del ARN ribosómico (figura 10A, grupo 22). Usando estos agrupamientos, pueden hipotetizarse funciones celulares para 61 factores de transcripción y otras proteínas parcialmente anotadas de funciones desconocidas. Por ejemplo, los factores de transcripción Z3CH13 y CHD8 son ambos miembros del grupo de motilidad celular, lo que sugiere su papel potencial en la regulación transcripcional de los genes de motilidad celular.

La figura 9 muestra mediciones simultáneas de 1001 genes en células individuales usando MERFISH con un código MHD2 de 14 bits. La figura 9A muestra las localizaciones de todas las moléculas individuales detectadas en una célula coloreadas basándose en sus palabras binarias medidas. Parte insertada: la imagen fluorescente compuesta y coloreada en falso de las 14 rondas de hibridación para la subregión recuadrada con círculos numerados que indican moléculas potenciales de ARN. Los círculos indican moléculas no identificables, cuyas palabras binarias no coinciden con ninguna de las palabras de código MHD2 de 14 bits. En la figura 19A se muestran imágenes de la ronda de hibridación individual. La figura 9B es un gráfico de dispersión del número de copias promedio por célula medido en los experimentos de 1001 genes frente a la abundancia medida mediante secuenciación masiva. Los símbolos superiores corresponden al 73 % de los genes detectados con razones de confianza superiores a la razón máxima observada para los controles de identificaciones erróneas. El coeficiente de correlación de Pearson es de 0,76 con un valor de p de 3×10^{-133} . Los símbolos inferiores corresponden al 27 % restante de genes. El coeficiente de correlación de Pearson es de 0,65 con un valor de p de 3×10^{-33} . La figura 9C es un gráfico de dispersión del número de copias promedio de los 107 genes compartidos tanto en la medición de 1001 genes con el código MHD2 como en la medición de 140 genes con el código MHD4. El coeficiente de correlación de Pearson es de 0,89 con un valor de p de 9×10^{-30} . La línea discontinua corresponde a la línea $y = x$.

La figura 10 muestra el análisis de covariación de las especies de ARN medidas en las mediciones de 1001 genes. La figura 10A es una matriz de todos los coeficientes de correlación por pares de la variación de expresión entre células para los genes medidos mostrados con el árbol de agrupamiento jerárquico. Los ~100 grupos identificados de genes correlacionados se indican mediante sombreado en el árbol. A la derecha se muestra un aumento de cuatro de los grupos descritos en el texto. La figura 10B es un enriquecimiento de 20 términos de GO seleccionados, enriquecidos de manera estadísticamente significativa en los cuatro grupos.

La figura 19 muestra la decodificación y la evaluación de errores de los experimentos de 1001 genes. La figura 19A muestra imágenes de la subregión recuadrada de la célula en la figura 9A para cada una de las 14 rondas de hibridación. El panel final es una imagen compuesta de estas 14 rondas. Los círculos indican puntos fluorescentes que se han identificado como posibles moléculas de ARN. Algunos círculos en la imagen compuesta indican moléculas no identificables, cuyas palabras binarias no coinciden con ninguna de las palabras de código MHD2 de 14 bits. La figura 9B muestra la palabra binaria correspondiente para cada uno de los puntos identificados en la figura 9A con la especie de ARN a la que se decodifica, "no identificado" implica que la palabra binaria medida no coincide con ninguna de las 1001 palabras de código. La figura 19C muestra las razones de confianza normalizadas medidas para las 985 especies de ARN (izquierda) y las 16 palabras de control de identificaciones erróneas no dirigidas a ningún ARN (derecha). La razón de confianza normalizada se define como en la figura 6F. La figura 19D muestra un histograma de la reducción en la abundancia detectada de 107 genes presentes tanto en los experimentos de 1001 genes como en los experimentos de 140 genes. "Disminución en veces en el número de copias" se define como el número promedio de moléculas de ARN por célula para cada especie medido en los experimentos de 140 genes dividido entre el número promedio correspondiente medido en los experimentos de 1001 genes.

La figura 18C es una comparación de los números de copias de ARN promedio por célula medidos en los experimentos de 1001 genes usando el código MHD2 con los determinados mediante smFISH convencional para 10 genes. La razón promedio del número de copias medido usando las mediciones de MHD2 con respecto al medido

mediante smFISH convencional fue de 0,30 +/- 0,05 (media +/- EEM en 10 genes). La línea discontinua corresponde a la línea $y = x$ y la línea de puntos corresponde a la línea $y = 0,30x$.

EJEMPLO 9

Los ejemplos anteriores ilustran un esquema de detección altamente multiplexado para la obtención de imágenes de ARN a nivel de sistema en células individuales. Usando el marcaje combinatorio, la hibridación secuencial y la obtención de imágenes, y dos esquemas diferentes de codificación robustos frente a errores, se obtuvieron simultáneamente imágenes o bien de 140 o bien de 1001 genes en cientos de células fibroblásticas humanas individuales. De los dos esquemas de codificación presentados en este caso, el código MHD4 es capaz tanto de detectar como de corregir errores y, por tanto, puede proporcionar una mayor tasa de identificaciones correctas y una menor tasa de identificaciones erróneas que el código MHD2, que sólo puede detectar errores, pero no corregirlos. MHD2, por otro lado, proporciona un ajuste a escala más rápido del grado de multiplexación con el número de bits que MHD4. También pueden usarse otros esquemas de codificación robustos frente a errores para este tipo de obtención de imágenes multiplexada, y los experimentadores pueden establecer el equilibrio entre la precisión de la detección y la facilidad de multiplexación basándose en los requisitos específicos de los experimentos.

Aumentando el número de bits en las palabras de código, debería ser posible aumentar adicionalmente el número de especies de ARN detectables usando MERFISH con, por ejemplo, un código MHD4 o MHD2. Por ejemplo, el uso del código MHD4 con 32 bits totales y cuatro o seis bits "1" aumentaría el número de especies de ARN abordables a 1.240 o 27.776, respectivamente. Esta última es la escala aproximada del transcriptoma humano. Las tasas de identificaciones correctas e identificaciones erróneas predichas siguen siendo razonables para el código MHD4 de 32 bits (mostradas en las figuras 5C y 5D para el código MHD4 con cuatro bits "1" y se calcularon tasas similares para el código MHD4 con seis bits "1"). Si se desean mediciones más precisas, un aumento adicional en el número de bits permitiría el uso de esquemas de codificación con una distancia de Hamming superior a 4, mejorando adicionalmente la capacidad de detección y corrección de errores. Mientras que un aumento en el número de bits añadiendo más rondas de hibridación aumentaría el tiempo de recogida de datos y potencialmente conduciría a la degradación de la muestra, estos problemas podrían mitigarse mediante la utilización de múltiples colores para leer múltiples bits en cada ronda de hibridación.

A medida que aumenta el grado de multiplexación, es importante tener en cuenta el aumento potencial de la densidad de ARN que deben resolverse en cada ronda de obtención de imágenes. Basándose en los resultados de la obtención de imágenes y de la secuenciación, puede estimarse que la inclusión de todo el transcriptoma de las células IMR90 conduciría a una densidad total de ARN de ~200 moléculas/micrómetro³. Usando los métodos de obtención de imágenes y análisis actuales, podrían resolverse 2-3 moléculas/micrómetro³ por ronda de hibridación, lo que alcanzaría una densidad total de ARN de ~20 moléculas/micrómetro³ después de 32 rondas de hibridación. Esta densidad debería permitir la obtención simultánea de imágenes de todos los genes, excepto del 10 % más expresado, o la inclusión de un subconjunto de genes con niveles de expresión aún más elevados. Utilizando algoritmos de análisis de imagen más avanzados para resolver mejor las imágenes solapantes de moléculas individuales, tales como la detección comprimida, es posible ampliar la densidad que puede resolverse en ~4 veces y permitir así que se obtengan imágenes de todos los genes juntos, excepto el 2 % más expresado.

Estos ejemplos han ilustrado la utilidad de los datos derivados de la obtención de imágenes de ARN altamente multiplexada mediante el uso de análisis de covariación y correlación para revelar distintos patrones de distribución subcelular de los ARN, restringir las redes de regulación génica y predecir las funciones de muchos genes previamente no anotados o parcialmente anotados con funciones desconocidas. Dada su capacidad para cuantificar ARN en una amplia gama de abundancias sin sesgo de amplificación mientras se preserva el contexto nativo, sistemas y métodos como MERFISH permitirán muchas aplicaciones de análisis transcriptómicos *in situ* de células individuales en cultivo o tejidos complejos.

EJEMPLO 10

A continuación se presentan diversos materiales y métodos usados en los ejemplos anteriores.

Diseño de la sonda. A cada especie de ARN del conjunto diana se le asignó aleatoriamente una palabra de código binaria de entre las 140 palabras de código posibles del código MHD4 de 16 bits o de entre las 1001 palabras de código posibles del código MHD2 de 14 bits.

Se usaron agrupaciones de oligonucleótidos sintetizadas en matriz como moldes para preparar las sondas de codificación. La molécula molde para cada sonda de codificación contenía tres componentes: i) una secuencia de direccionamiento central para la hibridación *in situ* con el ARN diana, ii) dos secuencias de lectura flanqueantes diseñadas para hibridarse con cada una de las dos sondas de lectura distintas, y iii) dos secuencias de cebadores flanqueantes para permitir la amplificación enzimática de las sondas (figura 13). Las secuencias de lectura se tomaron de las 16 secuencias de lectura posibles, cada una correspondiente a una ronda de hibridación. Las secuencias de lectura se asignaron a las sondas de codificación de manera que, para cualquier especie de ARN,

cada una de las 4 secuencias de lectura se distribuyera uniformemente a lo largo del ARN diana y apareciera con la misma frecuencia. Las moléculas molde para la biblioteca de 140 genes también incluían una región de cebado común de 20 nucleótidos (nt) entre el primer cebador de PCR y la primera secuencia de lectura. Esta secuencia de cebado se usó para la etapa de transcripción inversa descrita a continuación.

Se incorporaron múltiples experimentos en una única agrupación de oligonucleótidos sintetizada en matriz, y se usó la PCR para amplificar selectivamente sólo los oligonucleótidos necesarios para un experimento específico. Las secuencias de cebadores para esta reacción PCR indexada se generaron a partir de un conjunto de secuencias ortogonales de 25 nt. Estas secuencias se recortaron a 20 nt y se seleccionaron por i) un estrecho intervalo de temperatura de fusión (de 70 °C a 80 °C), ii) la ausencia de repeticiones consecutivas de 3 o más nucleótidos idénticos, y iii) la presencia de una pinza GC, es decir, una de las dos bases terminales 3' debe ser G o C. Para mejorar adicionalmente la especificidad, se cribaron estas secuencias con respecto al transcriptoma humano usando BLAST+, y se eliminaron los cebadores con 14 o más bases contiguas de homología. Por último, volvió a usarse BLAST+ para identificar y excluir los cebadores que tenían una región homóloga de 11 nt en el extremo 3' de cualquier otro cebador o una región homóloga de 5 nt en el extremo 3' del promotor T7. Las secuencias del cebador directo (cebador 1) se determinaron tal como se describió anteriormente, mientras que los cebadores inversos contienen cada uno una secuencia de 20 nt tal como se describió anteriormente más una secuencia de promotor T7 de 20 nt para facilitar la amplificación mediante transcripción *in vitro* (cebador 2). A continuación se enumeran las secuencias de cebadores usadas en los experimentos de 140 y 1001 genes.

Tabla 2

<u>Nombre del experimento</u>	<u>Secuencia de cebador 1</u> <u>(cebador de índice 1)</u>	<u>Secuencia de cebador 2</u> <u>(promotor T7 más el complemento inverso</u> <u>del cebador de índice 2)</u>
Libro de código 1 de 140 genes	GTTCGGTCGGCACTTGGGTG C (SEQ ID NO: 18)	TAATACGACTCACTATAGGGAAAGCCGG TTCATCCGGTGG (SEQ ID NO: 21)
Libro de código 2 de 140 genes	CGATGCGCCAATTCGGGTTC (SEQ ID NO: 19)	TAATACGACTCACTATAGGGTGATCATC GCTCGCGGGTTG (SEQ ID NO: 22)
1001 genes	CGCGGGCTATATGCGAACC G (SEQ ID NO: 20)	TAATACGACTCACTATAGGGCGTGGAGG GCATACAACGC (SEQ ID NO: 23)

Las secuencias de lectura de 30 nt de longitud se crearon concatenando fragmentos del mismo conjunto de cebadores ortogonales generado anteriormente combinando un cebador de 20 nt con un fragmento de 10 nt de otro. A continuación, se cribaron estas secuencias de lectura mediante BLAST+ para determinar su ortogonalidad con las secuencias de cebadores de índice y otras secuencias de lectura (no más de 11 nt de homología) y de posibles sitios de unión inespecíficos en el genoma humano (no más de 14 nt de homología). Se usaron sondas de lectura marcadas con fluorescencia con secuencias complementarias a las secuencias de lectura para investigar estas secuencias de lectura, una en cada ronda de hibridación. A continuación se enumeran todas las secuencias de sondas de lectura usadas.

Tabla 3

<u>Bit</u>	<u>Sondas de lectura</u>	
1	CGCAACGCTTGGGACGGTTCCAATCGGATC/3Cy5Sp/	SEQ ID NO: 24
2	CGAATGCTCTGGCCTCGAACGAACGATAGC/3Cy5Sp/	SEQ ID NO: 25
3	ACAAATCCGACCAGATCGGACGATCATGGG/3Cy5Sp/	SEQ ID NO: 26
4	CAAGTATGCAGCGCGATTGACCGTCTCGTT/3Cy5Sp/	SEQ ID NO: 27
5	GCGGGAAGCACGTGGATTAGGGCATCGACC/3Cy5Sp/	SEQ ID NO: 28
6	AAGTCGTACGCCGATGCGCAGCAATTCAC/3Cy5Sp/	SEQ ID NO: 29
7	CGAAACATCGGCCACGGTCCCGTTGAACTT/3Cy5Sp/	SEQ ID NO: 30
8	ACGAATCCACCGTCCAGCGCGTCAAACAGA/3Cy5Sp/	SEQ ID NO: 31
9	CGCGAAATCCCCGTAACGAGCGTCCCTTGC/3Cy5Sp/	SEQ ID NO: 32
10	GCATGAGTTGCCTGGCGTTGCGACGACTAA/3Cy5Sp/	SEQ ID NO: 33
11	CCGTCGTCTCCGGTCCACCGTTGCGCTTAC/3Cy5Sp/	SEQ ID NO: 34
12	GGCCAATGGCCCAGGTCCGTCACGCAATTT/3Cy5Sp/	SEQ ID NO: 35
13	TTGATCGAATCGGAGCGTAGCGGAATCTGC/3Cy5Sp/	SEQ ID NO: 36
14	CGCGCGGATCCGCTTGTCGGGAACGGATAC/3Cy5Sp/	SEQ ID NO: 37
15	GCCTCGATTACGACGGATGTAATTCGGCCG/3Cy5Sp/	SEQ ID NO: 38
16	GCCCGTATTCCCGCTTGCGAGTAGGGCAAT/3Cy5Sp/	SEQ ID NO: 39

Las sondas de lectura usadas para las bibliotecas de 140 genes fueron las sondas 1 a 16. Las sondas de lectura usadas para el experimento de 1001 genes fueron las sondas 1 a 14. /3Cy5Sp/ indica una modificación de Cy5 en 3'.

Para diseñar las secuencias de direccionamiento centrales de las sondas de codificación, se cumplieron la abundancia de diferentes transcritos en células IMR90 usando Cufflinks v2.1, datos de ARN total del proyecto ENCODE y anotaciones del genoma humano de Gencode v18. Las sondas se diseñaron a partir de modelos de genes correspondientes a la isoforma más abundante usando OligoArray2.1 con las siguientes restricciones: la región de la secuencia diana tiene una longitud de 30 nt; las temperaturas de fusión de la región hibridada de la sonda y la diana de ARN celular es superior a 70 °C; ausente de dianas de hibridación cruzada con temperaturas de fusión superiores a 72 °C; ausente de estructuras secundarias internas previstas con temperaturas de fusión superiores a 76 °C; y ausente de repeticiones contiguas de 6 o más nucleótidos idénticos. Las temperaturas de fusión se ajustaron para optimizar la especificidad de estas sondas y minimizar la estructura secundaria sin dejar de producir un número suficiente de sondas para las bibliotecas. Para disminuir el coste computacional, las isoformas se dividieron en regiones de 1 kb para el diseño de las sondas. Usando BLAST+, se rechazaron todas las sondas potenciales que se mapearon a más de una especie de ARN celular. Se mantuvieron las sondas con múltiples dianas en el mismo ARN.

Para cada gen en los experimentos de 140 genes, se generaron 198 secuencias putativas de sondas de codificación concatenando los cebadores de índice, las secuencias de lectura y las regiones de direccionamiento apropiados, tal como se muestra en la figura 13. Para abordar la posibilidad de que la concatenación de estas secuencias introdujera nuevas regiones de homología con ARN inespecífica, se usó BLAST+ para el cribado de estas secuencias putativas con respecto a todas las secuencias humanas de ARNr y ARNt, así como a genes con alta expresión (genes con FPKM > 10.000). Se retiraron las sondas con más de 14 nt de homología con ARNr o ARNt o más de 17 nt de homología con genes con alta expresión. Después de estos cortes, había ~192 (con una desviación estándar de 2) sondas por gen para ambos libros de código MHD4 usados en los experimentos de 140 genes. Se usó el mismo protocolo para los experimentos de 1001 genes, tal como sigue: partiendo de 96 secuencias de direccionamiento putativas por gen, se obtuvieron ~94 (con una desviación estándar de 6) sondas de codificación por gen después de estos cortes de homología adicionales. El número de sondas de codificación por ARN se redujo para los experimentos de 1001 genes, de modo que estas sondas pudieran sintetizarse a partir de una única agrupación de oligonucleótidos de 100.000 miembros, en lugar de dos agrupaciones separadas. Cada sonda de codificación se diseñó para contener dos de las cuatro secuencias de lectura asociadas a cada palabra de código, por lo que sólo la mitad de las sondas de codificación unidas pueden unirse a la sonda de lectura durante cualquier ronda de hibridación. Se usaron ~192 o ~94 sondas de codificación por ARN para obtener relaciones señal-fondo elevadas para moléculas de ARN individuales. El número de sondas de codificación por ARN podría reducirse sustancialmente, pero seguir permitiendo la identificación de moléculas de ARN individuales. Además, el aumento del número de secuencias de lectura por sonda de codificación o el uso de métodos de seccionamiento óptico para reducir el fondo de fluorescencia pueden permitir una mayor reducción del número de sondas de codificación por ARN.

Se diseñaron dos tipos de controles de identificaciones erróneas. El primer control (palabras en blanco) no estaba representada con sondas de codificación. El segundo tipo de control (palabras sin diana) tenía sondas de codificación que no se dirigían a ningún ARN celular. Las regiones de direccionamiento de estas sondas estaban compuestas por secuencias aleatorias de nucleótidos sujetas a las mismas restricciones usadas para diseñar las secuencias de direccionamiento a ARN descritas anteriormente. Además, se cribaron estas secuencias aleatorias con respecto al transcriptoma humano para garantizar que no contienen homología significativa (>14 nt) con ningún ARN humano. Las mediciones de 140 genes contenían 5 palabras en blanco y 5 palabras sin diana. Las mediciones de 1001 genes contenían 11 palabras en blanco y 5 palabras sin diana.

Síntesis de sondas. Las sondas de codificación se sintetizaron usando las siguientes etapas, y este protocolo de síntesis se ilustra en la figura 13.

Etapas 1: la agrupación de oligonucleótidos molde (CustomArray) se amplificó mediante PCR de ciclo limitado en un dispositivo Bio-Rad CFX96 usando secuencias de cebadores específicas para el conjunto de sondas deseado. Para facilitar la posterior amplificación mediante transcripción *in vitro*, el cebador inverso contenía el promotor T7. Todos los cebadores se sintetizaron por IDT. Esta reacción se purificó en columna (Zymo DNA Clean and Concentrator; D4003).

Etapas 2: los productos de PCR purificados se amplificaron ~200 veces más y se convirtieron en ARN mediante una transcripción *in vitro* de alto rendimiento siguiendo las instrucciones del fabricante (New England Biolabs, E2040S). Cada reacción de 20 microlitros contenía ~1 microgramo de ADN molde del anterior, 10 mM de cada NTP, 1x tampón de reacción, 1x inhibidor de ARNasa (Promega RNasin, N2611) y 2 microlitros de la polimerasa T7. Esta reacción se incubó a 37 °C durante 4 horas para maximizar el rendimiento. Esta reacción no se purificó antes de las etapas siguientes.

Etapas 3: a continuación, los productos de ARN de la reacción de transcripción *in vitro* anterior se convirtieron de nuevo en ADN mediante una reacción de transcripción inversa. Cada reacción de 50 microlitros contenía el ARN no purificado producido en la etapa 2, complementado con 1,6 mM de cada dNTP, 2 nmol de un cebador de transcripción inversa, 300 unidades de transcriptasa inversa Maxima H- (Thermo Scientific, EP0751), 60 unidades de RNasin y una concentración final 1x del tampón Maxima RT. Esta reacción se incubó a 50 °C durante 45 minutos y la transcriptasa inversa se inactivó a 85 °C durante 5 minutos. Los moldes para las bibliotecas de 140 genes contienen una región de cebado común para esta etapa de transcripción inversa; por tanto, se usó un único cebador para esta etapa al crear estas sondas. Su secuencia fue CGGGTTTAGCGCCGGAAATG (SEQ ID NO: 40). No se incluyó una región de cebado común para la biblioteca de 1001 genes, por lo que la transcripción inversa se realizó con el cebador directo: CGCGGGCTATATGCGAACCG (SEQ ID NO: 20).

Etapas 4: para retirar el ARN molde, se añadieron 20 microlitros de EDTA 0,25 M y NaOH 0,5 N a la reacción anterior para hidrolizar selectivamente el ARN, y la muestra se incubó a 95 °C durante 10 minutos. A continuación, esta reacción se purificó inmediatamente mediante purificación en columna usando una columna de 100 microgramos de capacidad (Zymo Research, D4030) y el protocolo Zymo Oligo Clean and Concentrator. Las sondas finales se eluyeron en 100 microlitros de agua desionizada libre de ARNasa, se evaporaron en un concentrador de vacío y luego se resuspendieron en 10 µl de tampón de hibridación de codificación (véase a continuación). Las sondas se almacenaron a -20 °C. Se usaron electroforesis en gel de poli(acrilamida) desnaturante y espectroscopia de absorción para confirmar la calidad de las sondas y revelaron que este protocolo de síntesis de sondas convierte el 90-100 % del cebador de transcripción inversa en una sonda de longitud completa y que, de la sonda construida, se recupera el 70-80 % durante la etapa de purificación.

Las sondas de lectura marcadas con fluorescencia tienen secuencias complementarias a las secuencias de lectura descritas anteriormente y un colorante Cy5 unido en el extremo 3'. Estas sondas se obtuvieron a partir de IDT y se purificaron mediante HPLC.

Preparación de muestras y marcaje con sondas de codificación. En este trabajo se usaron fibroblastos primarios humanos (Colección Americana de Cultivos Tipo, IMR90). Estas células son relativamente grandes y planas, lo que facilita la obtención de imágenes de campo amplio sin necesidad de seccionamiento óptico. Las células se cultivaron con medio esencial mínimo de Eagle. Las células se sembraron en cubreobjetos de 22 mm, n.º 1,5 (Bioptechs, 0420-0323-2) a 350.000 células/cubreobjetos y se incubaron a 37 °C con el 5 % de CO₂ durante 48-96 horas dentro de placas de Petri. Las células se fijaron durante 20 minutos en paraformaldehído al 4 % (Electron Microscopy Sciences, 15714) en solución salina tamponada con fosfato 1x (PBS; Ambion, AM9625) a temperatura ambiente, se redujeron durante 5 minutos con borohidruro de sodio al 0,1 % p/v (Sigma, 480886) en agua para reducir la fluorescencia de fondo, se lavaron tres veces con PBS 1x helado, se permeabilizaron durante 2 minutos con Triton al 0,5 % v/v (Sigma, T8787) en PBS 1x a temperatura ambiente y se lavaron tres veces con PBS 1x helado.

Las células se incubaron durante 5 minutos en un tampón de lavado de codificación que comprendía 2x tampón salino-citrato de sodio (SSC) (Ambion, AM9763), formamida al 30 % v/v (Ambion, AM9342) y complejo vanadilo-ribonucleósido 2 mM (NEB, S1402S). Se añadieron 10 microlitros de sondas de codificación de 100 micromolar (experimentos de 140 genes) o de 200 micromolar (experimentos de 1001 genes) en tampón de hibridación de

codificación al cubreobjetos que contenía las células y se extendieron uniformemente colocando otro cubreobjetos encima de la muestra. A continuación, las muestras se incubaron en una cámara húmeda dentro de un horno de hibridación a 37 °C durante 18-36 horas. El tampón de hibridación de codificación se compone de tampón de lavado de codificación complementado con ARNt de levadura 1 mg/ml (Life Technologies, 15401-011) y sulfato de dextrano al 10 % p/v (Sigma, D8906-50G).

A continuación, se lavaron las células con tampón de lavado de codificación primaria, se incubaron a 47 °C durante 10 minutos y se repitió este lavado un total de tres veces. Una dilución 1:1000 de perlas fluorescentes anaranjadas modificadas con carboxilato de 0,2 micrómetros de diámetro (Life Technologies, F-8809) en 2xSSC se sonicó durante 3 minutos y luego se incubó con la muestra durante 5 minutos. Las perlas se usaron como marcadores fiduciales para alinear las imágenes obtenidas de múltiples rondas sucesivas de hibridación, tal como se describe a continuación. La muestra se lavó una vez con 2xSSC y, a continuación, se fijó posteriormente con paraformaldehído al 4 % v/v en 2xSSC a temperatura ambiente durante 30 minutos. A continuación, la muestra se lavó tres veces con 2xSSC y o bien se sometió inmediatamente a obtención de imágenes o bien se almacenó durante no más de 12 horas a 4 °C antes de la obtención de imágenes. Todas las disoluciones se prepararon libres de ARNasa.

Obtención de imágenes de MERFISH con múltiples rondas sucesivas de hibridación. El cubreobjetos de la muestra se montó en una cámara de flujo FCS2 de Biotech, y el flujo a través de esta cámara se controló mediante un sistema fluido casero compuesto por tres válvulas de 8 vías controladas por ordenador (Hamilton, MVP y HVXM 8-5) y una bomba peristáltica controlada por ordenador (Rainin, Dynamax RP-1). Se obtuvieron imágenes de la muestra en un microscopio casero construido con un cuerpo de Olympus IX-71 y un objetivo de inmersión en aceite de 1,45 NA, 100x y configurado para excitación de incidencia oblicua. El objetivo se calentó hasta 37 °C con un calentador de objetivos Biotech. Se mantuvo un enfoque constante durante todo el procedimiento de obtención de imágenes con un sistema de autoenfoco casero. La iluminación se proporcionó a 641 nm, 561 nm y 405 nm usando láseres de estado sólido (MPB communications, VFL-P500-642; Coherent, 561-200CWCDRH; y Coherent, 1069413/AT) para la excitación de las sondas de lectura marcadas con Cy5, las perlas fiduciales, y las contratincciones nucleares, respectivamente. Estas líneas se combinaron con un filtro dicróico personalizado (Chroma, zy405/488/561/647/752RP-UF1) y la emisión se filtró con un filtro dicróico personalizado (Chroma, ZET405/488/561/647-656/752m). La fluorescencia se separó con un dispositivo QuadView (Photometrics) usando los filtros dicróicos T5601pxr, T6501pxr, 750dcxxr (Chroma) y los filtros de emisión ET525/50m, WT59550m-2f, ET700/75m, HQ7701p (Chroma) y se obtuvieron imágenes con una cámara EMCCD (Andor, iXon-897). La cámara se configuró de modo que un píxel correspondiera a 167 nm en el plano de la muestra. Todo el sistema estaba totalmente automatizado, de modo que la obtención de imágenes y la manipulación de fluidos se realizaron durante todo el experimento sin intervención del usuario.

La hibridación secuencial, la obtención de imágenes, y el blanqueo se realizaron tal como sigue. Se hizo fluir 1 ml de 10 nM de la sonda de lectura marcada con fluorescencia apropiada en tampón de hibridación de lectura (2xSSC; formamida al 10 % v/v; sulfato de dextrano al 10% p/v, y complejo vanadilo-ribonucleósido 2 mM) a través de la muestra, se detuvo el flujo y se incubó la muestra durante 15 minutos. A continuación, se hicieron fluir 2 ml de tampón de lavado de lectura (2xSSC, formamida al 20 % v/v; y complejo vanadilo-ribonucleósido 2 mM) a través de la muestra, se detuvo el flujo y se incubó durante 3 minutos. Se hicieron fluir 2 ml de tampón de obtención de imágenes que comprendía 2xSSC, TrisHCl 50 mM pH 8, glucosa al 10 % p/v, Trolox 2 mM (Sigma- Aldrich, 238813), glucosa oxidasa 0,5 mg/ml (Sigma-Aldrich, G2133), y catalasa 40 microgramos/ml (Sigma-Aldrich, C30) a través de la muestra. A continuación, se detuvo el flujo y luego se expusieron de 75 a 100 regiones a ~25 mW de luz de 642 nm y 1 mW de luz de 561 nm y se obtuvieron imágenes. Cada región era de 40 micrómetros por 40 micrómetros. Las potencias del láser se midieron en el puerto posterior del microscopio. Dado que el tampón de obtención de imágenes es sensible al oxígeno, los ~50 ml de tampón de obtención de imágenes usados para un solo experimento se prepararon justo al comienzo del experimento y luego se almacenaron bajo una capa de aceite mineral durante toda la medición. El tampón almacenado de este modo se mantuvo estable durante más de 24 horas.

Después de la obtención de imágenes, la fluorescencia de las sondas de lectura se extinguió mediante fotoblanqueo. La muestra se lavó con 2 ml de tampón de fotoblanqueo (2xSSC y complejo vanadilo-ribonucleósido 2 mM) y cada región de la muestra de la que se obtuvieron imágenes se expuso a 200 mW de luz de 641 nm durante 3 s. Para confirmar la eficacia de este tratamiento de fotoblanqueo, se introdujo de nuevo el tampón de obtención de imágenes y se obtuvieron imágenes de la muestra tal como se describió anteriormente.

El procedimiento anterior de hibridación, obtención de imágenes, y fotoblanqueo se repitió 16 veces para las mediciones de 140 genes usando el código MHD4 o 14 veces para las mediciones de 1001 genes usando el código MHD2. Un experimento completo se completaba normalmente en ~20 horas.

Tras la finalización de la obtención de imágenes, se hicieron pasar 2 ml de una dilución 1:1000 de Hoescht (ENZ-52401) en 2xSSC a través de la cámara para marcar los núcleos de las células. A continuación, la muestra se lavó inmediatamente con 2 ml de 2xSSC seguido de 2 ml de tampón de obtención de imágenes. A continuación, se obtuvieron imágenes de cada región de la muestra una vez más con ~1 mW de luz de 405 nm.

Dado que se obtuvieron imágenes de las células usando la obtención de imágenes de campo amplio con iluminación

de incidencia oblicua, sin seccionamiento óptico ni escaneo z, se cuantificó la fracción de especies individuales de ARN que se encontraba fuera del intervalo axial de la geometría de obtención de imágenes para 6 especies diferentes de ARN usando smFISH convencional. Para este fin, estas células se seccionaron ópticamente recogiendo apilamientos de imágenes a diferentes profundidades focales a través de toda la profundidad de las células. Las imágenes se alinearon en planos focales consecutivos y luego se calculó para cada célula la fracción de ARN detectados en el apilamiento tridimensional pero no en el plano focal basal. Se encontró que sólo una pequeña fracción, 15 % +/- 1 % (media +/- EEM en seis especies diferentes de ARN), de moléculas de ARN estaba fuera del intervalo de obtención de imágenes de un plano focal fijo sin escaneo z. Estas mediciones también confirmaron que la geometría de excitación iluminaba toda la profundidad de las células. Cualquier técnica de seccionamiento óptico podría emplearse en MERFISH para permitir la obtención de imágenes de ARN en células o tejidos más gruesos.

Construcción de palabras medidas. Los puntos fluorescentes se identificaron y localizaron en cada imagen usando un algoritmo de ajuste multigaussiano que supone una gaussiana con una anchura uniforme de 167 nm. Este algoritmo permitió distinguir y ajustar individualmente los puntos parcialmente solapantes. Los puntos de ARN se distinguieron de la señal de fondo, es decir, de la señal procedente de sondas unidas de forma no específica, estableciendo el umbral de intensidad necesario para ajustar un punto con este software. Debido a la variación en el brillo de los puntos entre las rondas de hibridación, este umbral se ajustó adecuadamente para cada ronda de hibridación para minimizar el promedio combinado de las tasas de errores $1 \rightarrow 0$ y $0 \rightarrow 1$ en todas las rondas de hibridación (mediciones de 140 genes) o para maximizar la razón del número de palabras medidas con cuatro bits "1" con respecto a aquellas con tres o cinco bits "1" (mediciones de 1001 genes). La ubicación de las perlas fiduciales se identificó en cada trama usando un algoritmo de ajuste gaussiano simple más rápido.

Las imágenes de la misma región de la muestra en diferentes rondas de hibridación se registraron rotando y trasladando la imagen para alinear las dos perlas fiduciales dentro de la misma imagen que fueran más similares en cuanto a ubicación tras una alineación inicial aproximada mediante correlación de imágenes. Todas las imágenes se alinearon con respecto a un sistema de coordenadas establecido por las imágenes recogidas en la primera ronda de hibridación. La calidad de esta alineación se determinó a partir de la distancia residual entre cinco perlas fiduciales adicionales, y el error de alineación fue normalmente de ~20 nm.

Los puntos de fluorescencia en diferentes rondas de hibridación se conectaron en una sola cadena, correspondiente a una molécula potencial de ARN, si la distancia entre los puntos era inferior a 1 píxel (167 nm). Para cada cadena de puntos, la secuencia de encendido-apagado de las señales fluorescentes en todas las rondas de hibridación se usó para asignar una palabra binaria a la molécula de ARN potencial, en la que "1" se asignó a las rondas de hibridación que contenían una señal fluorescente por encima del umbral y "0" se asignó a las otras rondas de hibridación. A continuación, las palabras medidas se decodificaron en especies de ARN usando el código MHD4 de 16 bits o el código MHD2 de 14 bits comentados anteriormente. En el caso del código MHD4 de 16 bits, si la palabra binaria medida coincidía perfectamente con la palabra de código de un ARN específico o difería de la palabra de código en un solo bit, se asignaba a ese ARN. En el caso del código MHD2 de 14 bits, sólo si la palabra binaria medida coincidía perfectamente con la palabra clave de un ARN específico, se asignaba a ese ARN. Para determinar el número de copias por célula, se contó el número de cada especie de ARN en células individuales dentro de cada área de obtención de imágenes de 40 micrómetros por 40 micrómetros. Cabe señalar que este número representa la mayoría, pero no todas las moléculas de ARN dentro de una célula, ya que una fracción de la célula podría estar fuera del área de obtención de imágenes o de la profundidad focal. Podrían emplearse imágenes en mosaico de áreas adyacentes y planos focales adyacentes para mejorar la precisión del recuento.

En los experimentos de 140 genes, algunas regiones del núcleo celular contenían ocasionalmente demasiada señal de fluorescencia para identificar correctamente puntos individuales de ARN. En los experimentos de 1001 genes, el núcleo celular contenía en general demasiada señal fluorescente para permitir la identificación de moléculas individuales de ARN. Estas regiones brillantes se excluyeron de todos los análisis posteriores. Este trabajo se centra en los ARNm, que están enriquecidos en el citoplasma. Para estimar la fracción de ARNm que se pasa por alto al excluir la región del núcleo, se usó smFISH convencional para cuantificar la fracción de moléculas encontradas dentro del núcleo para seis especies diferentes de ARNm. Se descubrió que sólo el 5 % +/- 2 % (media +/- EEM de las seis especies de ARN) de estas moléculas de ARN se encuentran en el núcleo. El empleo de obtención de imágenes de superresolución y/o de seccionamiento óptico podría permitir posiblemente la identificación de moléculas individuales en estas densas regiones del núcleo, lo que sería particularmente útil para investigar aquellos ARN no codificantes que están enriquecidos en el núcleo.

Mediciones de smFISH de genes individuales. Se adquirieron agrupaciones de 48 sondas de oligonucleótidos marcadas con fluorescencia (Quasar 670) por ARN de Biosearch Technologies. Las secuencias de sonda de 30 nt se tomaron directamente de un subconjunto aleatorio de las regiones de direccionamiento usadas para las mediciones multiplexadas. Las células se fijaron y permeabilizaron tal como se describió anteriormente. Se añadieron 10 microlitros de sondas de oligonucleótidos de 250 nM en tampón de hibridación de codificación (descrito anteriormente) al cubreobjetos que contenía las células y se extendieron uniformemente colocando otro cubreobjetos encima de la muestra. A continuación, las muestras se incubaron en una cámara húmeda dentro de un horno de hibridación a 37 °C durante 18 horas. A continuación, las células se lavaron con tampón de lavado de codificación (descrito anteriormente) a 37 °C durante 10 minutos, y este lavado se repitió un total de tres veces. A

continuación, se lavó la muestra tres veces con 2xSSC y se obtuvieron imágenes en tampón de obtención de imágenes usando la misma geometría de obtención de imágenes descrita anteriormente para MERFISH.

Secuenciación de ARN masiva. El ARN total se extrajo de células IMR90 cultivadas tal como se indicó anteriormente usando el kit Zymo Quick RNA MiniPrep (R1054) según las instrucciones del fabricante. A continuación, se seleccionó el ARN poliA (NEB; E7490) y se construyó una biblioteca de secuenciación usando el kit de preparación de bibliotecas de ARN NEB Next Ultra (NEB; E7530), se amplificó con oligonucleótidos personalizados y se obtuvieron lecturas de 150 pb en un MiSeq. Estas secuencias se alinearon con el genoma humano (Gencode v18) y la abundancia de isoformas se calculó con cufflinks.

Cálculo de propiedades previstas de error y ajuste de escala de diferentes esquemas de codificación. Se derivaron expresiones analíticas para la dependencia del número de palabras de código posibles, la tasa de identificaciones correctas, y la tasa de identificaciones erróneas en N . La tasa de identificaciones correctas se define como la fracción de moléculas de ARN que se identifican correctamente. La tasa de identificaciones erróneas se define como la fracción de moléculas de ARN que se identifican erróneamente como una especie de ARN equivocada. Para los esquemas de codificación con capacidad de detección de errores, la tasa de identificaciones correctas y la tasa de identificaciones erróneas no suman 1 porque una fracción de las moléculas no detectadas correctamente pueden detectarse como errores y descartarse y, por tanto, no identificarse erróneamente como una especie equivocada. Estos cálculos suponen que la probabilidad de lectura errónea de bits es constante para todas las rondas de hibridación, pero difiere para los errores 1-- >0 y 0-- >1. Se usaron las tasas medias de errores 1-- >0 y 0-- >1 promedio medidas experimentalmente (10 % y 4 % respectivamente) para las estimaciones mostradas en las figuras 5B-5D. Para simplificar, la palabra correspondiente a todos los "0" no se retiró de los cálculos.

Para el esquema de codificación binaria simple en el que todas las palabras binarias posibles de N bits se asignan a especies de ARN singulares, el número de palabras de código posibles es 2^N . El número de palabras que podrían usarse para codificar ARN es en realidad $2^N - 1$ porque la palabra de código "00...0" no contiene fluorescencia detectable en ninguna ronda de hibridación, pero por simplicidad la palabra correspondiente a todos los "0" no se retiró de los cálculos posteriores. El error introducido por esta aproximación es insignificante. Para cualquier palabra dada con m "1" y $N-m$ "0" la probabilidad de medir esa palabra sin error (la fracción de ARN que se identifica correctamente) es:

$$(1 - p_1)^m (1 - p_0)^{N-m}, \quad (1)$$

donde p_1 es la tasa de error 1-- >0 y p_0 es la tasa de error 0-- >1 por bit. Dado que diferentes palabras en este sencillo esquema de codificación binaria pueden tener diferentes números de bits "1", la tasa de identificaciones correctas para diferentes palabras diferirá si $p_1 \neq p_0$. La tasa de identificaciones correctas promedio, notificada en la figura 5C, se determinó a partir del promedio ponderado del valor de la Ec. (1) para todas las palabras. Este promedio ponderado es:

$$\frac{1}{2^N} \sum_{m=0}^N \binom{N}{m} (1 - p_1)^m (1 - p_0)^{N-m}, \quad (2)$$

donde $\binom{N}{m}$ es el coeficiente binomial y corresponde al número de palabras con m bits "1" en este esquema de codificación. Dado que en este esquema de codificación cada error produce una palabra binaria que codifica para un ARN diferente, la tasa media de identificaciones erróneas para este esquema de codificación, notificada en la figura 5D, sigue directamente de (2):

$$1 - \frac{1}{2^N} \sum_{m=0}^N \binom{N}{m} (1 - p_1)^m (1 - p_0)^{N-m}. \quad (3)$$

Para calcular las propiedades de error y ajuste a escala del código de distancia de Hamming 4 extendido (HD4), en primer lugar se creó la matriz generadora para el número deseado de bits de datos usando métodos convencionales. La matriz generadora determina las palabras específicas que están presentes en un esquema de codificación dado y se usó para determinar directamente el número de palabras codificadas en función del número de bits. En este esquema de codificación, la tasa de identificaciones correctas corresponde a la fracción de palabras medidas sin error, así como a la fracción de palabras medidas con un error de un solo bit. Para palabras de código con m bits "1",

esta fracción se determina mediante la siguiente expresión:

$$(1-p_1)^m(1-p_0)^{N-m} + m p_1^1(1-p_1)^{m-1}(1-p_0)^{N-m} + (N-m)p_0^1(1-p_1)^m(1-p_0)^{N-m-1} \quad (4)$$

donde el primer término es la probabilidad de no cometer ningún error, el segundo término corresponde a la probabilidad total de cometer un error 1-- >0 en cualquiera de los m bits "1" sin cometer ningún otro error 0-- >1, y el último término corresponde a la probabilidad total de cometer un error 0-- >1 en cualquiera de los $N-m$ bits "0" sin cometer ningún error 1-- >0. Debido a que el número de bits "1" puede variar entre las palabras en este esquema de codificación, la tasa de identificaciones correctas promedio notificada en la figura 5C se calculó a partir de un promedio ponderado con respecto a la Ec. (4) para diferentes valores de m . El peso para cada término se determinó a partir del número de palabras que contienen m bits "1" tal como se determina a partir de la matriz generadora descrita anteriormente.

Dado que las palabras que codifican para ARN están separadas por una distancia de Hamming mínima de 4, se necesitan al menos 4 errores para convertir una palabra en otra. Si se aplica la corrección de errores, entonces 3 ó 5 errores también podrían convertir un ARN en otro. Por tanto, se estimó la tasa de identificaciones erróneas de todas las combinaciones posibles de errores de 3, 4 y 5 bits para palabras de código con m bits "1". Técnicamente, los errores >5 bits también podrían convertir un ARN en otro, pero la probabilidad de cometer tales errores es insignificante debido a la pequeña tasa de errores por bit. Esta expresión se aproximó con:

$$\begin{aligned} & \sum_{i=0}^4 \binom{m}{i} \binom{N-m}{4-i} p_1^i p_0^{4-i} (1-p_1)^{m-i} (1-p_0)^{N-m-(4-i)} + \\ & \sum_{i=0}^3 \binom{m}{i} \binom{N-m}{3-i} p_1^i p_0^{3-i} (1-p_1)^{m-i} (1-p_0)^{N-m-(3-i)} + \\ & \sum_{i=0}^5 \binom{m}{i} \binom{N-m}{5-i} p_1^i p_0^{5-i} (1-p_1)^{m-i} (1-p_0)^{N-m-(5-i)} + \end{aligned} \quad (5)$$

La primera suma corresponde a todas las formas en que pueden cometerse exactamente cuatro errores. De manera similar, la segunda y tercera sumas corresponden a todas las formas en las que pueden cometerse exactamente tres o cinco errores. La Ec. (5) proporciona un límite superior para la tasa de identificaciones erróneas porque no todos los errores de tres, cuatro o cinco bits producen una palabra que coincida con o se corrigiera a otra palabra legítima. De nuevo, dado que el número de bits "1" puede diferir entre palabras, la tasa de identificaciones erróneas promedio notificada en la figura 5D se calcula como un promedio ponderado de la Ec. (5) con respecto al número de palabras que tienen m bits "1".

Para generar el código MHD4 en el que el número de bits "1" de cada palabra de código se fija en 4, en primer lugar, se generaron los códigos HD4 tal como se describió anteriormente y, a continuación, se retiraron todas las palabras de código que no contenían cuatro "1". La tasa de identificaciones correctas de este código, notificada en la figura 5C, se calculó directamente a partir de la Ec. (4) pero con $m = 4$ porque todas las palabras de código de este código tienen cuatro bits "1". La tasa de identificaciones erróneas de este código, notificada en la figura 5D, se calculó modificando la Ec. (5) con las siguientes consideraciones: (i) el número de bits "1", m , se fijó en 4 y (ii) los errores que producen palabras que no contienen tres, cuatro o cinco bits "1" se excluyeron. Por tanto, la expresión de la Ec. (5) se simplificó a

$$\begin{aligned} & \binom{4}{2} \binom{N-4}{2} p_1^2 p_0^2 (1-p_1)^2 (1-p_0)^{N-6} + \\ & \binom{4}{1} \binom{N-4}{2} p_1^1 p_0^2 (1-p_1)^3 (1-p_0)^{N-6} + \binom{4}{2} \binom{N-4}{1} p_1^2 p_0^1 (1-p_1)^2 (1-p_0)^{N-5} \\ & \binom{4}{2} \binom{N-4}{3} p_1^2 p_0^3 (1-p_1)^2 (1-p_0)^{N-7} + \binom{4}{3} \binom{N-4}{2} p_1^3 p_0^2 (1-p_1)^1 (1-p_0)^{N-6} \end{aligned} \quad (6)$$

De nuevo, esta expresión es un límite superior de la tasa de identificaciones erróneas real, ya que no todas las palabras con cuatro "1" son palabras de código válidas.

- 5 Estimaciones de las tasas de errores $1 \rightarrow 0$ y $0 \rightarrow 1$ para cada ronda de hibridación. Para calcular la probabilidad de leer erróneamente un bit en una ronda de hibridación determinada, se usaron las propiedades de corrección de errores del código MHD4. Brevemente, las probabilidades de errores $1 \rightarrow 0$ ó $0 \rightarrow 1$ se obtuvieron de la siguiente manera. Siendo p_i la probabilidad de cometer un error en el $i^{\text{ésimo}}$ bit, es decir, en la ronda de hibridación $i^{\text{ésima}}$, y siendo A el número real de moléculas de ARN de la especie dada, entonces el número de coincidencias exactas

$$W_E = A \prod_{i=1}^{16} (1 - p_i)$$

para este ARN será y el número de coincidencias con corrección de errores de un bit para este

$$W_i = A \frac{p_i}{(1 - p_i)} \prod_{j=1}^{16} (1 - p_j)$$

- 10 ARN que corresponden a errores en el $i^{\text{ésimo}}$ bit será . Puede derivarse p_i directamente de

$$W_i / W_E = \frac{p_i}{(1 - p_i)}$$

la razón: . Esta razón supone que los recuentos con corrección de errores de un bit sólo se generaron a partir de errores de un bit de la palabra correcta y que la contaminación por errores múltiples de otras palabras de ARN es insignificante. Dado que la tasa de errores por ronda de hibridación es pequeña y que se necesitan al menos tres errores para convertir una palabra de codificación de ARN en una palabra que se identificaría erróneamente como otro ARN, la aproximación anterior debería ser buena.

- Para calcular las probabilidades de error $1 \rightarrow 0$ ó $0 \rightarrow 1$ promedio para cada una de las 16 rondas de hibridación, se usó el enfoque anterior para calcular las tasas de errores por bit para cada bit de cada gen, y estos errores se clasificaron basándose en si correspondían a un error $1 \rightarrow 0$ ó $0 \rightarrow 1$, y se tomó la media de estos errores para cada bit ponderado por el número de recuentos observados para el gen correspondiente.

- Estimaciones de la tasa de identificaciones correctas para especies de ARN individuales a partir de datos de obtención de imágenes reales. Con las estimaciones de las probabilidades de error $1 \rightarrow 0$ o $0 \rightarrow 1$ para cada ronda de hibridación determinadas anteriormente, es posible estimar la tasa de identificaciones correctas para cada ARN basándose en la palabra específica usada para codificarlo. Específicamente, la fracción de una especie de ARN que se identifica correctamente se determina mediante

$$\prod_{i=1}^N (1 - p_i) + \sum_{j=1}^N \frac{p_j}{(1 - p_j)} \prod_{i=1}^N (1 - p_i), \quad (7)$$

- 30 donde el primer término representa la probabilidad de observar una coincidencia exacta de la palabra de código y el segundo término representa la probabilidad de observar una coincidencia con corrección de errores (es decir, con un error de un bit). Los valores de la tasa de errores por bit p_i para cada especie de ARN están determinados por la palabra de código específica para ese ARN y las tasas de error $1 \rightarrow 0$ ó $0 \rightarrow 1$ medidas para cada ronda de hibridación. Si la palabra de código del ARN contiene un "1" en el $i^{\text{ésimo}}$ bit, entonces p_i se determina a partir de la
- 35 tasa de error $1 \rightarrow 0$ para la $i^{\text{ésima}}$ ronda de hibridación; si la palabra contiene un "0" en el $i^{\text{ésimo}}$ bit, p_i se determina a partir de la tasa de error $0 \rightarrow 1$ para la $i^{\text{ésima}}$ ronda de hibridación.

- Análisis de agrupamiento jerárquico de la covariación en la abundancia de ARN. El agrupamiento jerárquico de la covariación en la expresión génica tanto para los experimentos de 140 genes como para los de 1001 genes se llevó a cabo tal como sigue. En primer lugar, la distancia entre cada par de genes se determinó como 1 menos el coeficiente de correlación de Pearson de la variación entre células de los números de copias medidos de estas dos especies de ARN, ambos normalizados por el ARN total contado en la célula. Por tanto, los genes altamente correlacionados están "más cerca" entre sí y los genes altamente anticorrelacionados están "más lejos". A partir de estas distancias se construyó entonces un árbol de agrupamiento jerárquico aglomerativo usando el método de grupos de pares no ponderados con media aritmética (UPGMA). Específicamente, partiendo de genes individuales, se construyeron agrupamientos jerárquicos identificando los dos agrupamientos (o genes individuales) más próximos entre sí según la media aritmética de las distancias entre todos los pares de genes entre agrupamientos. A continuación, se agrupan los pares de agrupamientos (o genes individuales) con la menor distancia y se repite el procedimiento. A continuación, se ordena la matriz de correlaciones por pares basándose en el orden de los genes dentro de estos árboles.

Los grupos de genes con covariaciones sustanciales se identificaron seleccionando un umbral en el árbol de agrupamiento jerárquico (indicado por las líneas discontinuas en las figuras 7D y 10A) que produjo aproximadamente 10 grupos de genes cada uno de los cuales contiene al menos 4 miembros para los experimentos

de 140 genes o aproximadamente 100 grupos cada uno de los cuales contiene al menos 3 miembros para los experimentos de 1001 genes. Cabe señalar que puede cambiarse el umbral para identificar o bien grupos más pequeños con un acoplamiento más fuerte o bien grupos más grandes con un acoplamiento relativamente débil.

5 El valor de probabilidad de que un gen pertenezca a un grupo específico se determinó calculando la diferencia entre el coeficiente de correlación promedio entre ese gen y todos los demás miembros de ese grupo y el coeficiente de correlación promedio entre ese gen y todos los demás genes medidos fuera de ese grupo. La significación (valor de p) de esta diferencia se determinó mediante la prueba de la t de Student.

10 Dado que el agrupamiento jerárquico es intrínsecamente un análisis unidimensional, es decir, que cualquier gen dado sólo puede ser miembro de un único grupo, este análisis no permite identificar todos los grupos de genes correlacionados. Podrían usarse análisis de mayor dimensión, tal como el análisis de componentes principales o el agrupamiento de *k*-medias, para identificar más agrupamientos de genes covariantes.

15 Análisis de las distribuciones espaciales de ARN. Para identificar los genes que tienen distribuciones espaciales similares, cada una de las células medidas se subdividió en regiones de 2x2 y se calculó la fracción de cada especie de ARN presente en cada uno de estos grupos. Para controlar el hecho de que algunas regiones de la célula contienen naturalmente más ARN que otras, se calculó el enriquecimiento para cada gen, es decir, la razón de la fracción observada en una región dada para una especie de ARN determinada con respecto a la fracción promedio observada para todos los genes en esa misma región. Para cada par de especies de ARN, se determinó el coeficiente de correlación de Pearson de la variación entre regiones en el enriquecimiento de estas dos especies de ARN para cada célula y los coeficientes de correlación se promediaron con respecto a ~400 células de las cuales se obtuvieron imágenes en 7 conjuntos de datos independientes. A continuación, se agruparon las especies de ARN basándose en estos coeficientes de correlación promedio usando el mismo algoritmo de agrupamiento jerárquico descrito anteriormente. Debido al gran número de células usadas para el análisis, se descubrió que el agrupamiento espacial aproximado (2x2 regiones por célula) era suficiente para capturar la correlación espacial entre genes y que un agrupamiento más fino no producía grupos más significativamente correlacionados.

30 Para medir las distancias de los genes a los núcleos y al borde celular, en primer lugar, se usaron umbrales de brillo en las imágenes celulares para segmentar los núcleos y los bordes celulares identificados. A continuación, se determinó la distancia de cada molécula de ARN a la parte más cercana del núcleo y a la parte más cercana del borde celular. Para cada conjunto de datos, se calculó la distancia promedio para cada especie de ARN promediada en todas las células medidas. Estas distancias se promediaron para los genes del grupo I, los genes del grupo II o todos los genes. Sólo se usaron en este análisis las especies de ARN con al menos 10 recuentos por célula para minimizar el error estadístico en los valores de distancia.

40 Análisis de ontología génica (GO). Se seleccionaron grupos de genes a partir de los árboles jerárquicos tal como se comentó anteriormente. Se determinó una colección de términos de GO para todas las especies de ARN medidas, así como para las especies de ARN asociadas a cada grupo a partir de las anotaciones de GO humanas más recientes, usando tanto los términos de GO anotados como los términos inmediatamente anteriores o posteriores a las anotaciones encontradas. El enriquecimiento de estas anotaciones se calculó a partir de la razón de la fracción de genes dentro de cada grupo que tienen este término con respecto a la fracción de todos los genes medidos que tienen este término y el valor de p para este enriquecimiento se calculó mediante la función hipergeométrica. Sólo se consideraron los términos de GO enriquecidos de manera estadísticamente significativa con un valor de p inferior a 0,05.

50 Más generalmente, los expertos en la técnica apreciarán fácilmente que todos los parámetros, dimensiones, materiales, y configuraciones descritos en el presente documento se entienden a modo de ejemplo y que los parámetros, dimensiones, materiales, y/o configuraciones reales dependerán de la aplicación o aplicaciones específicas para las que se usen las enseñanzas de la presente invención. Por tanto, debe entenderse que las realizaciones anteriores se presentan únicamente a modo de ejemplo y que, dentro del alcance de las reivindicaciones adjuntas y equivalentes de las mismas, la invención puede practicarse de manera distinta a la descrita específicamente. La presente invención se refiere a cada característica, sistema, artículo, material, kit, y/o método individuales descritos en el presente documento. Además, cualquier combinación de dos o más de tales características, sistemas, artículos, materiales, kits, y/o métodos, si tales características, sistemas, artículos, materiales, kits, y/o métodos no son mutuamente incoherentes, se incluye dentro del alcance de las reivindicaciones.

60 Debe entenderse que todas las definiciones, tal como se definen y usan en el presente documento, prevalecen sobre las definiciones de diccionario, y/o los significados ordinarios de los términos definidos.

Los artículos indefinidos "un" y "una", tal como se usan en la memoria descriptiva y en las reivindicaciones, a menos que se indique claramente lo contrario, deben entenderse como "al menos uno".

65 La expresión "y/o", tal como se usa en la presente memoria descriptiva y en las reivindicaciones, debe entenderse como "uno o ambos" de los elementos así combinados, es decir, elementos que están presentes de manera conjunta en algunos casos y de manera disyuntiva en otros. Los elementos múltiples enumerados con "y/o" deben

interpretarse de la misma manera, es decir, "uno o más" de los elementos así combinados. Opcionalmente, pueden estar presentes otros elementos además de los elementos específicamente identificados por la cláusula "y/o", ya estén relacionados o no con los elementos específicamente identificados. Por tanto, como ejemplo no limitativo, una referencia a "A y/o B", cuando se usa junto con un lenguaje abierto tal como "que comprende" puede referirse, en una realización, a sólo A (incluyendo opcionalmente elementos distintos de B); en otra realización, a sólo B (incluyendo opcionalmente elementos distintos de A); en aún otra realización, tanto a A como a B (incluyendo opcionalmente otros elementos); etc.

Tal como se usa en la presente memoria descriptiva y en las reivindicaciones, debe entenderse que "o" tiene el mismo significado que "y/o", tal como se definió anteriormente. Por ejemplo, al separar elementos en una lista, "o" o "y/o" se interpretará como inclusivo, es decir, la inclusión de al menos uno, pero también incluyendo más de uno, de un número o lista de elementos y, opcionalmente, elementos adicionales no incluidos en la lista. Sólo los términos que indiquen claramente lo contrario, tales como "sólo uno de" o "exactamente uno de", o, cuando se usen en las reivindicaciones, "que consiste en", se referirán a la inclusión de exactamente un elemento de un número o lista de elementos. En general, el término "o", tal como se usa en el presente documento, sólo se interpretará en el sentido de que indica alternativas excluyentes (es decir, "uno u otro, pero no ambos") cuando vaya precedido de términos de exclusividad, tales como "o bien", "uno de", "sólo uno de", o "exactamente uno de". "Que consiste esencialmente en", cuando se usa en las reivindicaciones, tendrá su significado ordinario tal como se usa en el ámbito del derecho de patentes.

Tal como se usa en la presente memoria descriptiva y en las reivindicaciones, la expresión "al menos uno", en referencia a una lista de uno o más elementos, debe entenderse en el sentido de al menos un elemento seleccionado de uno cualquiera o más de los elementos de la lista de elementos, pero sin incluir necesariamente al menos uno de todos y cada uno de los elementos específicamente enumerados en la lista de elementos y sin excluir ninguna combinación de elementos de la lista de elementos. Esta definición también permite que, opcionalmente, puedan estar presentes elementos distintos de los elementos específicamente identificados dentro de la lista de elementos a los que se refiere la expresión "al menos uno", ya estén relacionados o no con los elementos específicamente identificados. Por tanto, a modo de ejemplo no limitativo, "al menos uno de A y B" (o, equivalentemente, "al menos uno de A o B", o, equivalentemente, "al menos uno de A y/o B") puede referirse, en una realización, a al menos uno, incluyendo opcionalmente más de uno, A, sin que B esté presente (y opcionalmente incluyendo elementos distintos de B); en otra realización, a al menos uno, incluyendo opcionalmente más de uno, B, sin que A esté presente (y opcionalmente incluyendo elementos distintos de A); en aún otra realización, a al menos uno, incluyendo opcionalmente más de uno, A, y al menos uno, incluyendo opcionalmente más de uno, B (y opcionalmente incluyendo otros elementos); etc.

REIVINDICACIONES

1. Método de determinación de la posición y/o el número de una pluralidad de dianas de ácido nucleico dentro de una célula u otra muestra usando una pluralidad de sondas de codificación de ácido nucleico primarias y una pluralidad de sondas de lectura de ácido nucleico secundarias, comprendiendo el método:
 - (i) exponer la célula u otra muestra a una pluralidad de sondas de codificación de ácido nucleico primarias, en donde las sondas de codificación de ácido nucleico primarias comprenden una primera porción que comprende una secuencia diana que se une de manera singular a una diana de ácido nucleico y una segunda porción que comprende una o más secuencias de lectura, en donde las secuencias de lectura se toman de un conjunto de secuencias ortogonales que tienen homología limitada unas con otras y con las especies de ácido nucleico en la célula u otra muestra y en donde la muestra comprende una pluralidad de las dianas de ácido nucleico a cada una de las cuales se les asigna una palabra de código en donde las palabras de código comprenden varias posiciones y valores para cada posición, y las palabras de código forman un código de comprobación de error y/o corrección de error;
 - (ii) determinar la unión de las sondas de codificación de ácido nucleico primarias a las dianas de ácido nucleico dentro de la célula u otra muestra mediante
 - (a) exponer la célula u otra muestra a sondas de lectura de ácido nucleico secundarias que comprenden, cada una, una entidad de señalización y una secuencia que es complementaria a una de las secuencias de lectura de las sondas de codificación de ácido nucleico primarias de manera que las sondas de lectura de ácido nucleico secundarias se hibridan con la secuencia de lectura en al menos parte de las sondas de ácido nucleico primarias y
 - (b) determinar la unión de las sondas de codificación de ácido nucleico primarias determinando la fluorescencia de cada una de las sondas de lectura de ácido nucleico secundarias dentro de la célula u otra muestra mediante obtención de imágenes de fluorescencia de la célula u otra muestra;
 - (iii) repetir las etapas (a) y (b) una o más veces usando una sonda de lectura de ácido nucleico secundaria diferente para al menos parte de las repeticiones;
 - (iv) crear las palabras de código basándose en la fluorescencia de las sondas de lectura de ácido nucleico secundarias; y
 - (v) para al menos parte de las palabras de código, hacer coincidir la palabra de código con una palabra de código válida asociada con la pluralidad de dianas de ácido nucleico en donde, si no se encuentra coincidencia o se encuentra coincidencia cuyo error puede corregirse, aplicar corrección de error a la palabra de código para formar una palabra de código válida.
2. Método según la reivindicación 1, en donde la primera entidad de señalización es (a) fluorescente, y/o (b) una proteína, y/o (c) un colorante, y/o (d) una nanopartícula.
3. Método según la reivindicación 1 ó 2, que comprende exponer la muestra a al menos 5 sondas de codificación de ácido nucleico primarias diferentes, preferiblemente al menos 10 sondas de codificación de ácido nucleico primarias diferentes, preferiblemente al menos 100 sondas de codificación de ácido nucleico primarias diferentes.
4. Método según una cualquiera de las reivindicaciones 1-3, que comprende exponer la muestra a una pluralidad de sondas de codificación de ácido nucleico primarias simultáneamente.
5. Método según una cualquiera de las reivindicaciones 1-3, que comprende exponer la muestra a una pluralidad de sondas de codificación de ácido nucleico primarias secuencialmente.
6. Método según una cualquiera de las reivindicaciones 1-5, en donde la pluralidad de sondas de codificación de ácido nucleico primarias comprende una combinación combinatoria de sondas de ácido nucleico con secuencias diferentes, preferiblemente en donde la combinación combinatoria de sondas de ácido nucleico se dirige a una combinación combinatoria de especies de ARN y/o secuencias de ADN en la muestra.
7. Método según una cualquiera de las reivindicaciones 1-6, en donde la pluralidad de sondas de codificación de ácido nucleico primarias comprende sondas de ácido nucleico distinguibles formadas a partir de la combinación combinatoria de una o más secuencias de lectura y que se toman de la una o más secuencias de lectura, preferiblemente en donde:

- (a) la secuencia diana es sustancialmente complementaria a una secuencia de ácidos nucleicos que codifica para una proteína; y/o
- (b) la secuencia diana es sustancialmente complementaria a ARN mensajero (ARNm);
- (c) la pluralidad de sondas de ácido nucleico comprende al menos 8 secuencias de lectura posibles; y/o
- (d) la pluralidad de sondas de ácido nucleico comprende no más de 32 secuencias de lectura posibles, preferiblemente no más de 16 secuencias de lectura posibles, preferiblemente no más de 8 secuencias de lectura posibles; y/o
- (e) la pluralidad de secuencias de lectura se distribuyen en la pluralidad de sondas de ácido nucleico para definir un código de detección de error y/o corrección de error; y/o
- (f) la secuencia diana de la pluralidad de sondas de ácido nucleico tiene una longitud promedio de entre 10 y 200 nucleótidos; y/o
- (g) la pluralidad de secuencias de lectura tienen una longitud promedio de entre 5 nucleótidos y 50 nucleótidos; y/o
- (h) al menos parte de la pluralidad de sondas de ácido nucleico comprenden no más de 10 secuencias de lectura, preferiblemente no más de 6 secuencias de lectura, preferiblemente no más de 4 secuencias de lectura, preferiblemente no más de 3 secuencias de lectura, preferiblemente no más de 2 secuencias de lectura.
8. Método según una cualquiera de las reivindicaciones 1-7, que comprende además inactivar una primera entidad de señalización antes de exponer la muestra a una segunda sonda de lectura de ácido nucleico secundaria, preferiblemente que comprende inactivar la primera entidad de señalización mediante:
- (a) fotoblanquear al menos parte de la primera entidad de señalización; y/o
- (b) blanquear químicamente al menos parte de la primera entidad de señalización; y/o
- (c) exponer la primera entidad de señalización a un reactante capaz de alterar la estructura de la entidad de señalización; y/o
- (d) eliminar al menos parte de la primera entidad de señalización; y/o
- (e) disociar la primera entidad de señalización a partir de la primera sonda de lectura; y/o
- (f) disociar la primera sonda de lectura que contiene la primera entidad de señalización a partir de la muestra; y/o
- (g) escindir químicamente la primera entidad de señalización a partir de la primera sonda de lectura; y/o
- (h) escindir enzimáticamente la primera entidad de señalización a partir de la primera sonda de lectura; y/o
- (i) exponer la entidad de señalización o la primera sonda de lectura a una enzima.
9. Método según una cualquiera de las reivindicaciones 1-8, que comprende determinar el centroide de la entidad de señalización usando un algoritmo para determinar emisores individuales no solapantes y/o emisores individuales parcialmente solapantes, preferiblemente que comprende determinar el centroide usando:
- (a) un algoritmo de probabilidad máxima; y/o
- (b) un algoritmo de mínimos cuadrados; y/o
- (c) un algoritmo bayesiano; y/o
- (d) un algoritmo de detección comprimida.
10. Método según una cualquiera de las reivindicaciones 1-9, que comprende además determinar un nivel de

confianza para la diana de ácido nucleico identificada, preferiblemente usando una razón del número de coincidencias exactas con respecto al número de coincidencias que tienen (a) uno o más errores de un bit o (b) exactamente un error de un bit con respecto a la palabra de código.

- 5 11. Método según las reivindicaciones 1-10, en donde al menos parte de la pluralidad de sondas de ácido nucleico:
 - (a) comprenden ADN; y/o
 - 10 (b) comprenden ARN; y/o
 - (c) comprenden ANP; y/o
 - (d) tienen una longitud promedio de entre 10 y 300 nucleótidos; y/o
 - 15 (e) están configuradas para unirse a un ácido nucleico dentro de la muestra, preferiblemente en donde al menos parte de la unión de las sondas de ácido nucleico a una diana dentro de la muestra es unión específica y/o es a través de apareamiento de bases de Watson-Crick; y/o
 - 20 (f) están configuradas para unirse a ARN; y/o
 - (g) están configuradas para unirse a ARN no codificante; y/o
 - (h) están configuradas para unirse a ARNm; y/o
 - 25 (i) están configuradas para unirse a ARN de transferencia (ARNt); y/o
 - (j) están configuradas para unirse a ARN ribosómico (ARNr); y/o
 - 30 (k) están configuradas para unirse a ARNlnc; y/o
 - (l) están configuradas para unirse a ARNnop; y/o
 - 35 (m) están configuradas para unirse a ADN; y/o
 - (n) están configuradas para unirse a ADN genómico.
12. Método según una cualquiera de las reivindicaciones 1-11, que comprende determinar la unión de las sondas de codificación de ácido nucleico primarias dentro de la célula u otra muestra a una resolución mejor que 300 nm, preferiblemente mejor que 100 nm, preferiblemente mejor que 80 nm, preferiblemente mejor que 50 nm.
13. Método según una cualquiera de las reivindicaciones 1-12, en donde la célula es una célula humana, y/o la célula está fijada.
14. Método según una cualquiera de las reivindicaciones 1-13, que comprende determinar la unión de las sondas de codificación de ácido nucleico primarias mediante:
 - (a) obtención de imágenes de al menos una porción de la célula u otra muestra; y/o
 - 50 (b) uso de una técnica de obtención de imágenes óptica; y/o
 - (c) uso de una técnica de obtención de imágenes de fluorescencia; y/o
 - 55 (d) técnica de obtención de imágenes de fluorescencia multicolor; y/o
 - (e) técnica de obtención de imágenes de fluorescencia de superresolución, preferiblemente usando (i) microscopía de reconstrucción óptica estocástica (STORM); y/o (ii) microscopía de localización fotoactivada (PALM) o microscopía de localización por fotoactivación de fluorescencia (FPALM); y/o (iii) microscopía de agotamiento de emisión estimulada (STED); y/o (iv) microscopía de iluminación estructurada (SIM); y/o (v) microscopía de transición de fluorescencia ópticamente lineal saturada reversible (RESOLFT).
 - 60

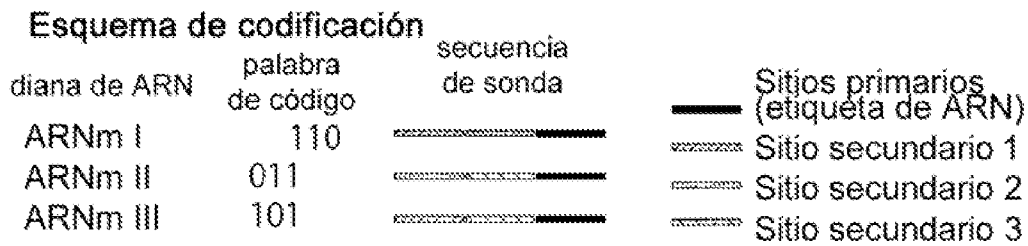


Fig. 1A

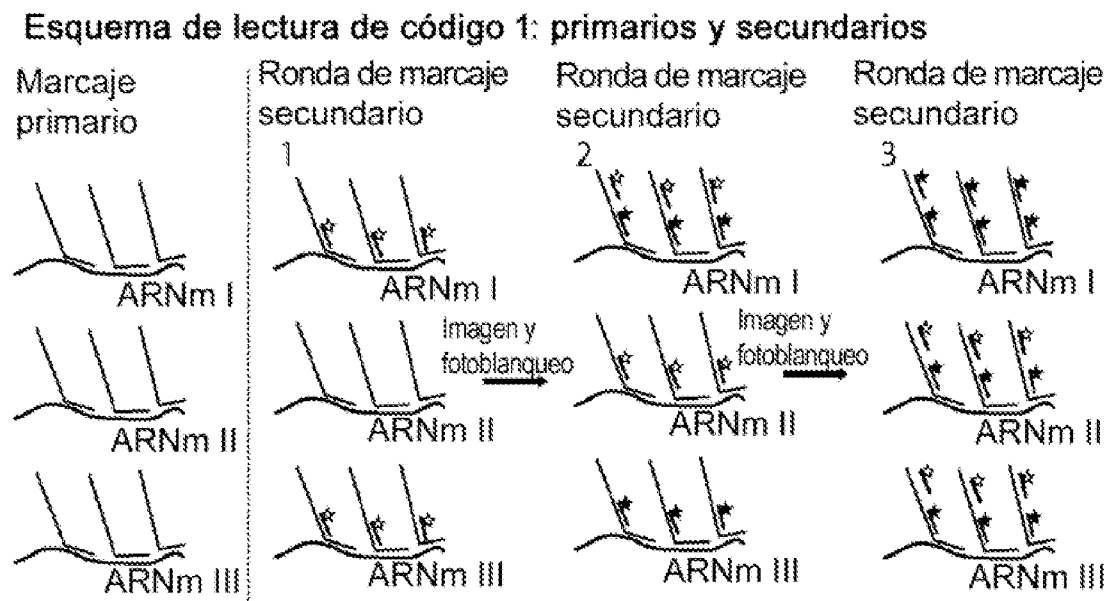


Fig. 1B

Esquema de lectura de código 2: sólo primarios

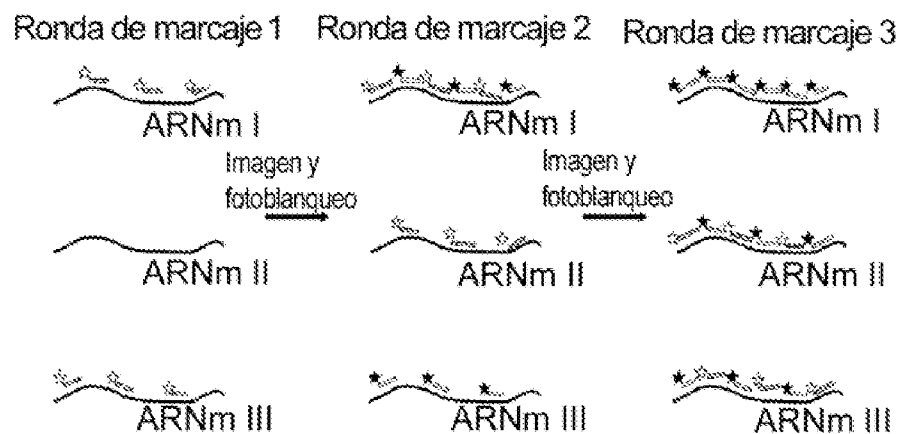


Fig. 1C

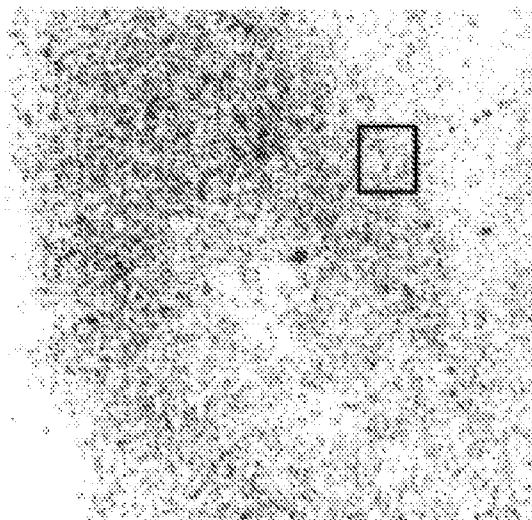


Fig. 2A

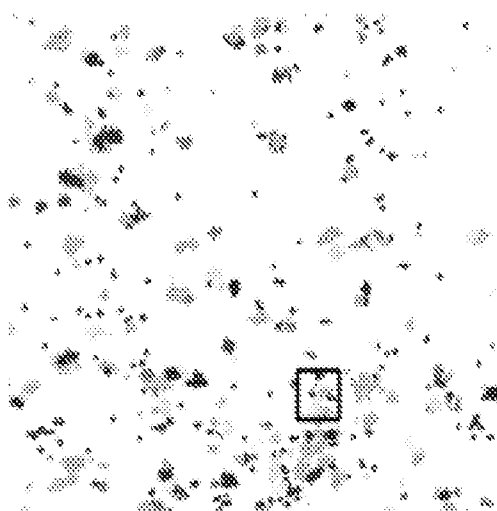


Fig. 2B



[0 1 0 1 1 1 0 0]

Fig. 2C

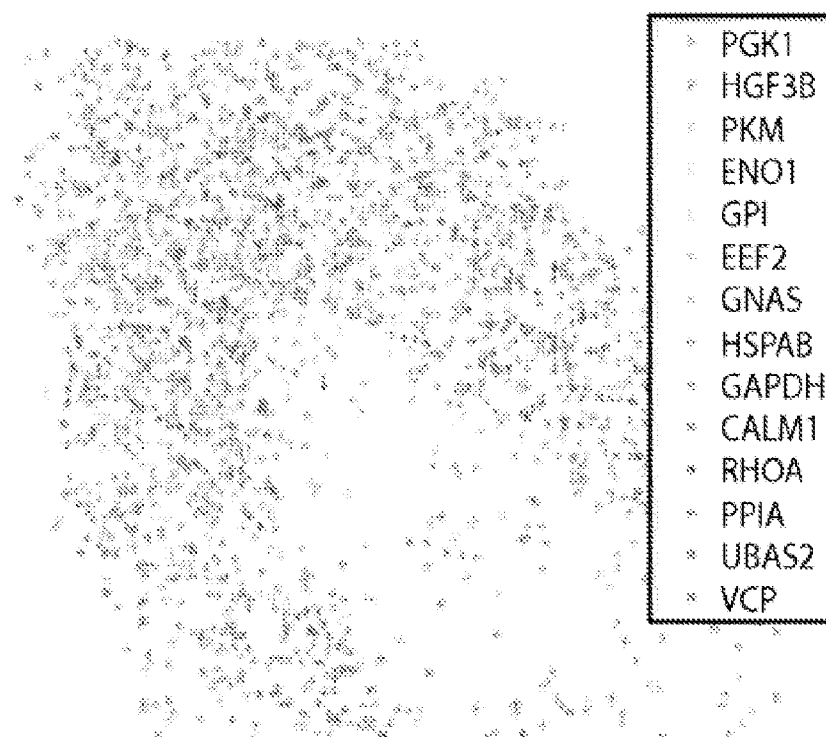


Fig. 2D

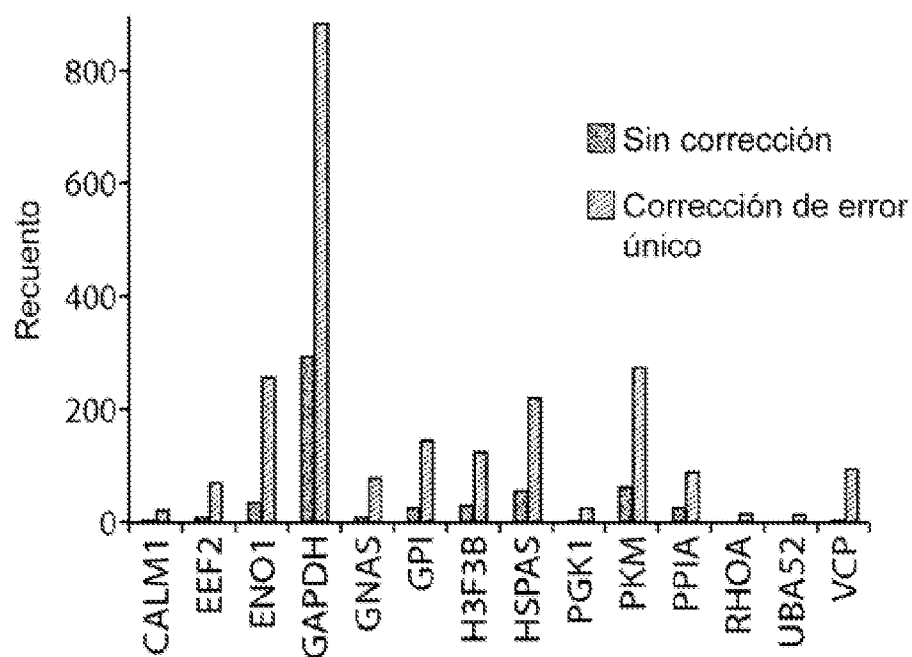


Fig. 2E

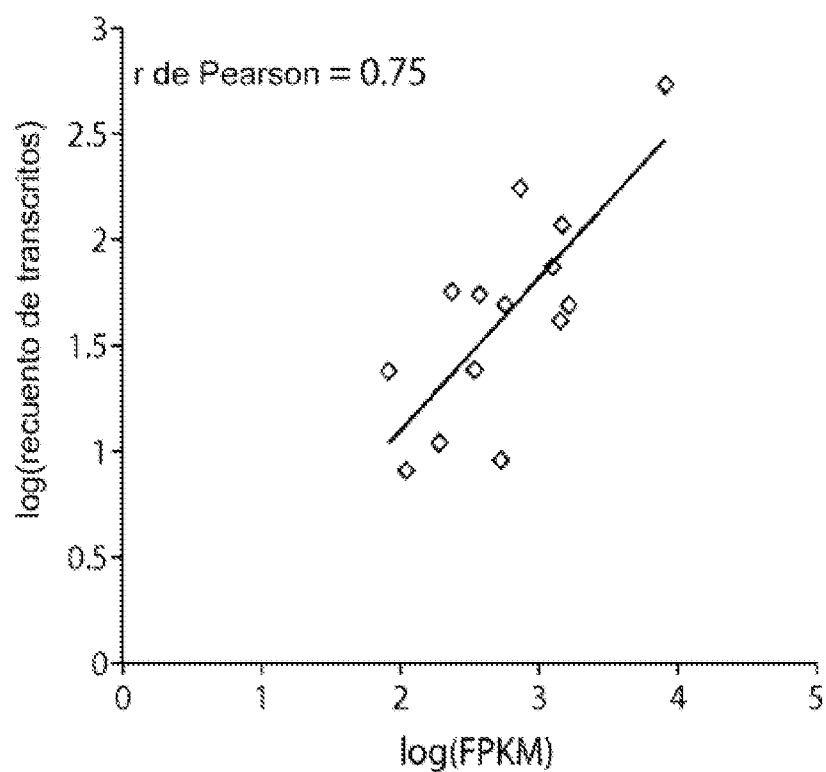


Fig. 2F

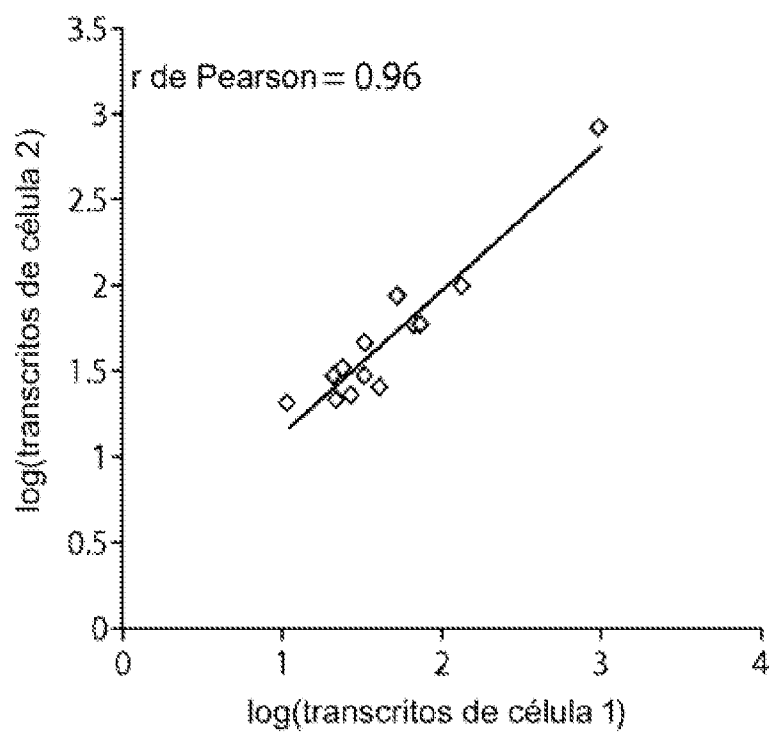


Fig. 2G

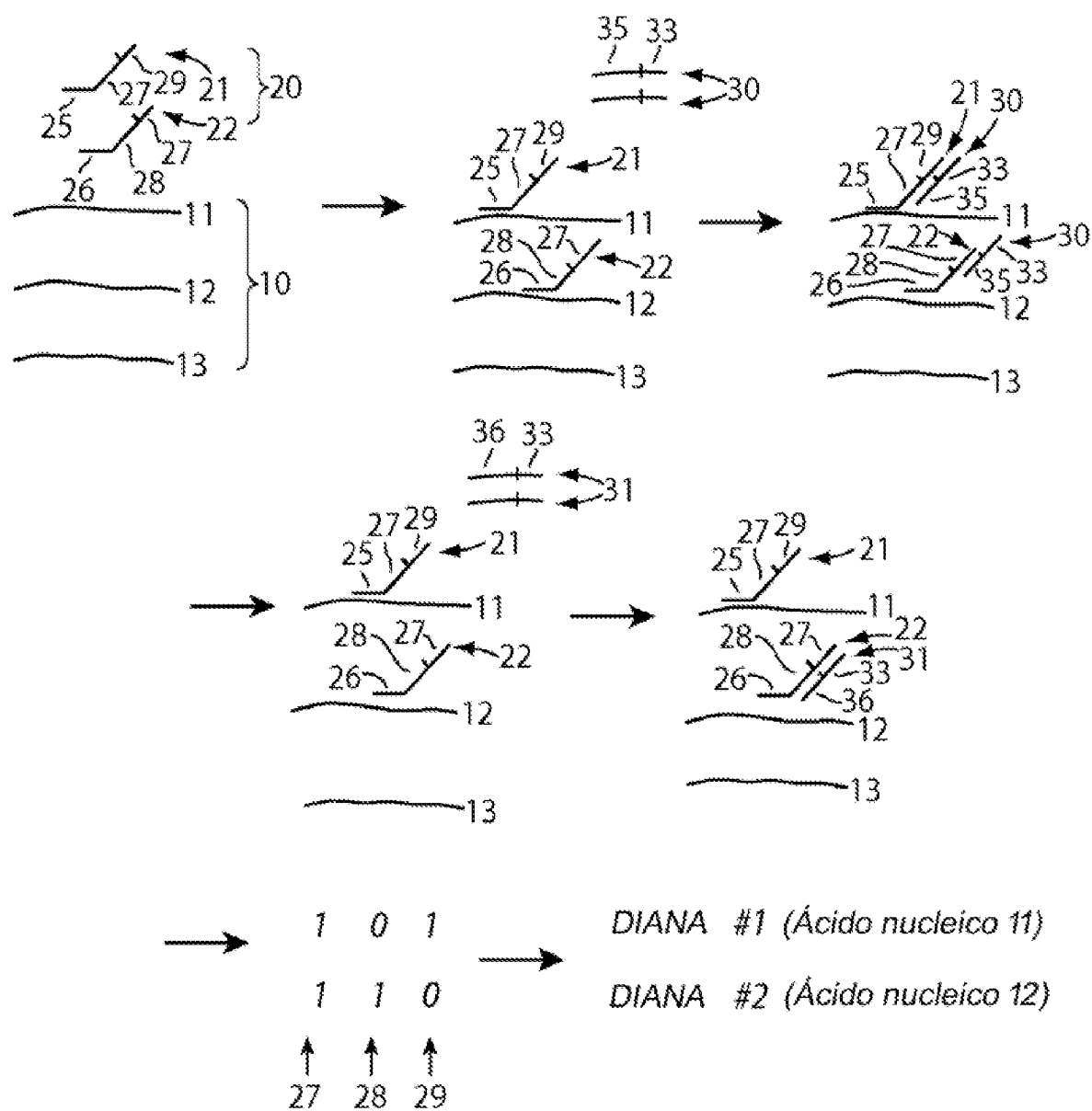


Fig. 3A

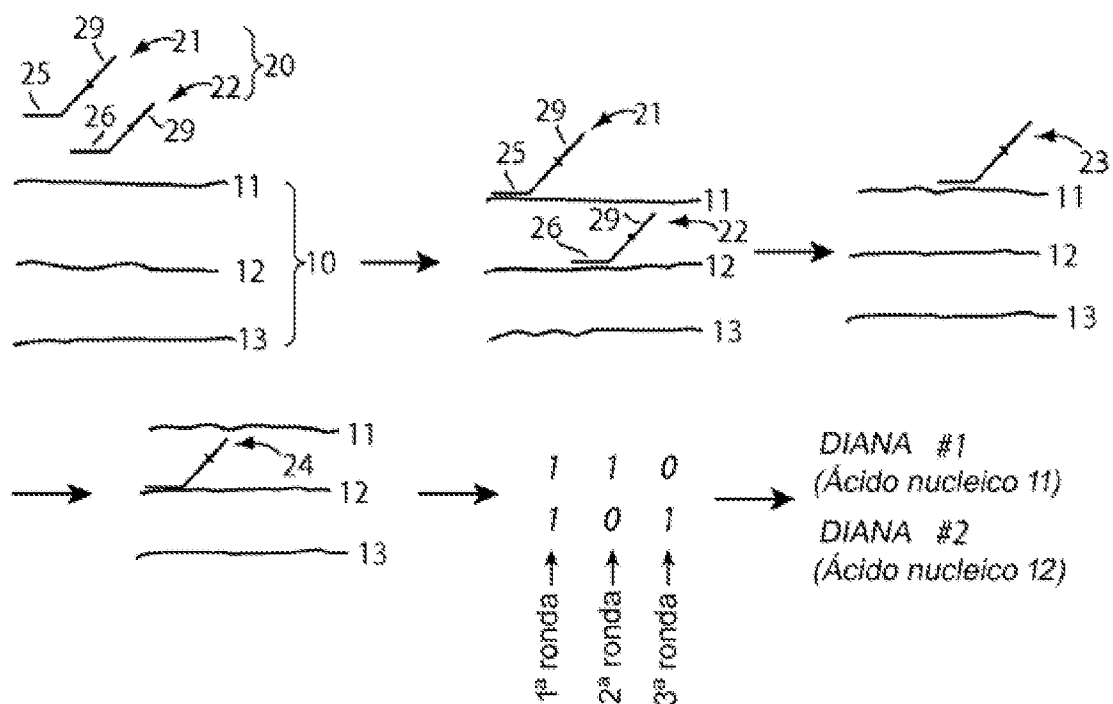


Fig. 3B

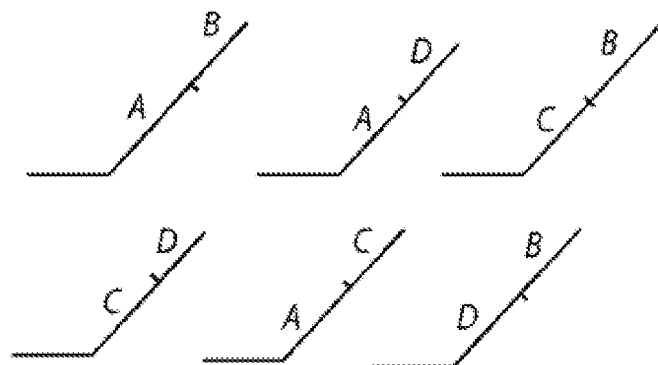


Fig. 4A

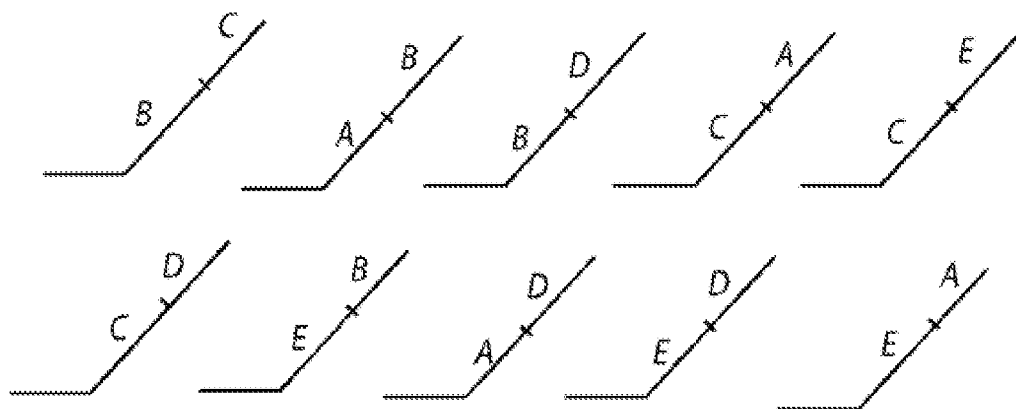


Fig. 4B

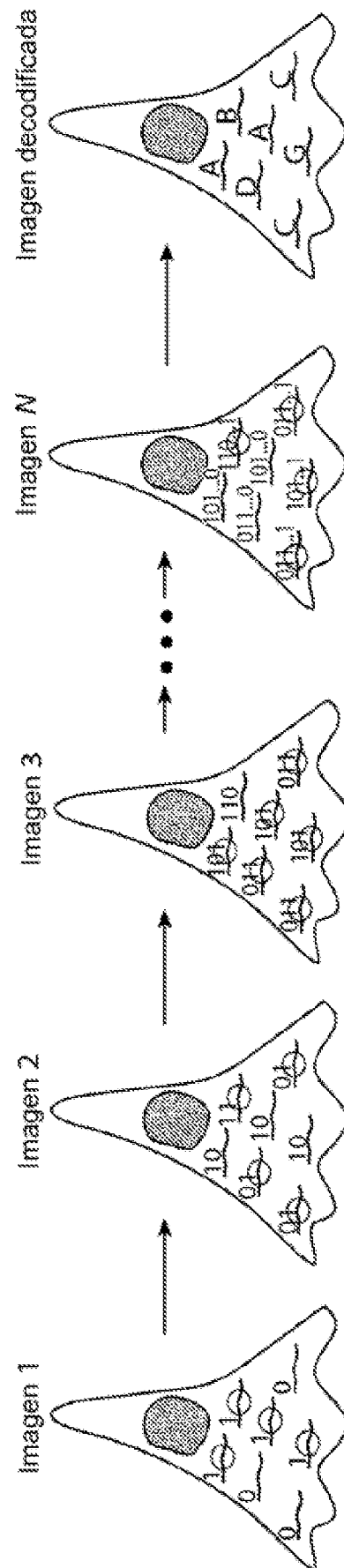


Fig. 5A

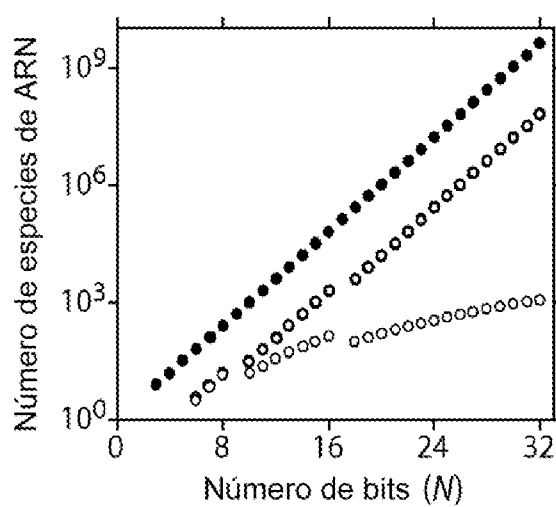


Fig. 5B

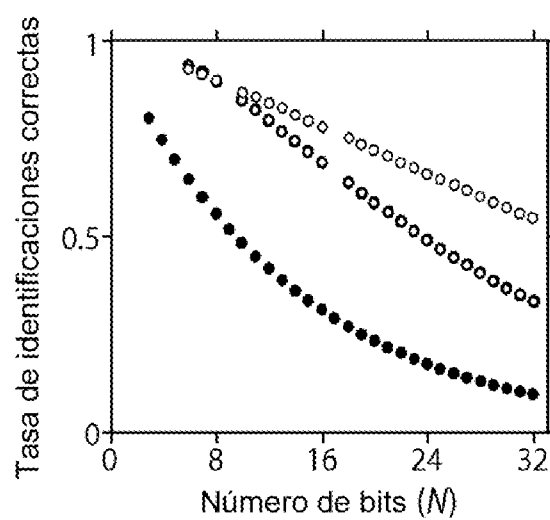


Fig. 5C

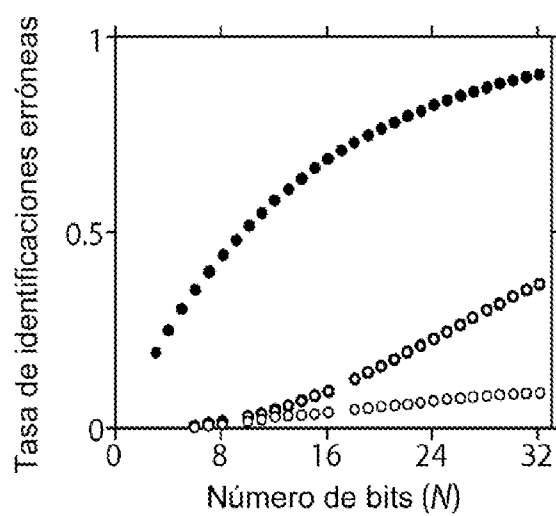


Fig. 5D

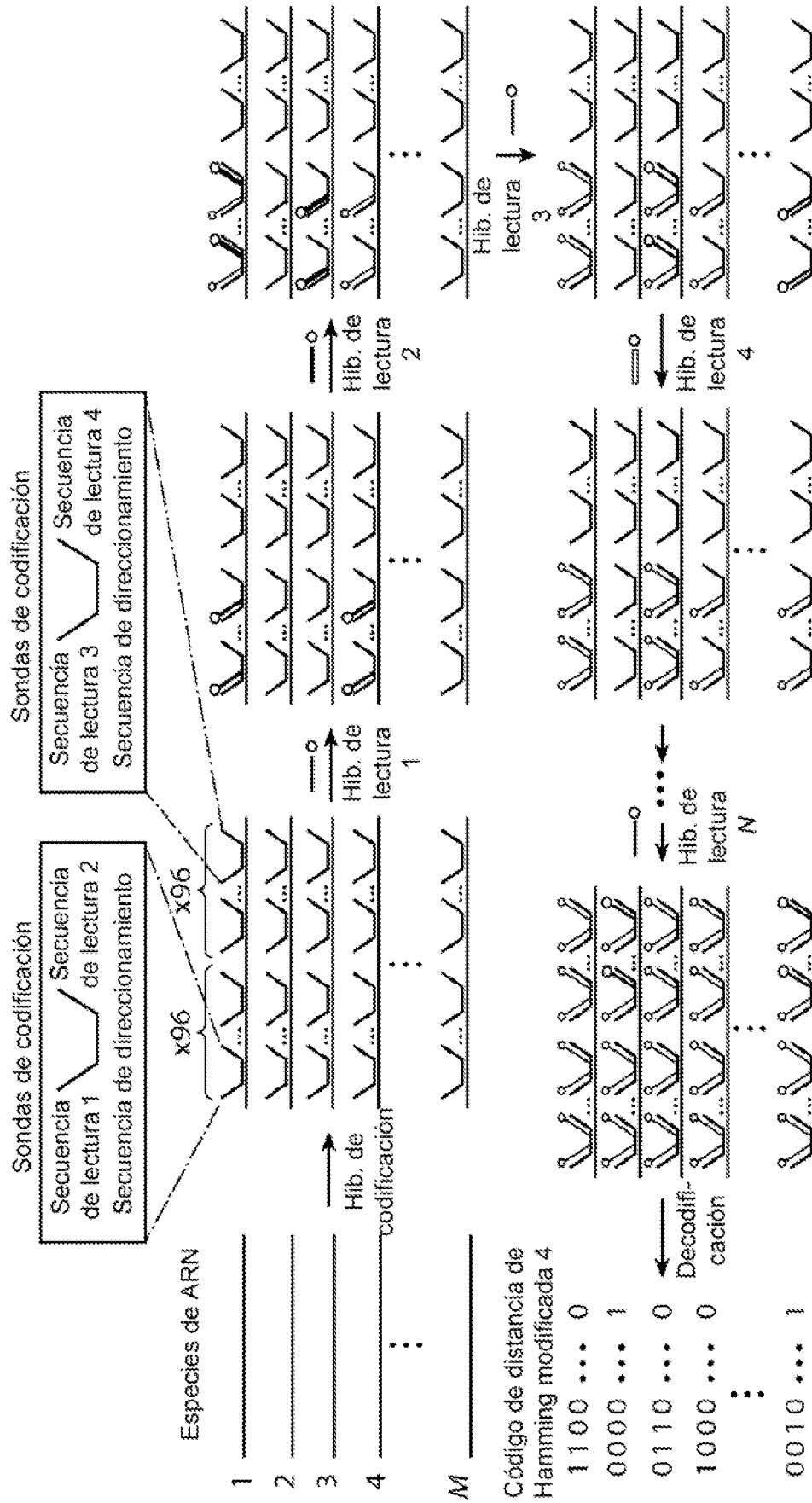


Fig. 5E

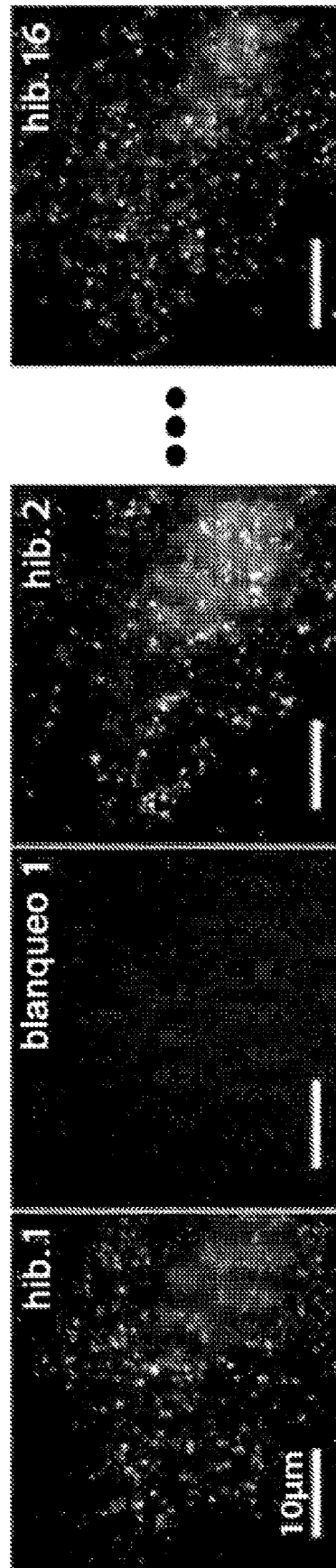


Fig. 6A

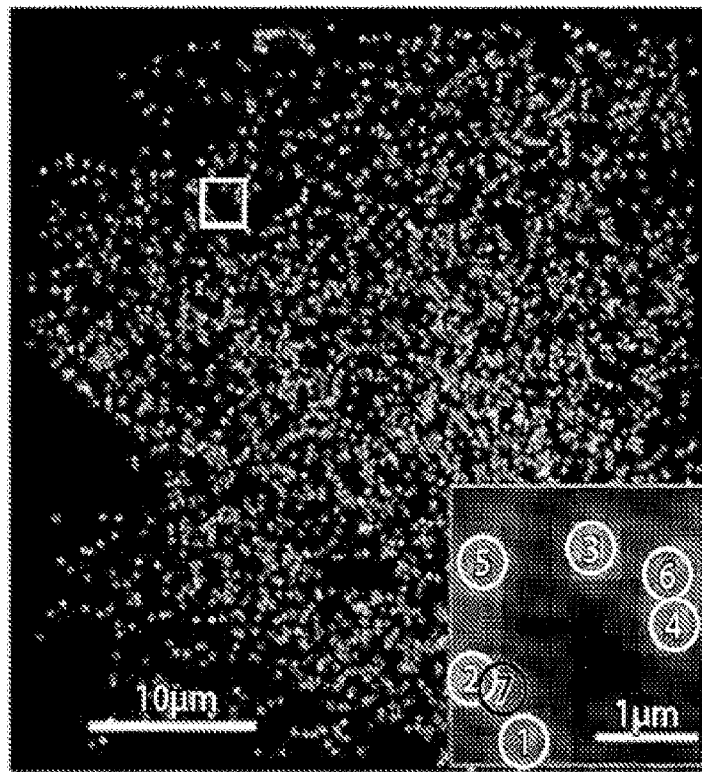


Fig. 6B

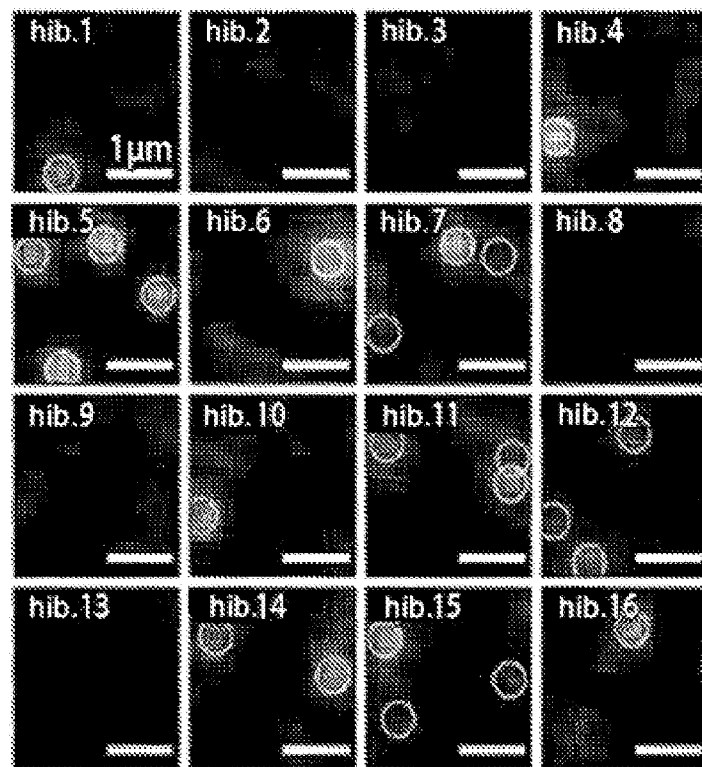


Fig. 6C

		Ronda de hibridación																Gen
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	
Número de puntos	1	1	0	0	0	1	0	0	0	0	0	0	1	0	1	0	0	DYNC1H1
	2	0	0	0	1	0	0	1	0	0	1	0	1	0	0	0	0	EGFR
	3	0	0	0	0	1	0	1	0	0	0	0	1	0	0	0	1	FLNA
	4	0	0	0	0	1	0	0	0	0	0	1	0	0	1	1	0	TLN1
	5	0	0	0	0	1	0	0	0	0	0	1	0	0	1	1	0	TLN1
	6	0	0	0	0	0	1	1	0	0	0	1	0	0	1	0	0	LRP1
	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	no identificado

Fig. 6D

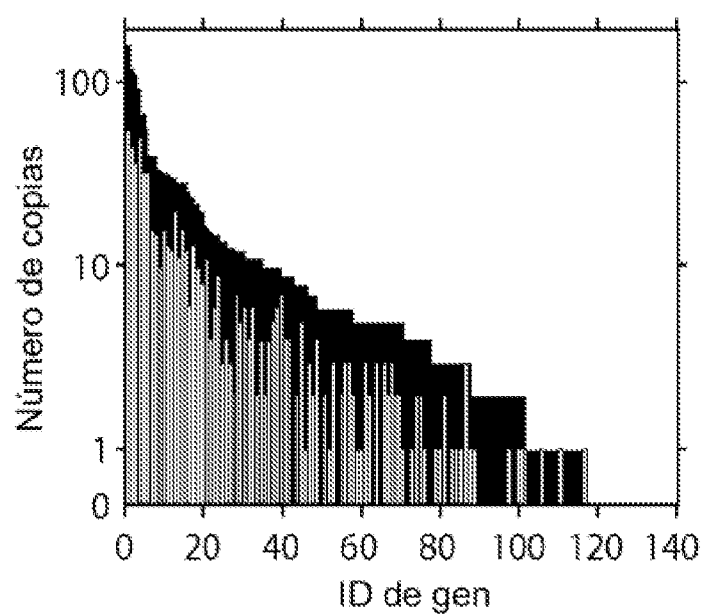


Fig. 6E

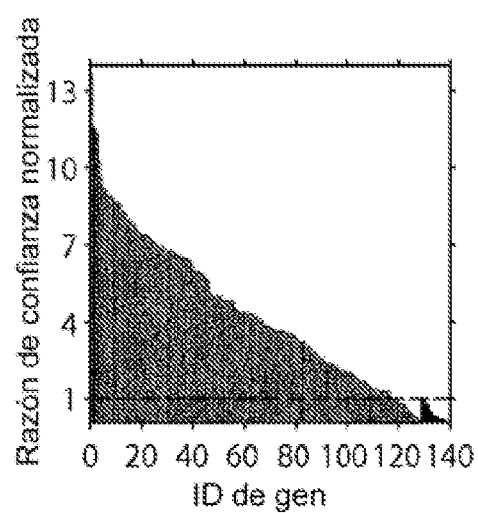


Fig. 6F

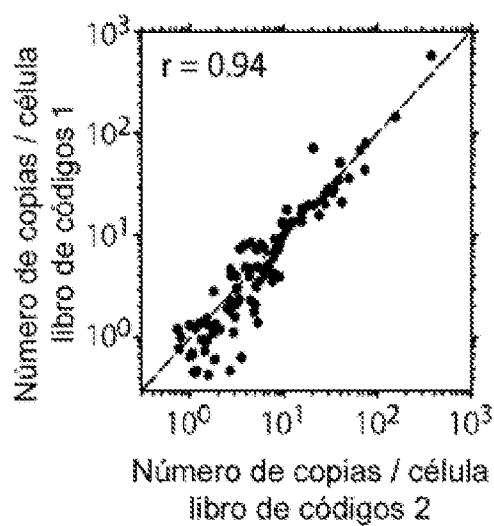


Fig. 6G

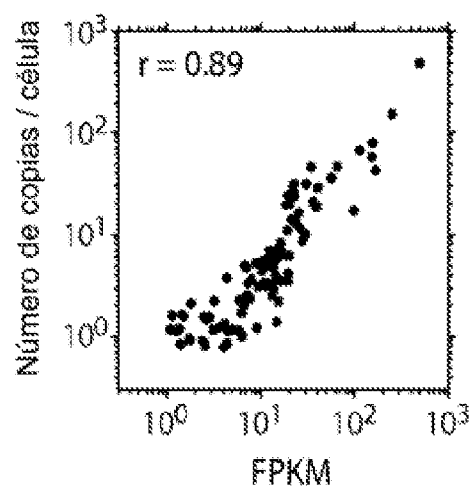


Fig. 6H

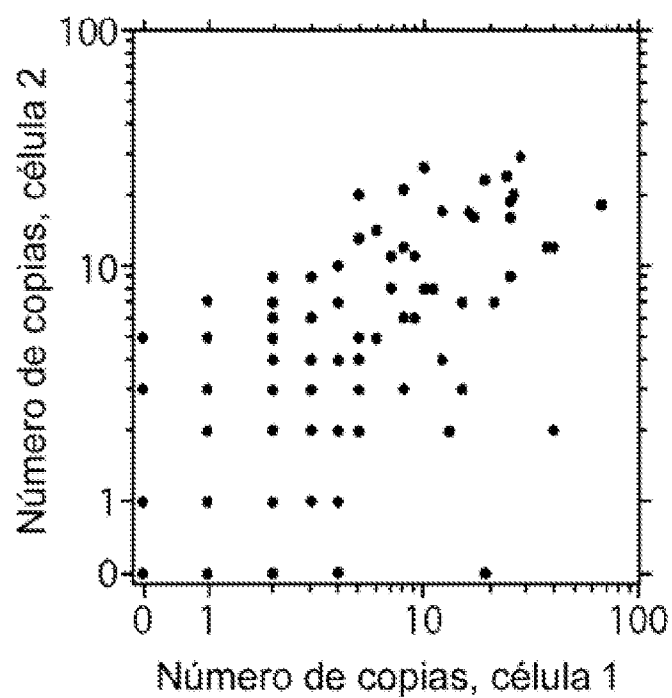


Fig. 7A

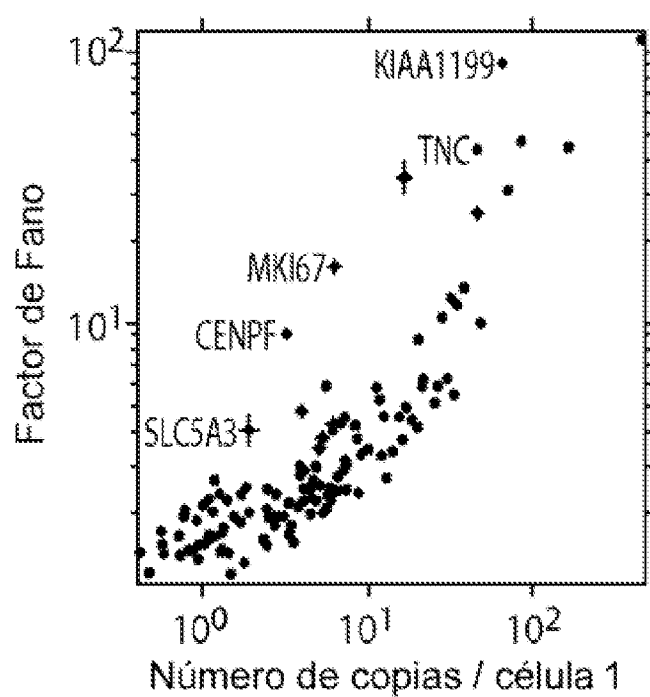


Fig. 7B

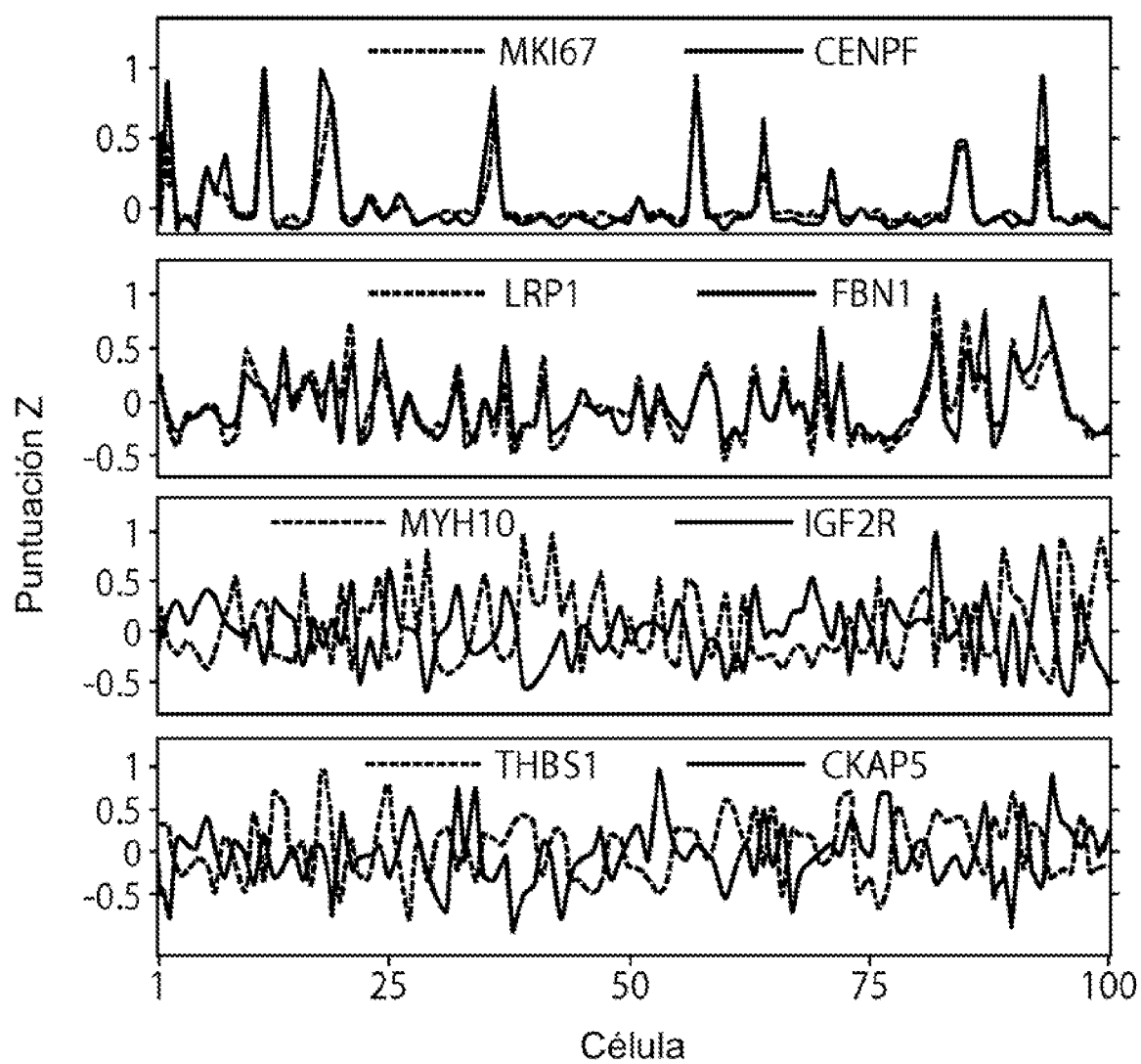


Fig. 7C

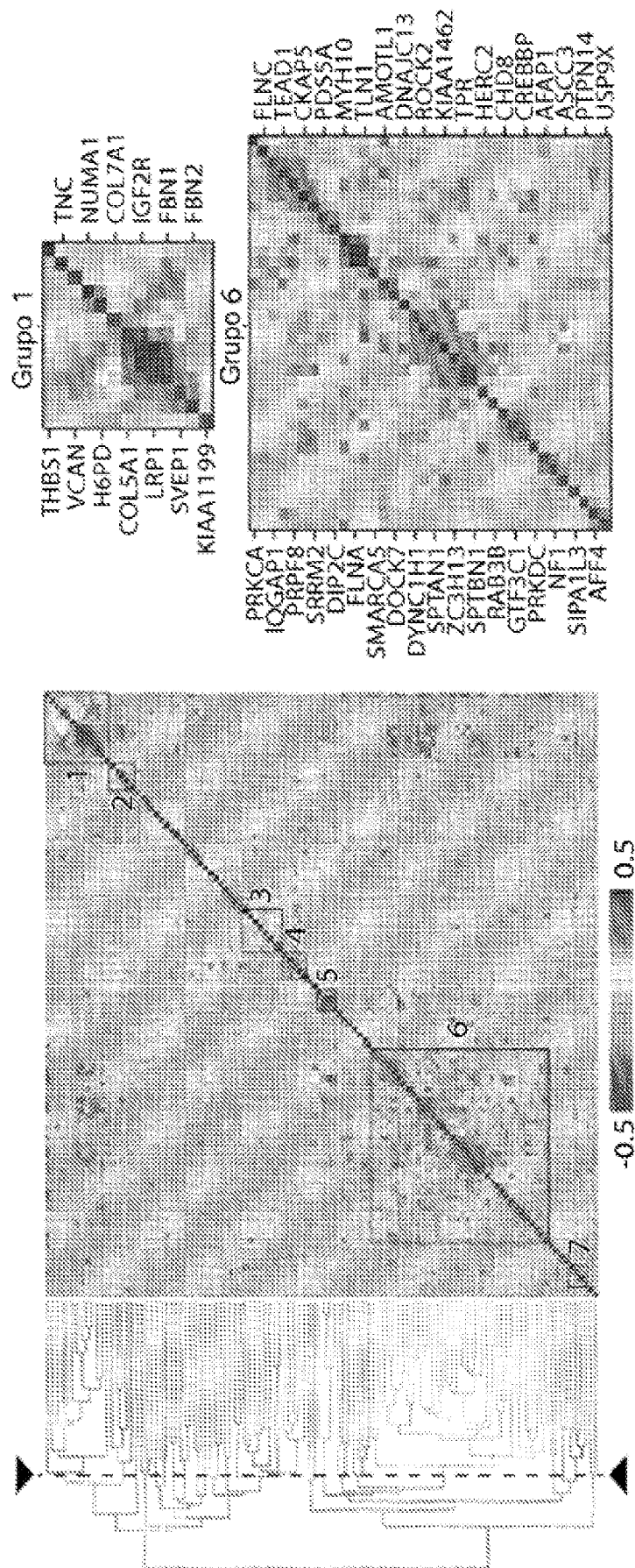


Fig. 7D

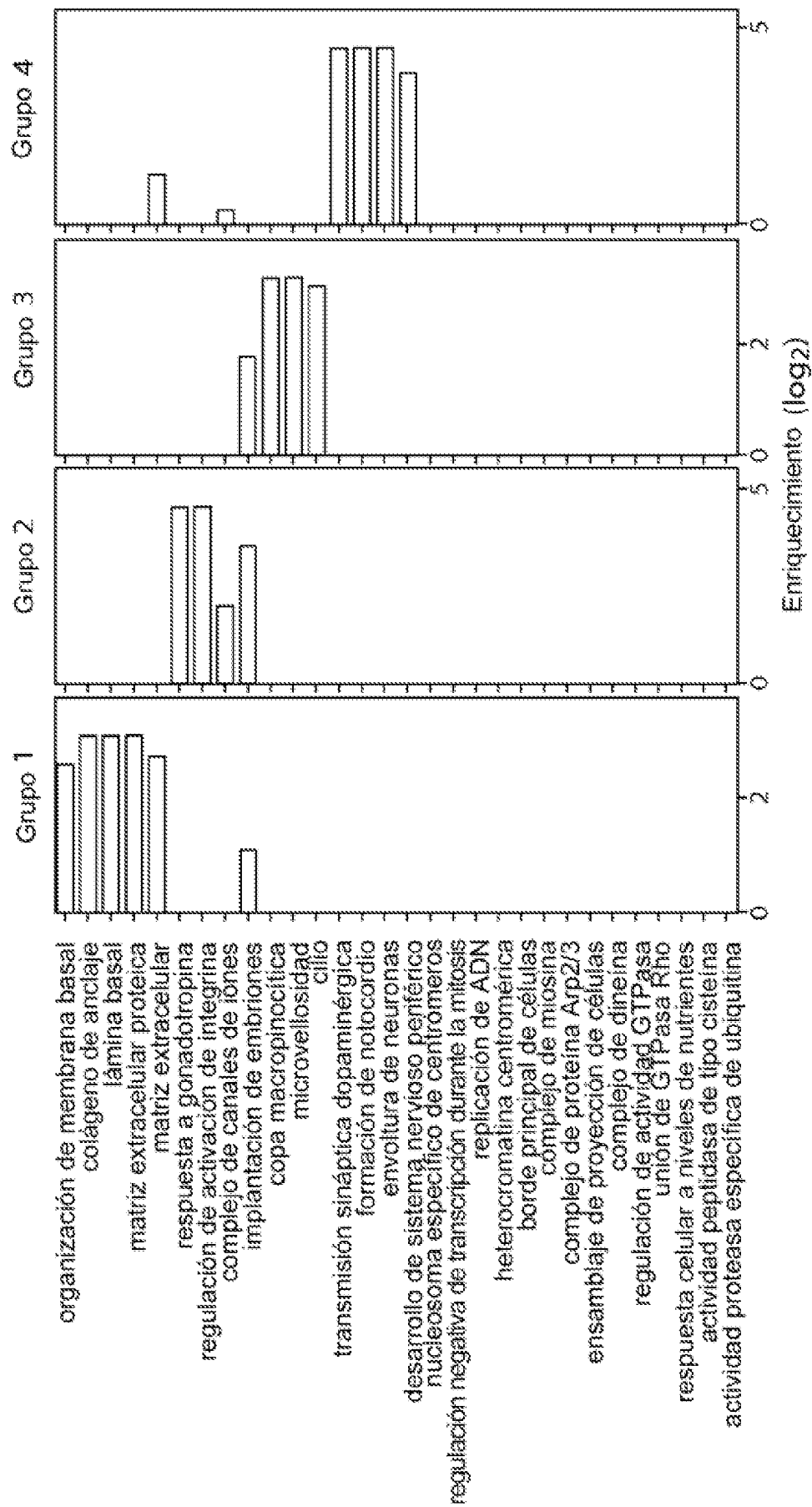


Fig. 7E

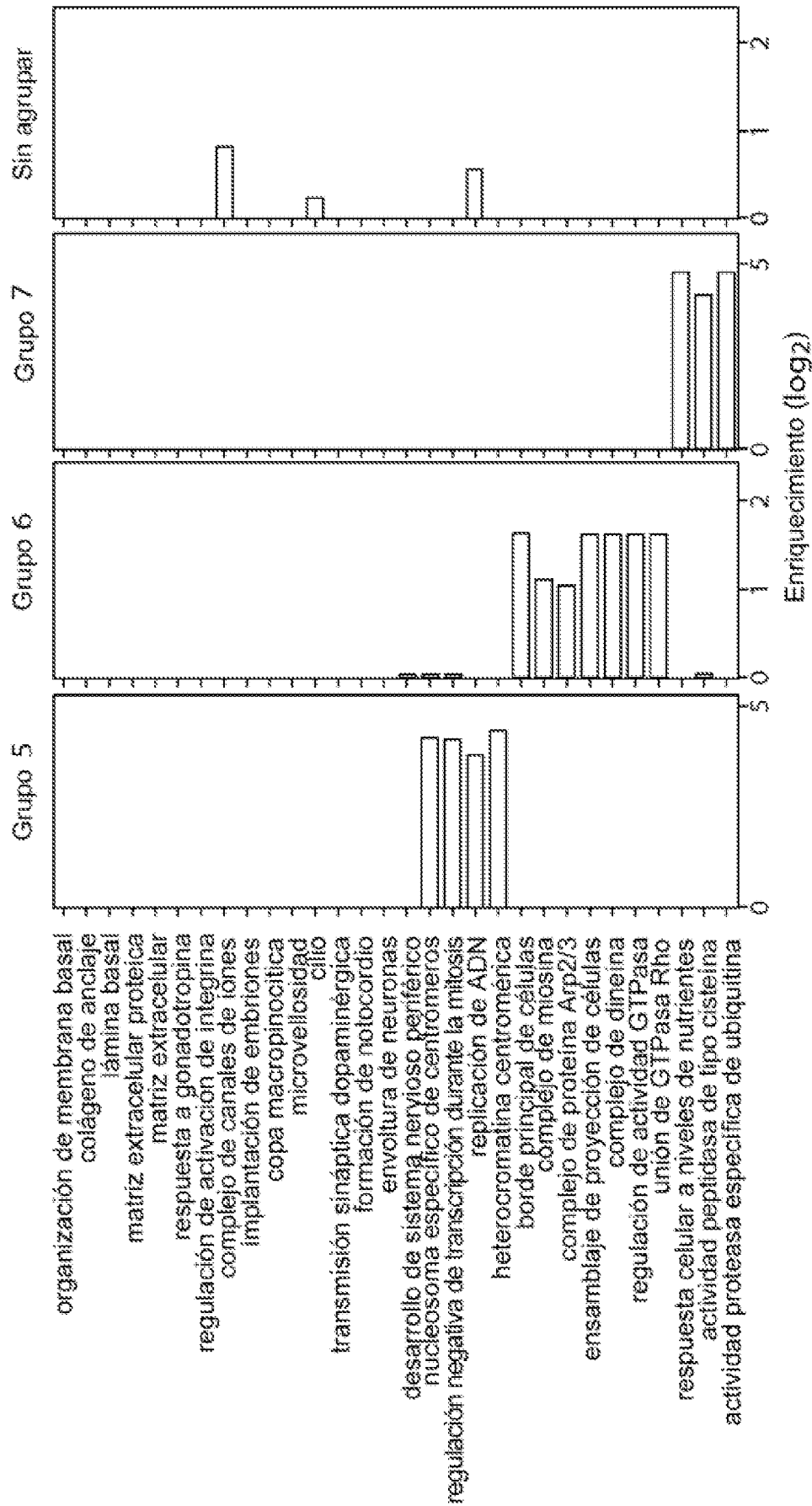


Fig. 7E Continuación

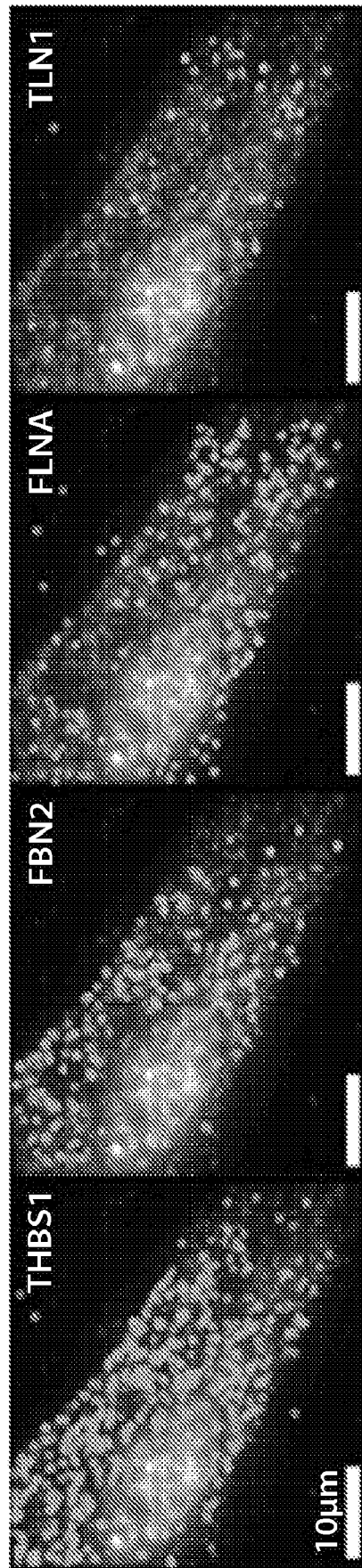


Fig. 8A

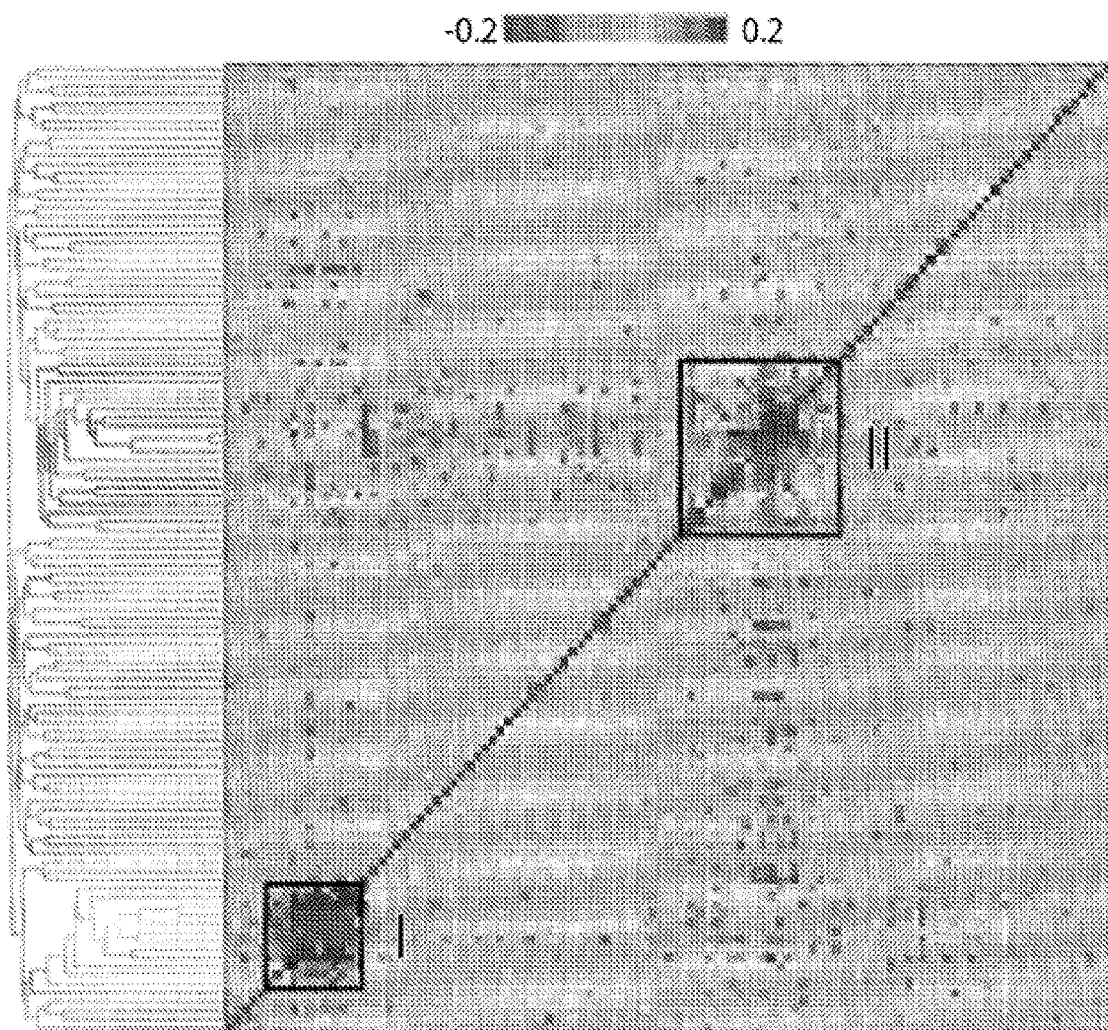


Fig. 8B

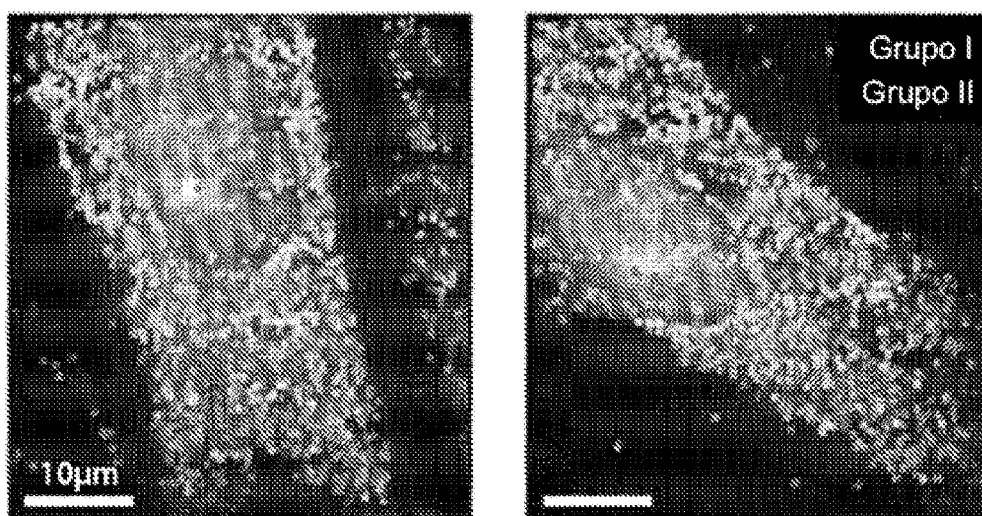


Fig. 8C

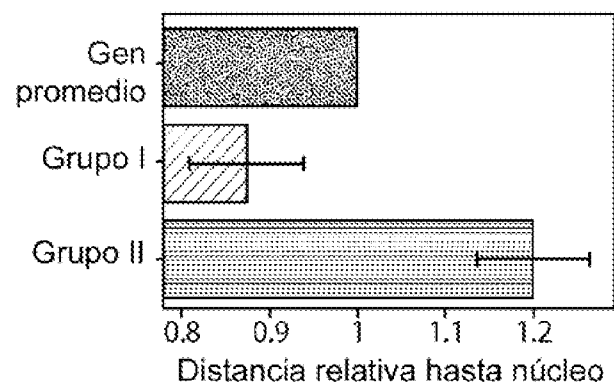
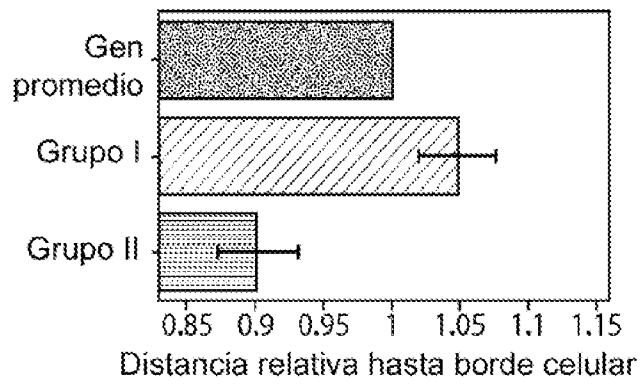


Fig. 8D

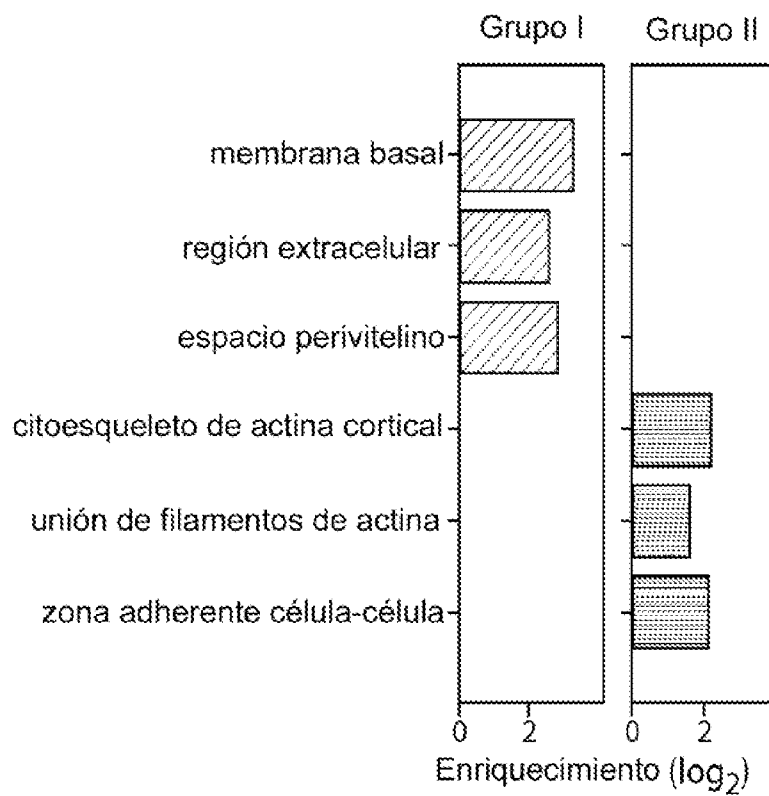


Fig. 8E

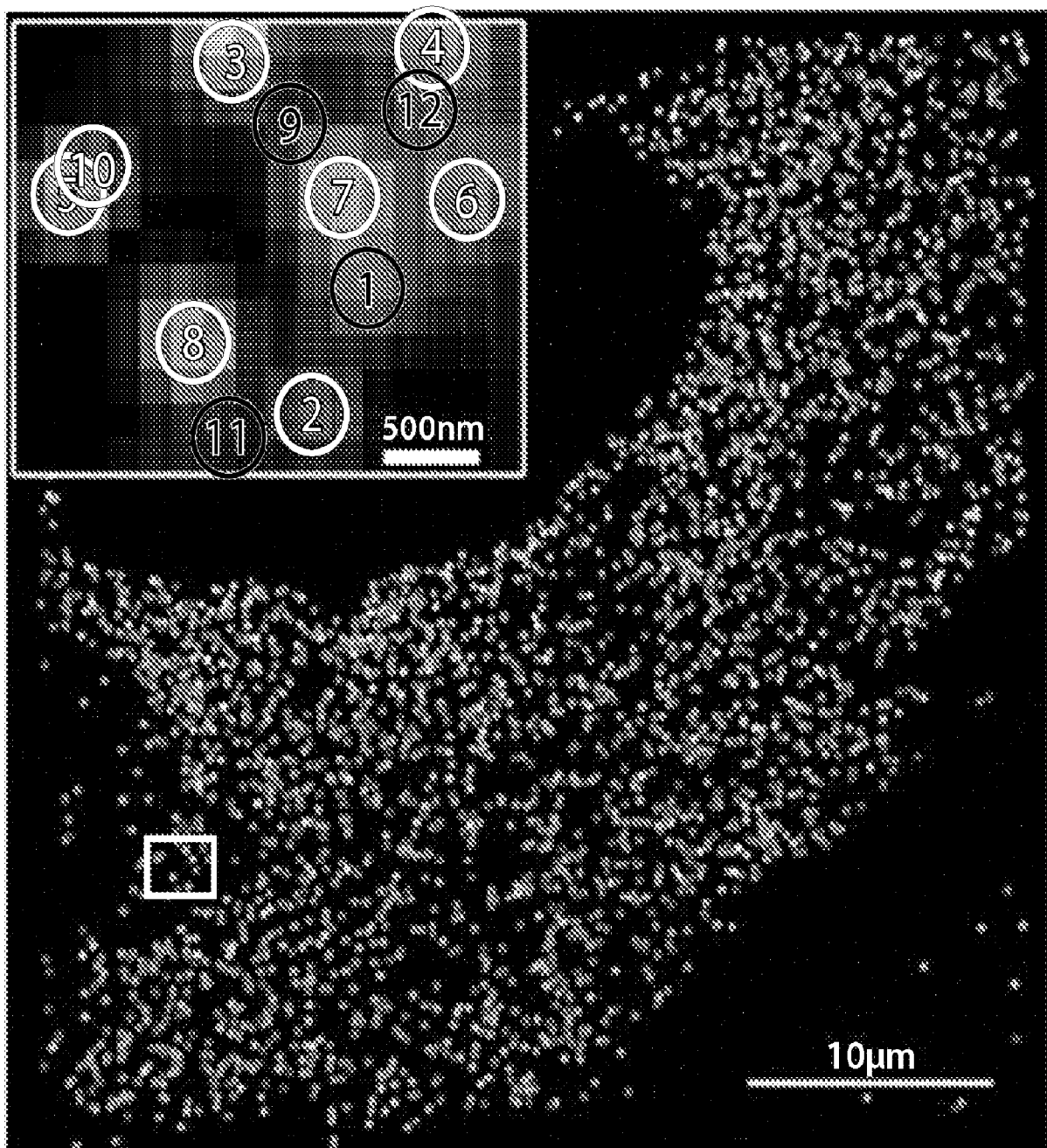


Fig. 9A

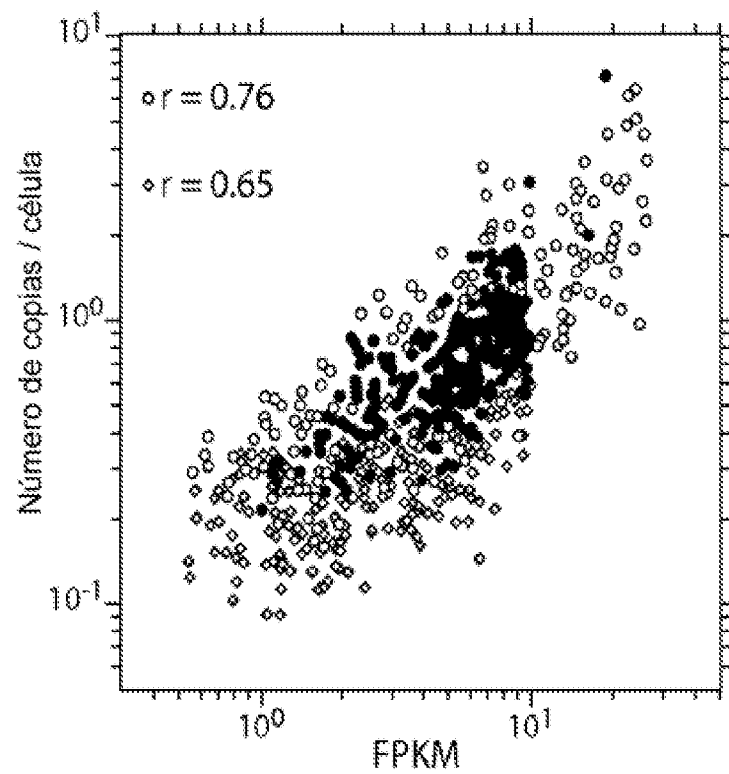


Fig. 9B

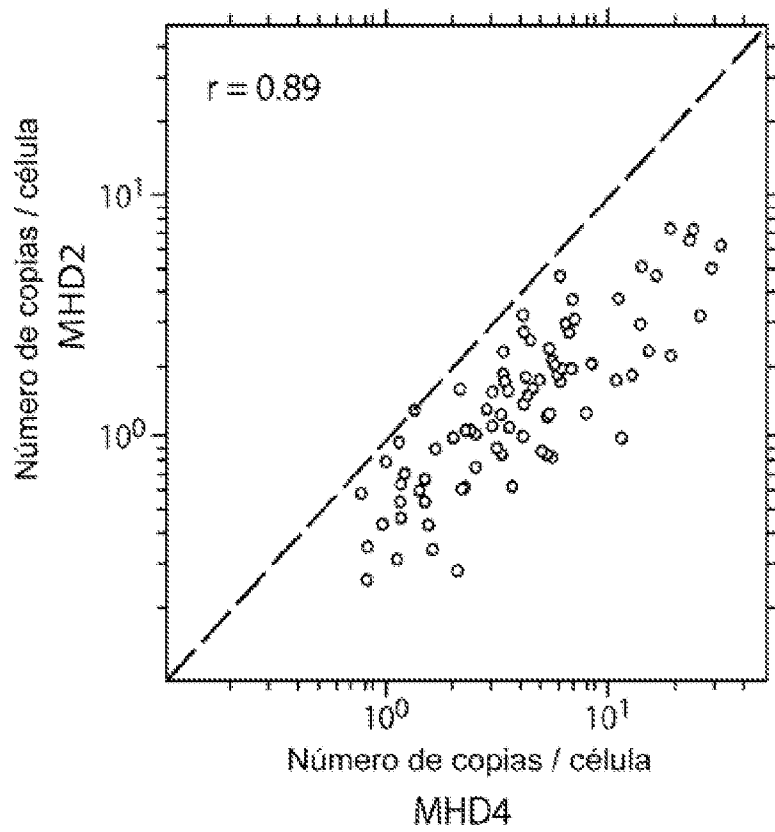


Fig. 9C

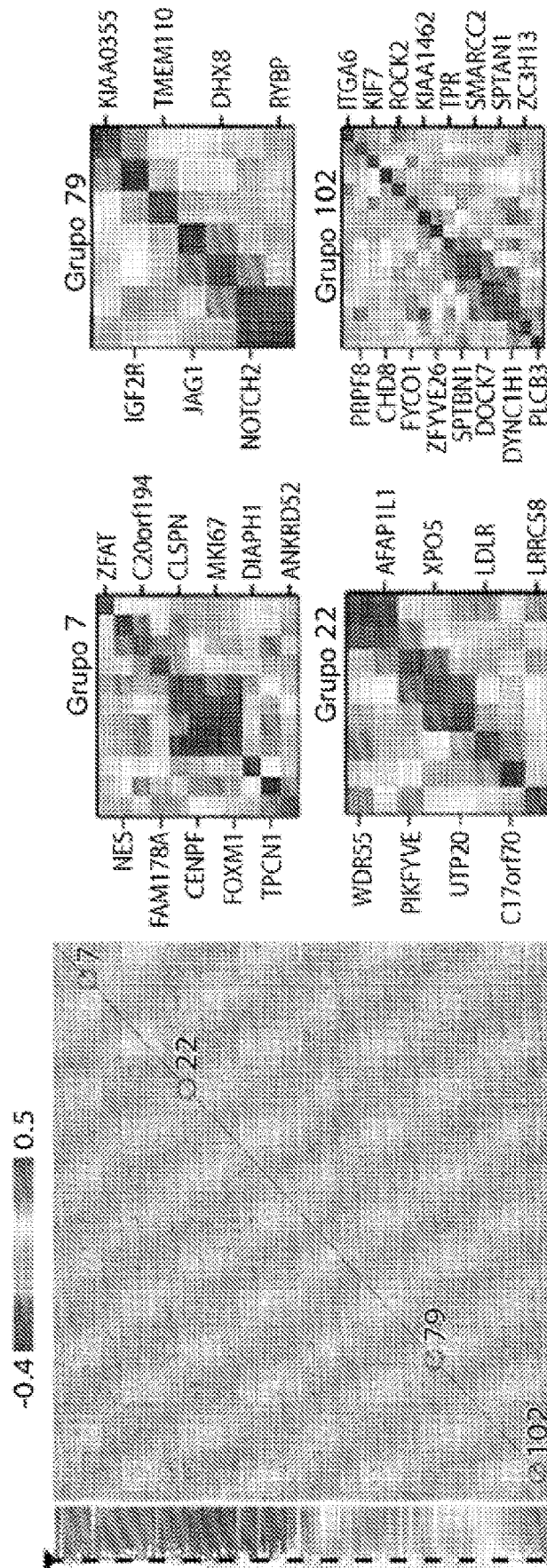


Fig. 10A

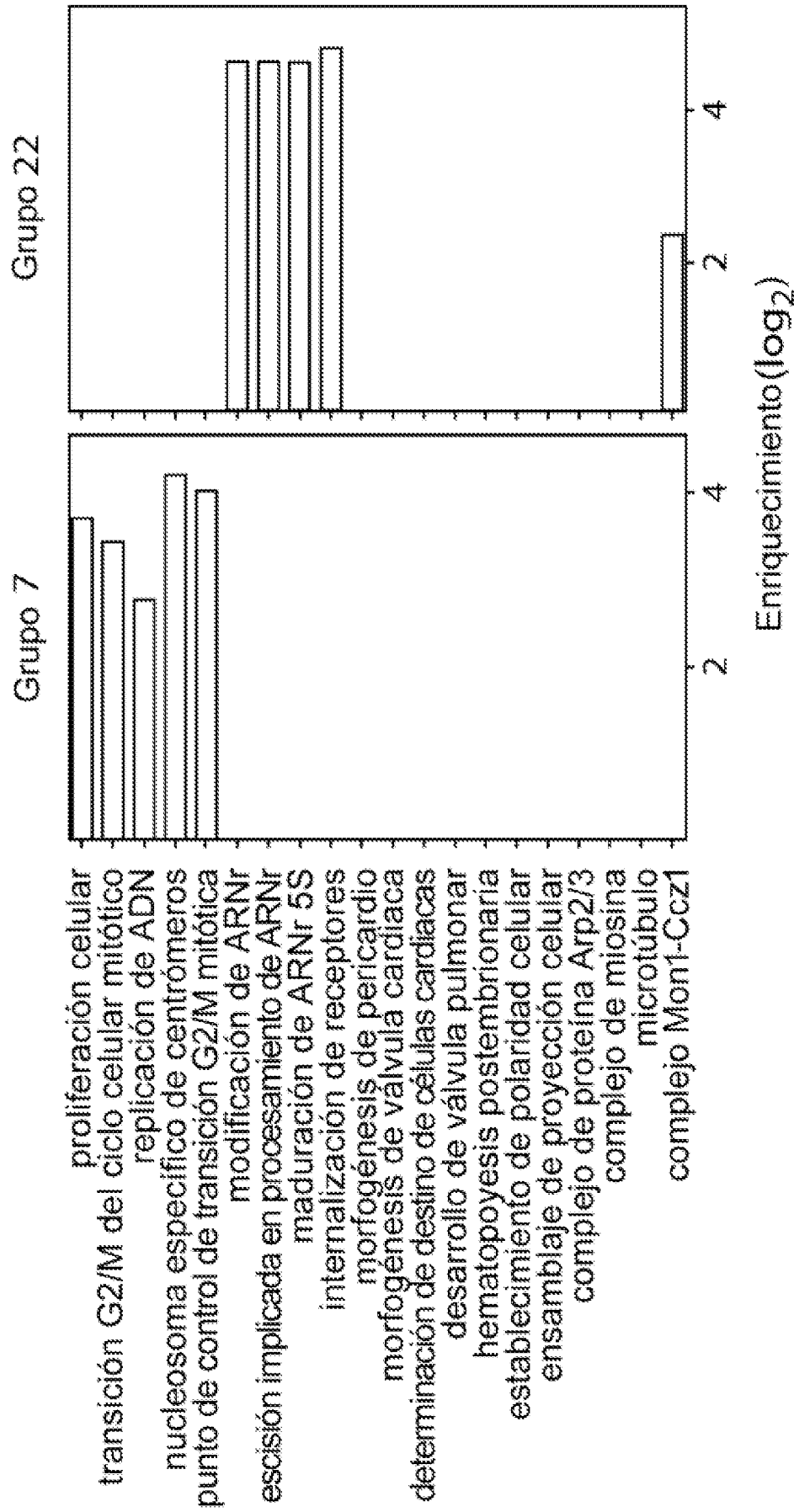


Fig. 10B

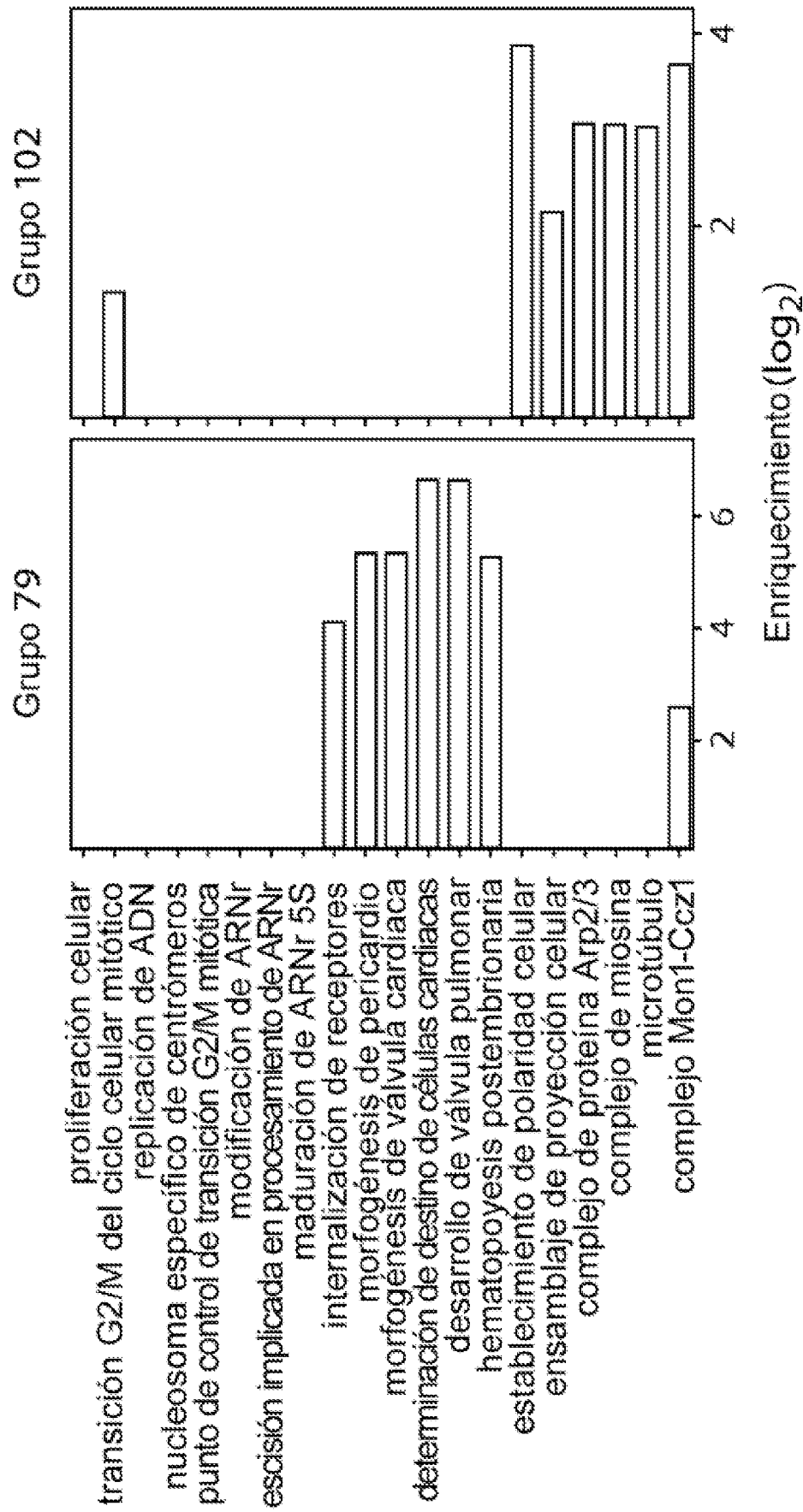


Fig. 10B Continuación

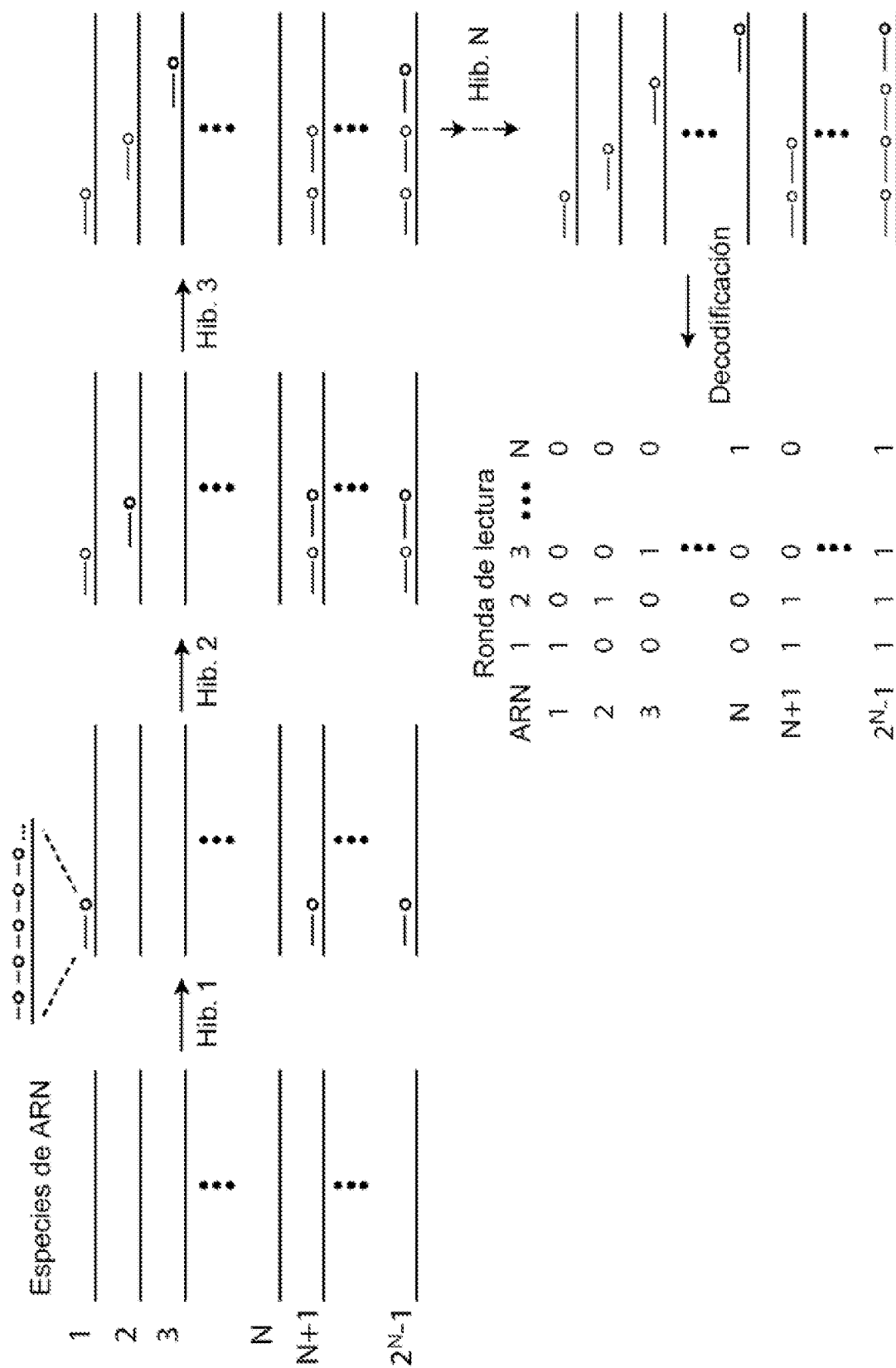


Fig. 11

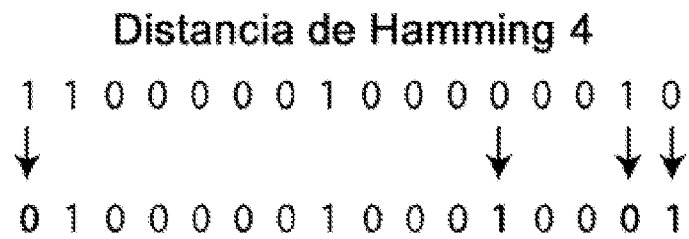


Fig. 12A

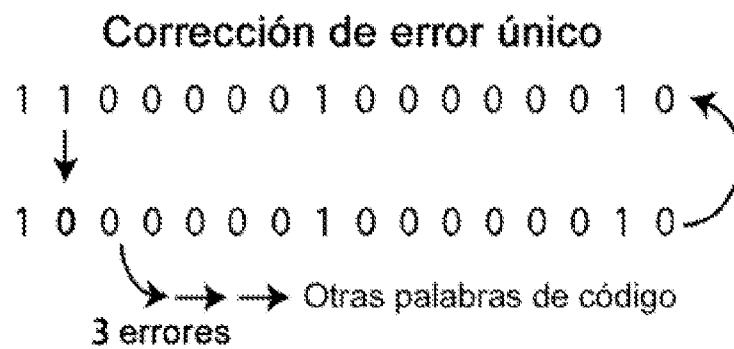


Fig. 12B

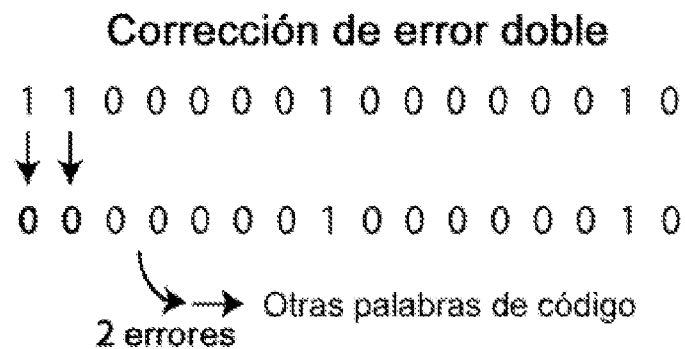


Fig. 12C

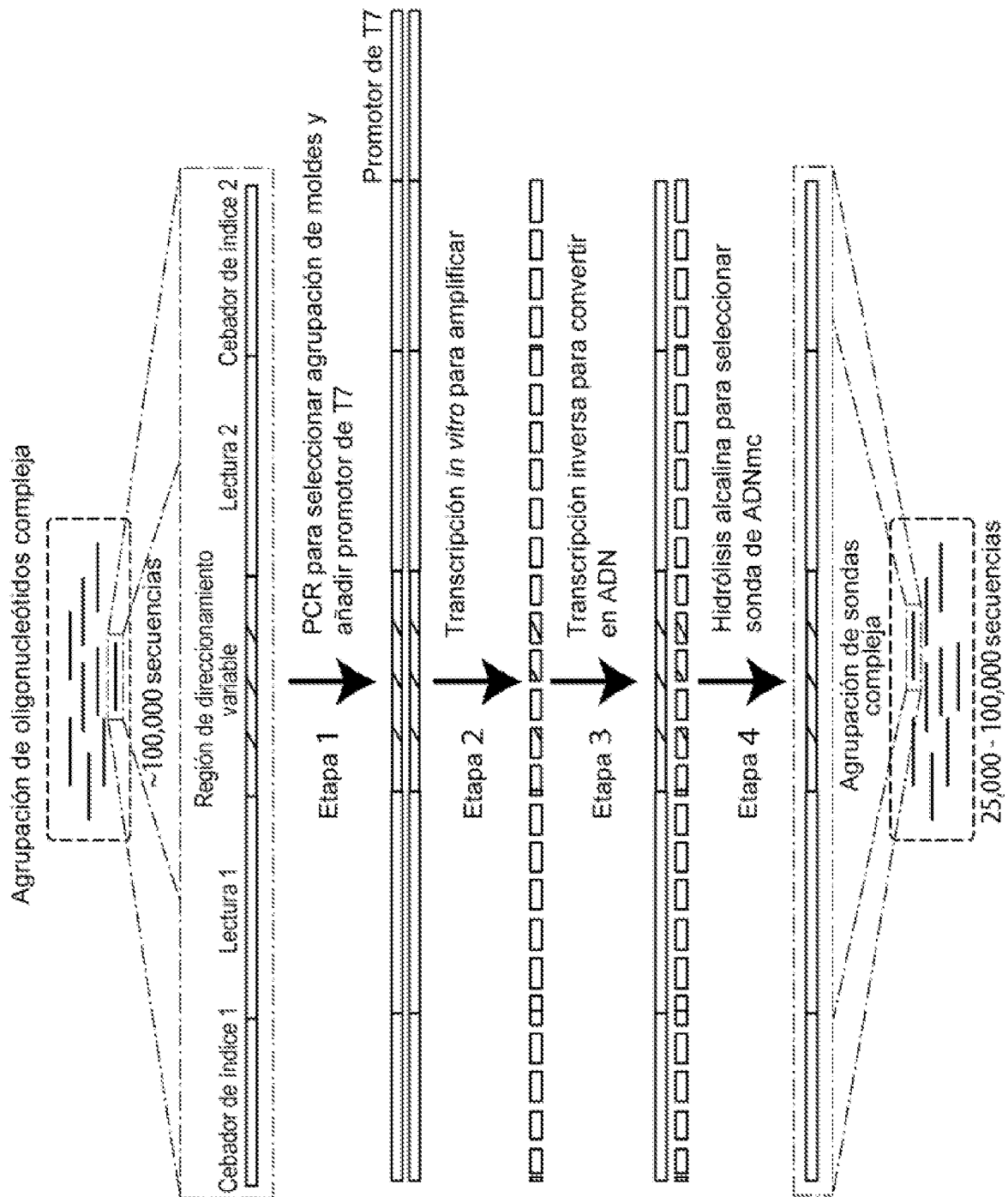


Fig. 13

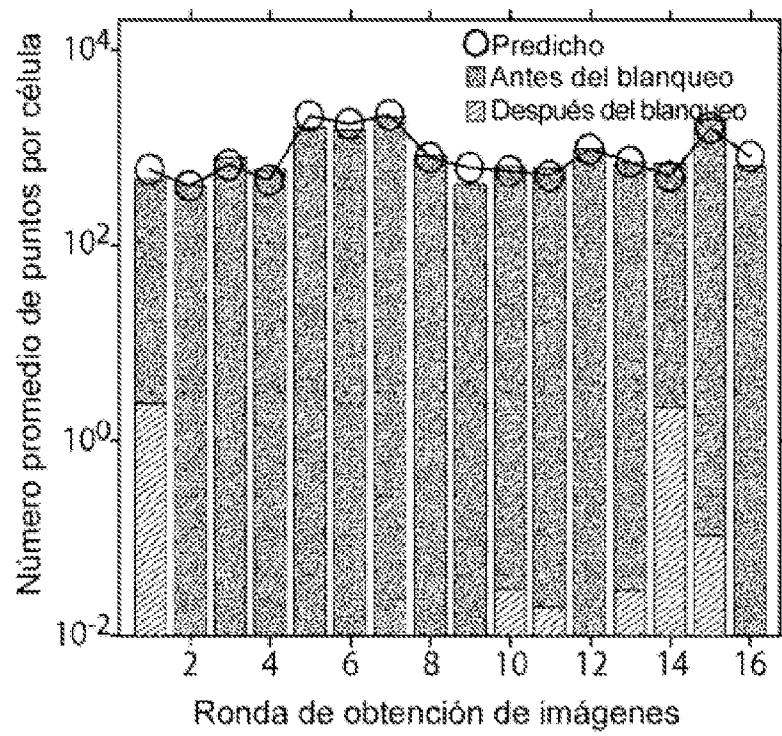


Fig. 14A

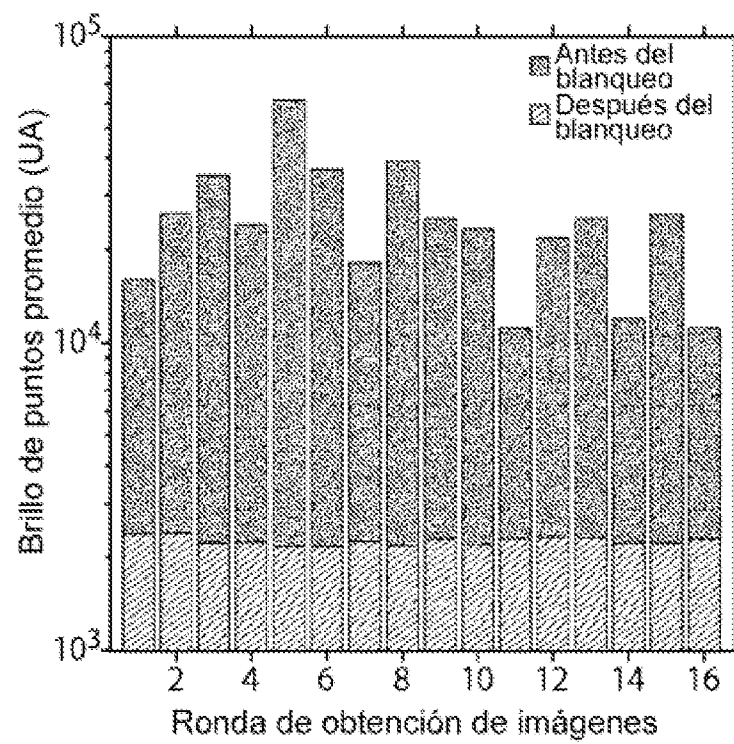


Fig. 14B

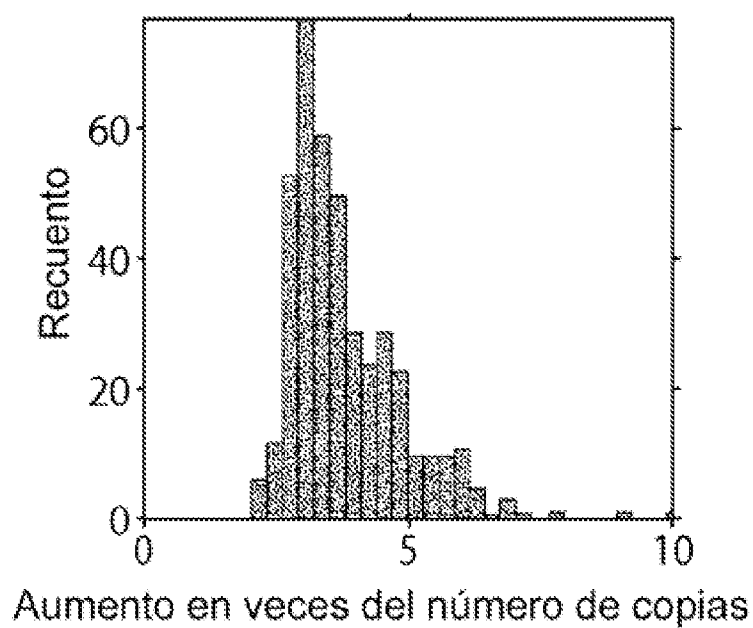


Fig. 15A

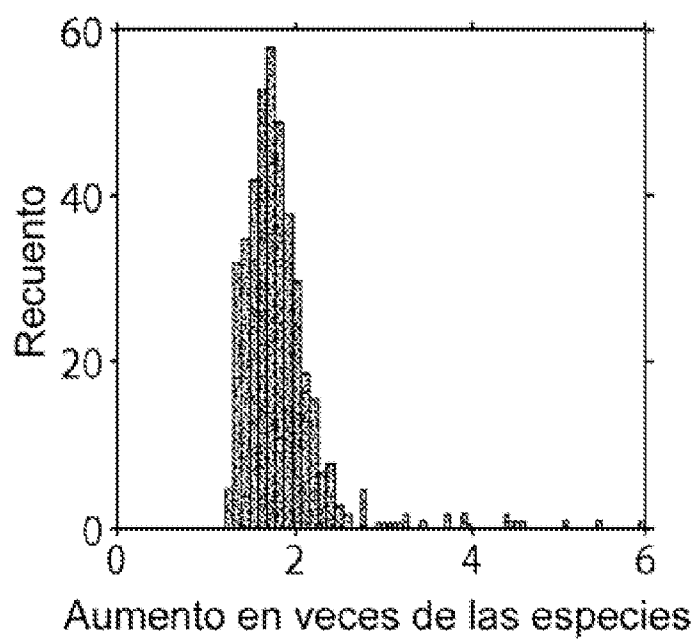


Fig. 15B

Libro de códigos 1

FLNA

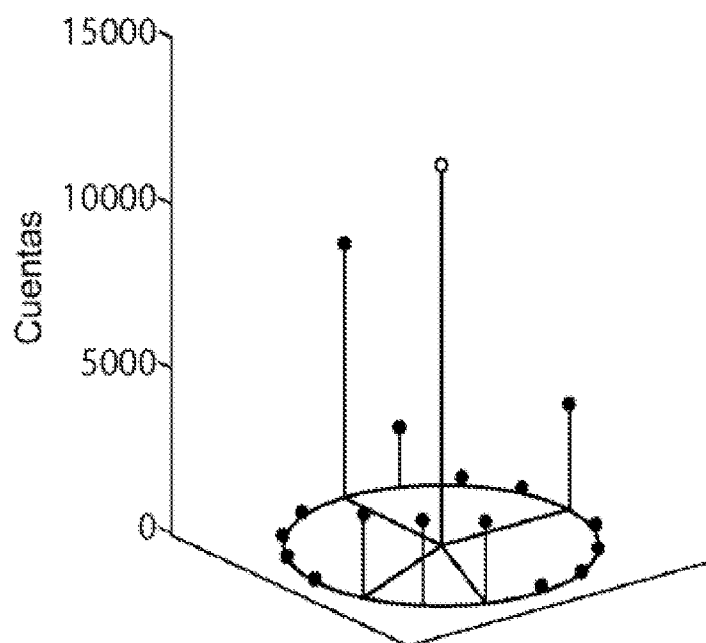


Fig. 16A

Control de identificaciones erróneas

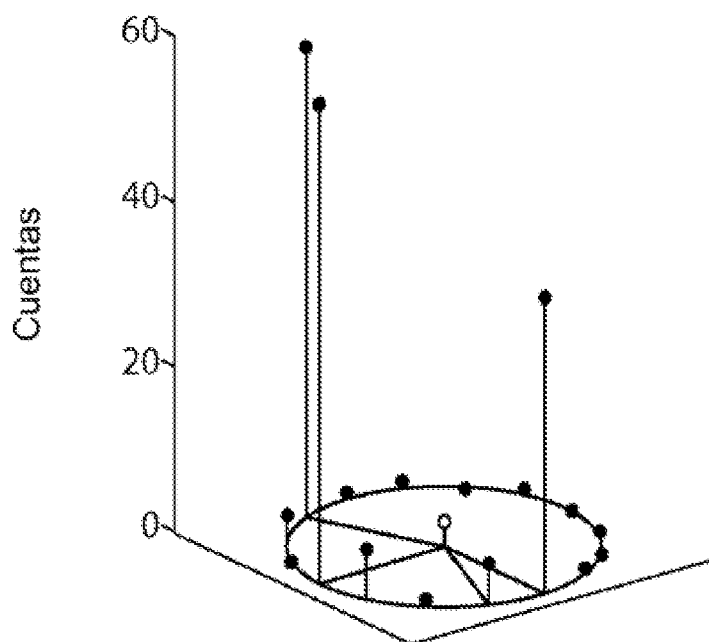


Fig. 16B

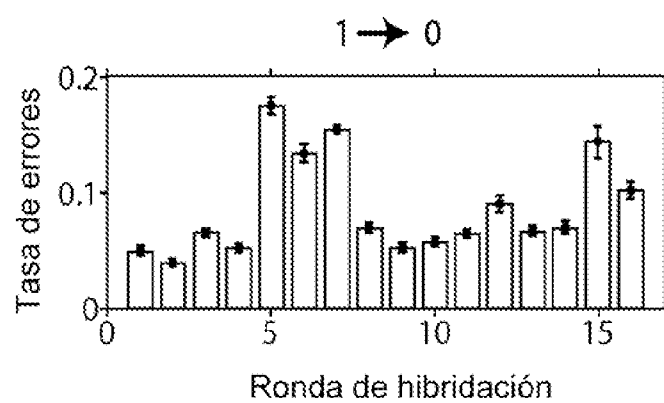


Fig. 16C

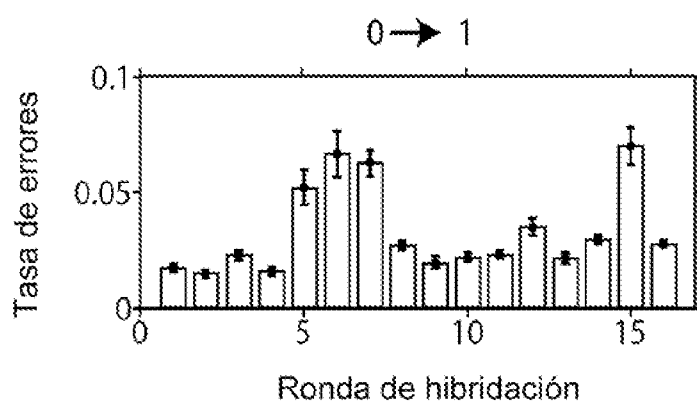


Fig. 16D

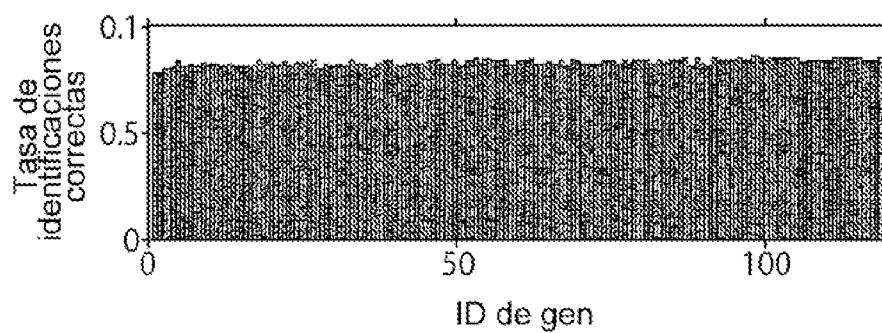


Fig. 16E

Libro de códigos 2

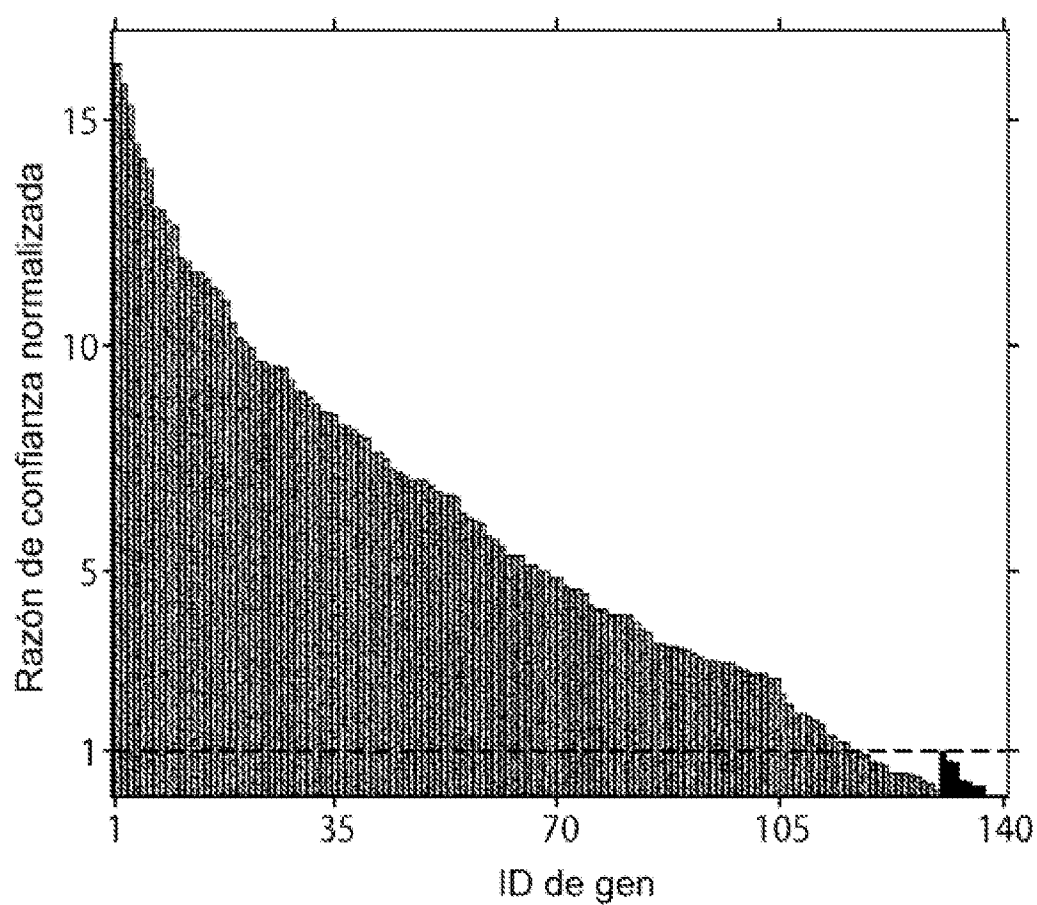


Fig. 17A

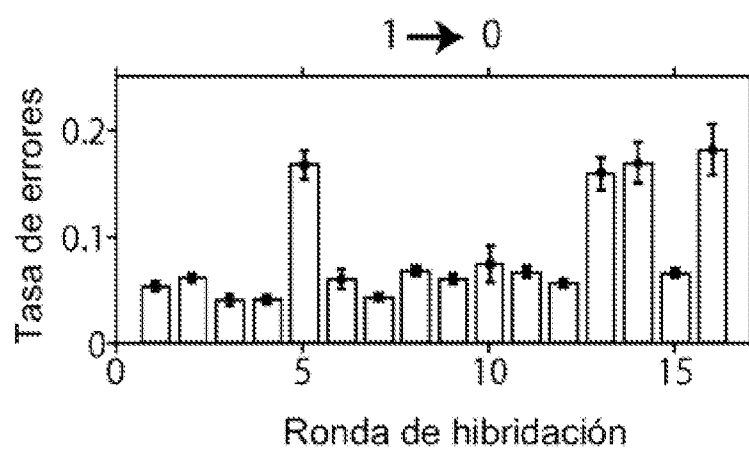


Fig. 17B

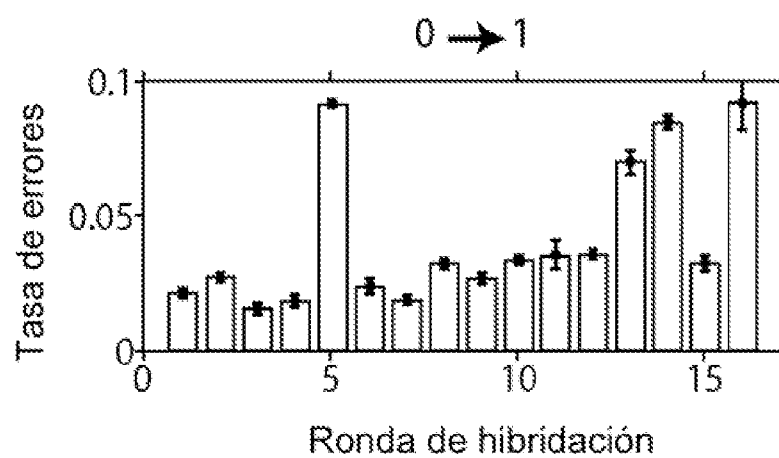


Fig. 17C

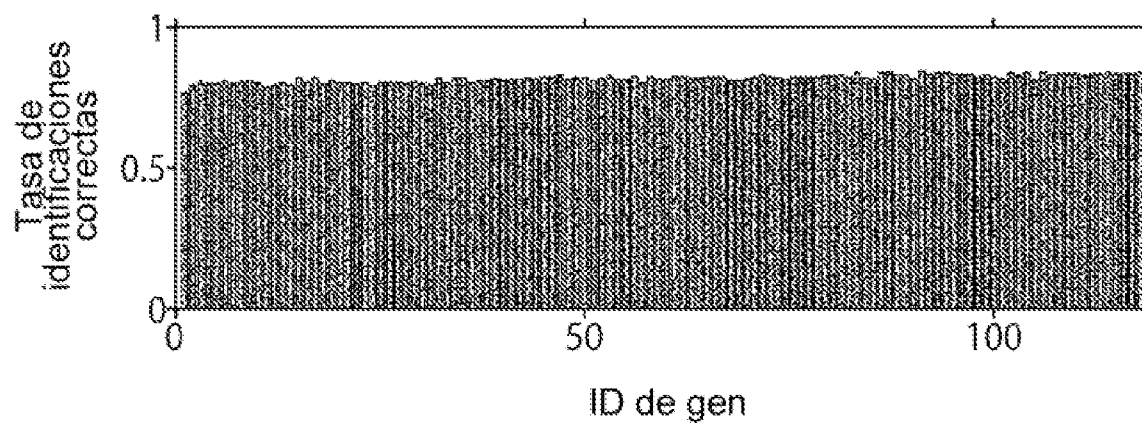


Fig. 17D

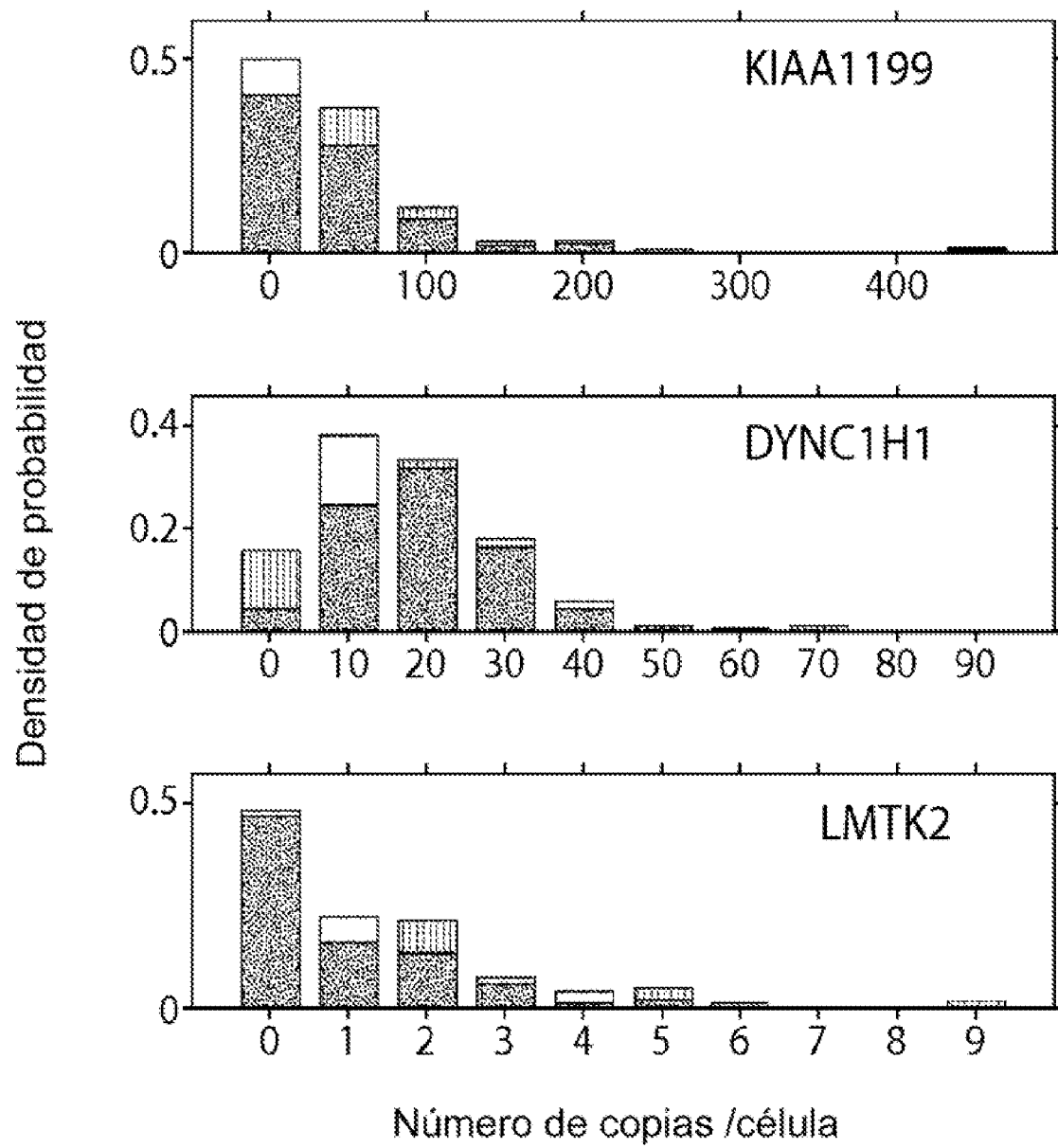


Fig. 18A

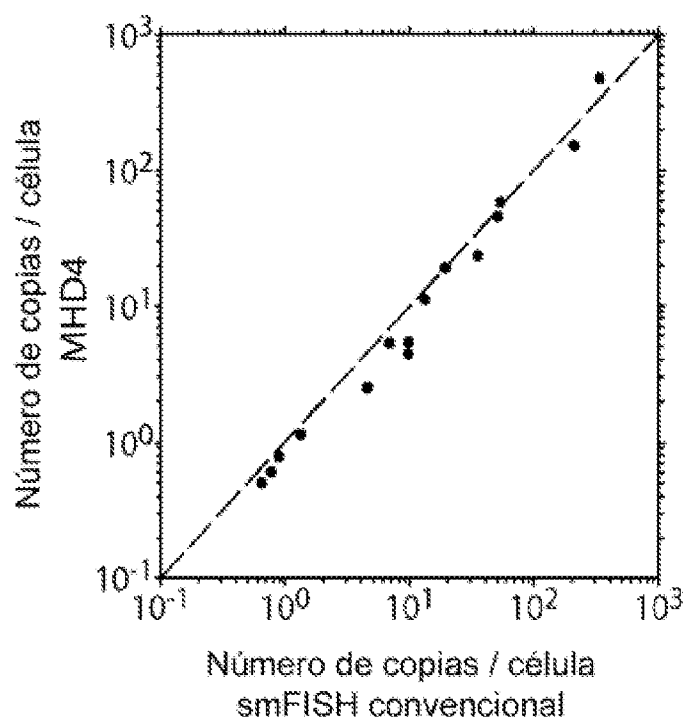


Fig. 18B

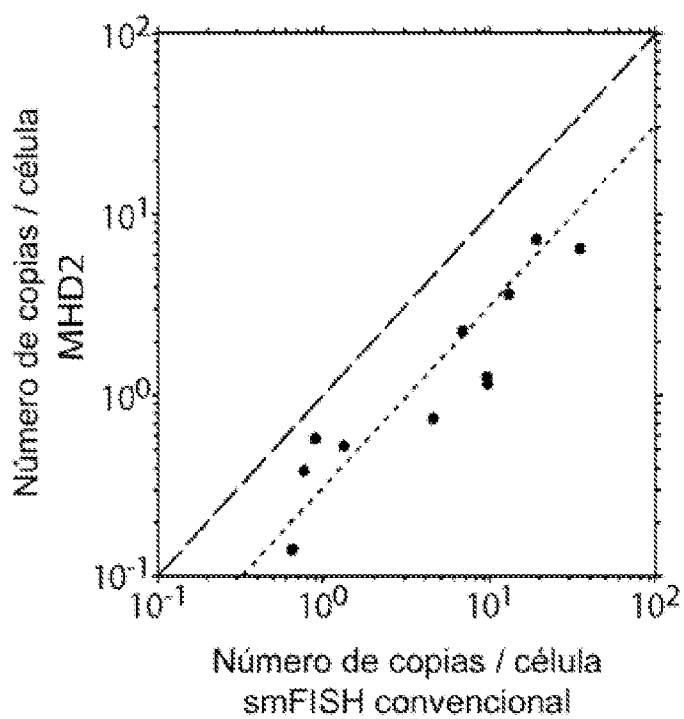


Fig. 18C

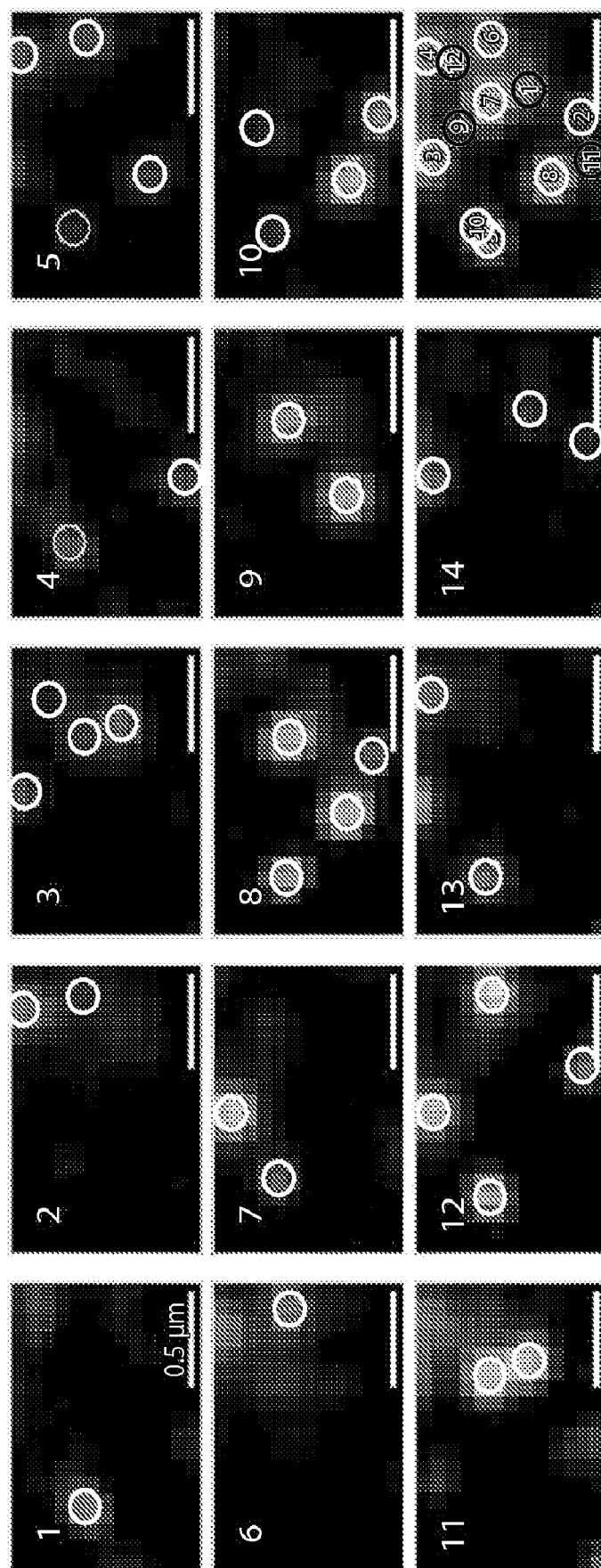


Fig. 19A

		Ronda de hibridación														Gen
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	
1	0	0	1	0	0	0	0	0	0	0	0	1	0	0	1	no identificado
2	0	0	0	0	0	0	0	0	1	0	1	0	1	0	1	FLNC
3	0	0	1	0	0	0	0	1	0	0	0	0	1	0	1	ZFPM2
4	0	1	0	0	1	1	1	0	0	0	0	0	0	1	0	LTBP2
5	1	0	0	0	0	0	0	0	1	0	0	0	1	1	0	TP53BP1
6	0	1	0	0	1	1	1	0	0	0	0	0	1	0	0	INTS1
7	0	0	1	0	0	0	0	0	1	1	0	1	0	0	0	PINK1-AS
8	0	0	0	0	1	0	0	0	1	1	1	0	0	0	0	MLLT4
9	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	no identificado
10	0	0	0	1	1	0	0	1	0	0	1	0	0	0	0	TRIM14
11	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	no identificado
12	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	no identificado

Fig. 19B

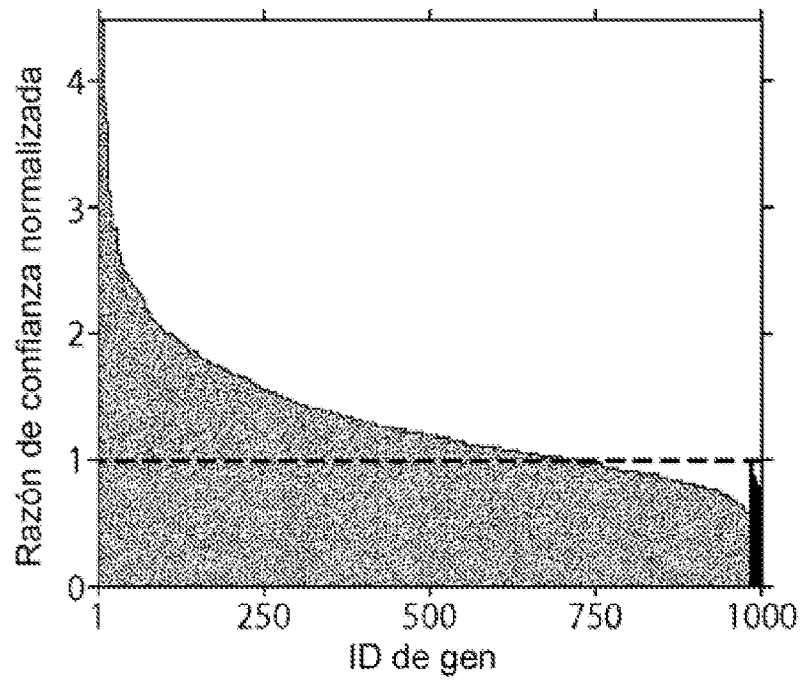


Fig. 19C

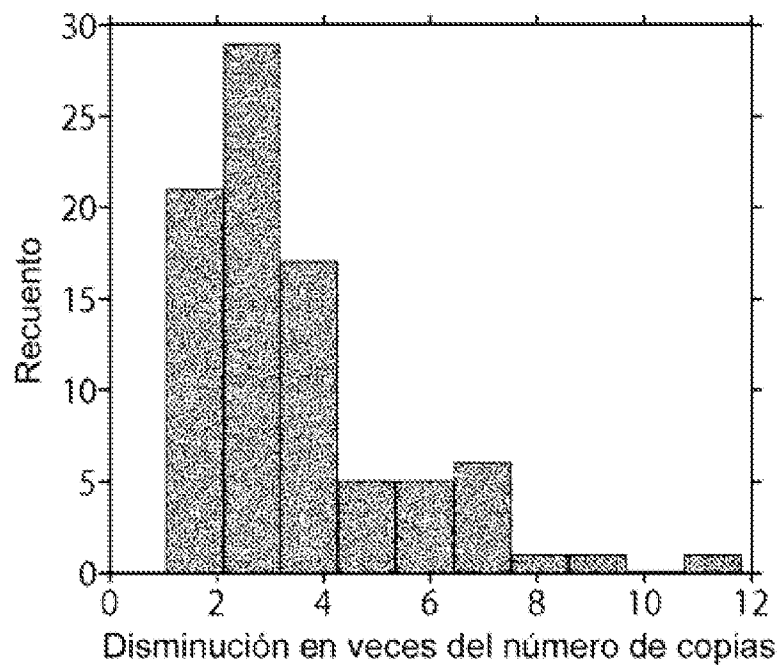


Fig. 19D

Libro de códigos 1 para los experimentos de 140 genes

<u>Genes</u>	<u>Palabras de código</u>
THBS1	0 0 0 0 1 1 1 0 0 0 0 0 0 0 1 0
FLNA	0 0 0 0 1 0 1 0 0 0 0 0 1 0 0 0 1
TNC	1 0 0 0 1 0 1 0 0 0 0 0 0 1 0 0 0
FBN2	0 0 1 0 1 0 1 1 0 0 0 0 0 0 0 0 0
KIAA1199	0 0 0 0 1 1 0 1 0 0 0 0 1 0 0 0 0
COL5A1	0 0 0 0 0 1 0 0 1 1 0 0 0 0 0 1 0
MALAT1	0 0 0 0 0 1 1 0 1 0 0 0 0 1 0 0 0
TLN1	0 0 0 0 1 0 0 0 0 0 0 1 0 0 1 1 0
FBN1	0 0 1 0 0 1 0 1 0 0 0 0 0 0 0 1 0
SRRM2	0 0 1 0 0 0 1 0 0 0 0 0 0 0 0 1 1
TEAD1	0 0 1 0 0 1 1 0 0 0 0 0 1 0 0 0 0
IQGAP1	0 0 0 0 0 0 1 1 0 0 0 0 1 0 0 1 0
MYH10	0 0 1 0 1 0 0 0 0 0 0 0 1 0 0 1 0
PAPPA	0 0 0 1 0 1 0 0 0 0 0 1 0 0 0 1 0
NR4A1	0 0 0 1 1 0 1 0 0 0 0 1 0 0 0 0 0
LRP1	0 0 0 0 0 1 1 0 0 0 0 1 0 0 1 0 0
SLC38A1	1 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 1 0
SMARCA5	0 0 0 0 0 1 1 1 0 0 0 0 0 0 0 0 1
CKAP5	0 1 0 0 0 0 1 0 0 0 0 1 0 0 0 1 0
PIEZO1	1 0 0 0 1 1 0 0 1 0 0 0 0 0 0 0 0
VCAN	0 0 0 0 1 0 1 0 1 1 0 0 0 0 0 0 0
PRRC2B	0 1 0 1 0 1 1 0 0 0 0 0 0 0 0 0 0
SON	1 0 0 0 0 1 0 0 0 0 0 0 0 1 0 1 0
SLC7A11	1 0 0 0 0 0 1 0 1 0 0 0 0 0 0 1 0
GPR107	0 1 0 0 0 1 0 0 0 0 0 0 0 0 0 1 1 0
KIAA1462	0 1 0 1 1 0 0 0 0 0 0 0 0 0 0 0 1 0
sin diana 001	0 0 0 0 1 0 0 0 0 1 0 0 0 0 1 0 1 0
blanco 001	0 1 0 0 1 1 0 0 0 0 0 1 0 0 0 0 0
ROCK2	0 1 0 0 1 0 1 0 0 0 0 0 0 0 0 1 0 0
SPTBN1	0 0 0 1 1 1 0 0 0 0 0 0 0 0 0 1 0 0
FLNC	0 0 0 1 0 0 1 0 0 0 0 0 0 0 0 1 1 0
PRPF8	0 0 0 0 0 1 0 0 0 0 0 0 1 0 0 1 1
IGF2R	0 0 0 0 1 1 0 0 0 0 1 0 0 1 0 0 0
FOSB	1 0 0 0 0 1 1 0 0 1 0 0 0 0 0 0 0
UBR5	0 0 0 0 1 0 0 1 0 0 0 0 0 0 0 1 1
SVEP1	0 0 0 0 0 0 1 0 0 1 0 0 1 0 0 1 0 1 0
CBX5	0 0 1 0 1 1 0 0 0 0 0 0 0 0 0 0 0 1
PDSSA	1 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 1
NUMA1	0 0 0 0 0 0 1 0 0 1 1 0 0 0 0 0 1
PLXNA1	0 0 0 1 0 0 0 0 0 0 0 0 0 1 1 0 1 0

Fig. 20-1

Libro de códigos 1 para los experimentos de 140 genes

<u>Genes</u>	<u>Palabras de código</u>
ALPK2	0 1 0 0 0 1 0 0 0 1 0 1 0 0 0 0
AFAP1	0 0 0 0 0 0 1 0 1 0 1 1 0 0 0 0
SPTAN1	0 0 1 1 1 0 0 0 0 0 0 0 1 0 0 0
DOCK7	0 1 1 0 0 0 0 0 0 0 0 0 1 0 1 0
NOTCH2	0 0 0 0 0 0 0 0 1 0 0 0 0 1 1 1
DYNC1H1	1 0 0 0 1 0 0 0 0 0 0 1 0 1 0 0
ITGA2	1 0 1 0 0 0 0 0 0 0 0 0 0 1 1 0
PRKDC	0 0 0 1 1 0 0 0 0 1 0 0 0 0 0 1
GTF3C1	0 0 1 0 0 1 0 0 0 0 0 0 1 1 0 0
THSD4	0 0 0 0 0 1 0 0 0 1 0 0 0 1 0 1
PTPN14	0 1 0 0 1 0 0 1 0 1 0 0 0 0 0 0
LUZP1	0 0 0 0 1 0 0 0 1 0 1 0 0 0 0 1
AMOTL1	1 0 0 0 0 0 1 1 0 0 1 0 0 0 0 0
MAN1A2	1 0 1 0 1 0 0 0 0 0 1 0 0 0 0 0
sin diana 002	1 0 0 1 1 0 0 1 0 0 0 0 0 0 0 0
blanco 002	1 1 0 0 0 0 0 1 0 0 0 0 0 0 1 0
USP9X	0 0 0 0 1 0 0 0 0 1 1 1 0 0 0 0
CREBBP	0 1 0 0 0 0 1 0 0 0 0 0 1 0 0 1
KPNA4	0 1 1 0 0 0 1 0 0 1 0 0 0 0 0 0
EGFR	0 0 0 1 0 0 1 0 0 1 0 1 0 0 0 0
ZC3H13	0 0 0 1 0 0 1 1 0 0 0 0 1 0 0 0
TNRC6A	0 0 0 1 0 1 0 1 1 0 0 0 0 0 0 0
TPR	0 0 0 0 0 1 0 0 1 0 0 1 0 1 0 0
COL7A1	0 0 0 1 0 1 0 0 0 0 0 0 1 0 0 1
ASCC3	0 1 0 0 0 0 1 1 1 0 0 0 0 0 0 0
USP34	0 0 0 0 0 0 0 0 0 0 1 0 1 0 1 1
CENPF	0 0 1 0 0 0 1 0 1 0 0 0 0 1 0 0
KIDINS220	0 0 0 0 0 1 0 0 0 0 1 1 1 0 0 0
SCUBE3	1 0 0 0 0 1 0 1 0 0 0 0 0 1 0 0
NRIP1	0 0 0 0 0 0 0 0 0 1 0 1 0 1 1 0
NUP98	1 0 0 1 0 0 0 0 0 0 0 0 0 0 1 1
ZFC3H1	0 0 0 0 0 0 0 1 1 0 1 0 0 0 1 0
FAM208B	0 1 0 0 0 0 0 0 0 1 0 0 0 0 1 1
RAB3B	0 0 0 1 0 0 0 1 0 1 0 0 0 0 1 0
AFF4	1 0 0 0 0 0 0 0 0 0 1 1 0 0 1 0
PRKCA	0 0 1 0 0 0 1 0 0 0 1 0 1 0 0 0
FLNB	1 0 0 0 0 0 1 0 0 0 0 0 0 1 0 1
CHD8	1 0 0 0 0 1 0 0 0 0 1 0 0 0 0 1
ANKRD52	0 1 0 0 1 0 0 0 0 0 0 1 1 0 0 0
MKI67	0 0 0 0 0 0 1 0 0 0 0 1 1 1 0 0

Fig. 20-2

Libro de códigos 1 para los experimentos de 140 genes

<u>Genes</u>	<u>Palabras de código</u>
DIP2C	0 1 1 0 1 0 0 0 1 0 0 0 0 0 0 0
SSH1	0 0 0 0 1 0 0 0 0 0 0 0 1 1 0 1
sin diana 003	0 0 1 0 0 1 0 0 1 0 1 0 0 0 0 0
blanco 003	1 1 1 0 0 1 0 0 0 0 0 0 0 0 0 0
DNAJC13	0 0 0 0 1 0 0 1 1 0 0 0 0 1 0 0
DSEL	0 1 0 0 0 0 0 0 1 0 0 1 0 0 1 0
PIK3CA	0 0 0 0 0 1 0 1 0 1 1 0 0 0 0 0
MED14	1 0 0 1 0 1 0 0 0 0 0 1 0 0 0 0
HERC2	0 1 0 0 0 1 0 0 1 0 0 0 0 0 0 1
H6PD	0 0 1 1 0 0 0 0 1 0 0 0 0 0 1 0
C14orf132	0 0 0 0 0 0 0 1 0 0 0 0 1 1 1 0
SIPA1L3	0 0 0 0 0 0 1 1 0 1 0 0 0 1 0 0
C17orf51	0 0 1 1 0 1 0 0 0 1 0 0 0 0 0 0
NF1	0 0 0 1 0 0 1 0 1 0 0 0 0 0 0 1
VPS13A	1 1 0 0 0 0 1 0 0 0 0 1 0 0 0 0
ABCA2	1 0 1 1 0 0 1 0 0 0 0 0 0 0 0 0
PCNX	0 1 0 0 0 1 0 1 0 0 0 0 1 0 0 0
NKTR	0 0 1 0 1 0 0 0 0 1 0 0 0 1 0 0
FYCO1	0 0 0 1 1 0 0 0 1 0 0 1 0 0 0 0
ATP11A	0 0 1 0 0 0 0 0 0 1 1 0 0 0 1 0
REV3L	0 0 0 0 1 0 0 1 0 0 1 0 1 0 0 0
CASC5	0 1 0 0 0 0 0 1 0 0 1 1 0 0 0 0
AKAP11	0 1 0 1 0 0 0 0 0 0 1 0 0 1 0 0
CEP250	0 0 1 0 0 0 0 0 0 1 0 0 1 0 0 1
USP24	0 0 0 0 0 0 0 0 1 0 1 0 1 1 0 0
AGPS	1 0 0 0 0 0 0 0 0 0 0 1 1 0 0 1
CBL	1 0 1 0 0 0 0 0 0 1 0 1 0 0 0 0
RNF169	0 0 0 1 0 0 0 0 0 0 1 1 0 0 0 1
TMOD2	0 0 0 1 0 0 0 0 0 1 0 0 1 1 0 0
FASN	0 0 0 0 0 0 0 0 1 1 0 1 0 0 0 1
sin diana 004	1 0 0 1 0 0 0 0 1 0 0 0 0 1 0 0
blanco 004	0 0 1 0 0 0 0 0 1 0 0 1 1 0 0 0
UBR2	0 1 1 0 0 0 0 0 0 1 0 0 0 0 0 1
PHIP	1 0 0 0 0 0 0 1 0 1 0 0 0 0 0 1
GTF3C4	0 0 0 0 0 0 0 1 0 1 0 1 1 0 0 0
QSER1	1 0 0 1 0 0 0 0 0 0 1 0 1 0 0 0
VPS13D	0 0 1 0 0 0 0 1 1 1 0 0 0 0 0 0
FRY	0 0 0 1 0 0 0 0 1 1 1 0 0 0 0 0
DIP2B	0 1 0 1 0 0 0 0 1 0 0 0 1 0 0 0
HIVEP2	1 1 0 1 0 0 0 0 0 1 0 0 0 0 0 0

Fig. 20-3

Libro de códigos 1 para los experimentos de 140 genes

<u>Genes</u>	<u>Palabras de código</u>
DOPEY1	1 0 0 0 0 0 0 0 1 1 0 0 1 0 0 0
ALMS1	0 1 1 0 0 0 0 1 0 0 0 0 0 1 0 0
HEATR5B	0 0 1 1 0 0 0 1 0 0 1 0 0 0 0 0
HIVEP1	0 0 1 0 0 0 0 1 0 0 0 1 0 0 0 1
ARL10	0 1 0 1 0 0 0 1 0 0 0 0 0 0 0 1
HELZ2	0 0 0 0 0 0 0 1 1 0 0 0 1 0 0 1
POLQ	1 0 1 0 0 0 0 0 1 0 0 0 0 0 0 1
CACNA1A	1 1 0 0 0 0 0 0 0 0 0 0 1 1 0 0
KIAA2018	0 1 0 0 0 0 0 0 0 0 0 1 0 1 0 1
LYST	1 0 0 0 0 0 0 1 1 0 0 1 0 0 0 0
ZKSCAN2	0 0 1 1 0 0 0 0 0 0 0 0 0 1 0 1
DIEXF	0 0 0 1 0 0 0 1 0 0 0 1 0 1 0 0
FZD4	1 0 1 0 0 0 0 1 0 0 0 0 1 0 0 0
SLC5A3	0 1 1 1 0 0 0 0 0 0 0 1 0 0 0 0
KIAA1147	1 1 0 0 0 0 0 0 1 0 1 0 0 0 0 0
LMTK2	0 1 0 0 0 0 0 0 0 1 1 0 1 0 0 0
STARD9	0 0 1 0 0 0 0 0 0 0 1 1 0 1 0 0
BRCA2	0 0 0 0 0 0 0 1 0 0 1 0 0 1 0 1
sin diana 005	1 0 0 0 0 0 0 0 0 1 1 0 0 1 0 0
blanco 005	1 1 0 0 0 0 0 0 1 0 1 0 0 0 0 0

Fig. 20-4

Libro de códigos 2 (aleatorizado) para los experimentos de 140 genes

<u>Genes</u>	<u>Palabras de código</u>
THBS1	0 0 0 0 1 0 0 0 0 0 0 0 1 1 0 1
FLNA	0 0 0 0 1 1 0 0 0 1 0 0 1 0 0 0
TNC	1 0 0 0 0 0 1 0 0 0 0 0 0 1 0 1
FBN2	0 1 0 0 0 0 0 0 0 0 0 0 1 0 1 0 1
KIAA1199	0 0 0 0 0 0 0 0 0 0 0 1 0 1 0 1 1
COL5A1	0 0 0 0 0 0 0 1 0 0 0 0 0 1 1 1 0
MALAT1	0 0 0 0 0 0 0 0 0 1 0 0 0 0 1 1 1
TLN1	0 0 0 1 0 0 0 0 0 0 1 0 0 1 1 0 0
FBN1	1 0 0 0 1 0 0 0 0 0 0 0 1 0 1 0 0
SRRM2	0 0 0 0 1 0 0 1 0 0 0 0 0 0 0 1 1
TEAD1	1 1 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0
IQGAP1	0 0 1 0 1 1 0 0 0 0 0 0 0 0 0 0 1
MYH10	0 0 0 1 1 0 0 0 0 0 1 0 0 0 0 0 1
PAPPA	0 0 0 0 1 0 0 1 0 0 1 0 1 0 1 0 0 0
NR4A1	0 0 0 0 1 0 0 0 1 0 1 0 0 0 0 0 1
LRP1	0 0 0 0 0 0 0 0 0 1 0 1 0 1 1 1 0 0
SLC38A1	0 0 0 0 0 1 0 0 0 1 0 0 0 1 0 1
SMARCA5	0 0 1 0 0 0 0 0 0 1 0 0 1 0 0 1
CKAP5	0 0 1 0 0 1 0 0 0 0 0 0 0 1 1 0 0
PIEZO1	0 0 0 0 0 0 0 1 0 0 1 0 0 1 0 1
VCAN	0 0 0 0 0 0 1 0 0 0 0 0 1 1 1 0 0
PRRC2B	1 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0 1
SON	0 0 1 0 1 0 0 0 0 1 0 0 0 1 0 0
SLC7A11	0 1 0 0 1 0 1 0 0 0 0 0 0 1 0 0
GPR107	0 0 0 0 1 0 1 0 0 0 0 0 1 0 0 0 1
KIAA1462	0 0 1 1 1 0 0 0 0 0 0 0 0 1 0 0 0
sin diana 001	1 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1
blanco 001	1 0 0 0 1 0 1 0 0 0 0 0 0 1 0 0 0
ROCK2	0 1 0 0 0 0 1 0 0 0 0 0 0 1 0 0 1
SPTBN1	0 0 0 0 1 0 0 0 0 0 0 1 0 0 1 1 0
FLNC	0 0 0 0 1 0 0 0 1 0 0 0 1 0 1 0
PRPF8	0 0 0 0 0 0 0 1 1 0 0 0 1 0 0 1
IGF2R	0 0 0 1 1 1 0 0 0 0 0 0 0 0 1 0 0
FOSB	0 0 1 1 0 0 0 0 0 0 0 0 0 0 1 0 1
UBR5	0 1 0 0 1 0 0 0 0 0 0 0 1 1 0 0 0
SVEP1	0 0 0 0 1 0 0 1 1 0 0 0 0 1 0 0
CBX5	0 0 0 1 0 1 0 0 0 0 0 0 0 1 0 0 1
PDSSA	1 0 0 1 1 0 0 1 0 0 0 0 0 0 0 0
NUMA1	0 1 0 0 0 1 0 1 0 0 0 0 0 1 0 0 0
PLXNA1	0 1 0 1 0 0 0 0 0 0 1 0 0 1 0 0

Fig. 20-5

Libro de códigos 2 (aleatorizado) para los experimentos de 140 genes

<u>Genes</u>	<u>Palabras de código</u>
ALPK2	0 1 1 0 0 0 0 0 0 0 0 0 1 0 1 0
AFAP1	1 0 0 0 0 0 0 0 0 0 1 1 0 0 1 0 0
SPTAN1	0 0 0 0 0 0 0 0 0 0 1 0 1 0 1 1 0
DOCK7	0 0 1 0 0 0 0 0 0 0 0 1 1 0 1 0 0
NOTCH2	0 0 0 0 0 1 1 0 1 0 0 0 1 0 0 0
DYNC1H1	0 1 1 0 1 0 0 0 1 0 0 0 0 0 0 0
ITGA2	0 0 0 1 1 0 0 0 1 0 0 1 0 0 0 0
PRKDC	0 0 0 0 0 0 0 0 1 1 0 1 0 0 0 1
GTF3C1	0 0 1 0 1 0 0 0 0 0 0 1 0 0 1 0
THSD4	1 0 0 0 0 0 0 1 0 1 0 0 0 0 0 1
PTPN14	0 0 0 0 0 0 1 1 0 1 0 0 0 1 0 0
LUZP1	1 0 1 0 0 0 0 0 0 0 0 0 0 1 1 0
AMOTL1	0 0 0 0 1 0 1 0 1 1 0 0 0 0 0 0
MAN1A2	0 1 0 1 0 0 0 1 0 0 0 0 0 0 0 1
sin diana 002	1 0 0 0 0 1 0 0 0 0 0 0 1 0 1 0
blanco 002	1 0 0 1 0 0 0 0 1 0 0 0 0 1 0 0
USP9X	0 0 0 0 0 0 0 1 0 1 0 1 1 0 0 0
CREBBP	0 1 0 0 0 1 0 0 1 0 0 0 0 0 0 1
KPNA4	1 0 0 1 0 0 0 0 0 0 0 0 0 0 1 1
EGFR	1 0 0 0 0 1 0 0 0 0 1 0 0 0 0 1
ZC3H13	1 0 0 0 1 1 0 0 1 0 0 0 0 0 0 0
TNRC6A	0 0 0 0 1 0 0 0 0 1 1 1 0 0 0 0
TPR	1 0 1 0 0 0 0 1 0 0 0 0 1 0 0 0
COL7A1	0 0 1 0 0 0 0 1 0 0 0 1 0 0 0 1
ASCC3	0 1 0 0 1 0 0 1 0 1 0 0 0 0 0 0
USP34	0 0 1 0 0 0 1 0 0 0 1 0 1 0 0 0
CENPF	0 0 1 0 0 0 0 0 1 0 0 1 1 0 0 0
KIDINS220	0 0 0 1 0 0 0 0 0 0 1 1 0 0 0 1
SCUBE3	0 0 0 0 0 0 1 0 0 1 1 0 0 0 0 1
NRIP1	0 0 0 1 1 0 1 0 0 0 1 0 0 0 0 0
NUP98	0 0 0 0 0 1 0 0 0 0 1 1 1 0 0 0
ZFC3H1	0 0 0 1 0 0 1 0 1 0 0 0 0 0 0 1
FAM208B	0 1 0 1 1 0 0 0 0 0 0 0 0 0 1 0
RAB3B	0 0 0 1 0 0 0 0 0 0 0 1 1 0 1 0
AFF4	0 1 0 0 0 0 0 0 0 1 0 0 0 0 1 1
PRKCA	0 1 0 0 0 1 0 0 0 0 0 0 0 1 1 0
FLNB	0 1 0 0 0 0 0 0 1 1 0 0 0 1 0 0
CHD8	0 0 0 0 0 1 0 0 0 0 0 1 0 0 1 1
ANKRD52	0 0 1 0 0 0 1 0 0 0 0 0 0 0 1 1
MKI67	0 1 0 0 1 1 0 0 0 0 1 0 0 0 0 0

Fig. 20-6

Libro de códigos 2 (aleatorizado) para los experimentos de 140 genes

<u>Genes</u>	<u>Palabras de código</u>
DIP2C	1 0 1 0 1 0 0 0 0 0 1 0 0 0 0 0
SSH1	0 1 0 1 0 0 0 0 0 1 0 0 0 1 0 0
sin diana 003	0 1 1 0 0 0 0 0 0 0 1 0 0 0 0 1
blanco 003	0 0 0 0 1 1 1 0 0 0 0 0 0 0 1 0
DNAJC13	0 0 0 1 0 0 1 1 0 0 0 0 1 0 0 0
DSEL	0 0 0 1 0 0 0 1 0 0 0 1 0 1 0 0
PIK3CA	0 0 0 0 0 1 1 0 0 0 1 0 0 1 0 0
MED14	0 0 0 0 0 1 1 1 0 0 0 0 0 0 0 1
HERC2	1 0 0 0 1 0 0 0 0 1 0 0 0 0 1 0
H6PD	0 0 1 0 0 0 1 0 1 0 0 0 0 1 0 0
C14orf132	0 0 0 0 0 1 0 0 1 0 0 1 0 1 0 0
SIPA1L3	1 0 1 0 0 0 0 0 0 1 0 0 0 0 0 1
C17orf51	0 1 0 0 0 0 0 0 0 1 1 0 1 0 0 0
NF1	1 0 0 0 0 0 0 0 0 1 1 0 0 1 0 0
VPS13A	0 0 1 0 1 0 1 1 0 0 0 0 0 0 0 0
ABCA2	0 0 0 0 0 0 1 0 0 1 0 0 1 0 1 0
PCNX	0 0 0 1 0 0 1 0 0 0 0 0 0 0 1 0
NKTR	0 1 1 0 0 0 0 1 0 0 0 0 0 1 0 0
FYCO1	0 0 0 0 1 1 0 1 0 0 0 1 0 0 0 0
ATP11A	1 0 0 1 0 0 0 0 0 0 0 1 0 1 0 0
REV3L	1 0 0 0 0 1 0 1 0 0 0 0 0 1 0 0
CASC5	0 0 0 0 0 1 0 1 0 1 1 0 0 0 0 0
AKAP11	0 0 1 0 0 1 0 0 1 0 1 0 0 0 0 0
CEP250	1 0 1 0 0 0 0 0 0 1 0 1 0 0 0 0
USP24	1 1 0 1 0 0 0 0 0 1 0 0 0 0 0 0
AGPS	1 0 0 0 0 0 0 0 0 0 0 1 1 0 0 1
CBL	0 1 0 0 0 0 1 1 1 0 0 0 0 0 0 0
RNF169	0 0 1 1 0 0 0 1 0 0 1 0 0 0 0 0
TMOD2	0 0 1 0 0 1 0 1 0 0 0 0 0 0 1 0
FASN	0 0 0 0 0 0 1 1 0 0 0 1 0 0 1 0
sin diana 004	0 1 1 0 0 0 1 0 0 1 0 0 0 0 0 0
blanco 004	1 1 0 0 0 0 0 1 0 0 0 0 0 0 1 0
UBR2	0 1 0 0 0 0 0 0 1 0 0 1 0 0 1 0
PHIP	0 0 1 0 0 1 1 0 0 0 0 1 0 0 0 0
GTF3C4	0 1 0 0 0 0 1 0 0 0 1 0 0 0 1 0
QSER1	1 0 0 0 0 0 1 1 0 0 1 0 0 0 0 0
VPS13D	1 1 0 0 0 0 1 0 0 0 0 1 0 0 0 0
FRY	1 0 0 0 0 0 1 0 1 0 0 0 0 0 1 0
DIP2B	0 1 0 0 0 1 0 0 0 1 0 1 0 0 0 0
HIVEP2	1 0 1 1 0 0 1 0 0 0 0 0 0 0 0 0

Fig. 20-7

Libro de códigos 2 (aleatorizado) para los experimentos de 140 genes

<u>Genes</u>	<u>Palabras de código</u>
DOPEY1	0 0 1 1 0 0 0 0 1 0 0 0 0 0 1 0
ALMS1	0 0 1 1 0 1 0 0 0 1 0 0 0 0 0 0
HEATR5B	0 0 1 0 0 0 0 1 1 1 0 0 0 0 0 0
HIVEP1	0 0 0 1 0 1 0 0 0 0 1 0 0 0 1 0
ARL10	0 0 0 0 0 1 0 0 1 1 0 0 0 0 1 0
HELZ2	0 0 0 1 0 0 0 1 0 1 0 0 0 0 1 0
POLQ	0 0 0 0 0 0 1 0 1 0 1 1 0 0 0 0
CACNA1A	1 0 0 0 0 1 1 0 0 1 0 0 0 0 0 0
KIAA2018	0 0 0 0 0 0 0 1 1 0 1 0 0 0 1 0
LYST	0 0 0 1 0 1 0 1 1 0 0 0 0 0 0 0
ZKSCAN2	0 1 0 1 0 1 1 0 0 0 0 0 0 0 0 0
DIEXF	0 1 0 0 0 0 0 1 0 0 1 1 0 0 0 0
FZD4	0 0 0 1 0 0 1 0 0 1 0 1 0 0 0 0
SLC5A3	1 0 0 0 0 0 0 1 1 0 0 1 0 0 0 0
KIAA1147	0 0 1 0 0 0 0 0 0 1 1 0 0 0 1 0
LMTK2	0 1 1 1 0 0 0 0 0 0 0 0 1 0 0 0
STARD9	0 0 0 1 0 0 0 0 1 1 1 0 0 0 0 0
BRCA2	1 0 0 1 0 1 0 0 0 0 0 1 0 0 0 0
sin diana 005	1 1 1 0 0 1 0 0 0 0 0 0 0 0 0 0
blanco 005	1 1 0 0 0 0 0 0 1 0 1 0 0 0 0 0

Fig. 20-8