

19



OFICINA ESPAÑOLA DE  
PATENTES Y MARCAS

ESPAÑA



11 Número de publicación: **2 615 733**

51 Int. Cl.:

**C12Q 1/68** (2006.01)

12

TRADUCCIÓN DE PATENTE EUROPEA

T3

86 Fecha de presentación y número de la solicitud internacional: **16.12.2011 PCT/US2011/065600**

87 Fecha y número de publicación internacional: **21.06.2012 WO2012083225**

96 Fecha de presentación y número de la solicitud europea: **16.12.2011 E 11848932 (7)**

97 Fecha y número de publicación de la concesión europea: **16.11.2016 EP 2652155**

54 Título: **Métodos para el análisis masivo paralelo de ácidos nucleicos en células individuales**

30 Prioridad:

**16.12.2010 US 459600 P**

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

**08.06.2017**

73 Titular/es:

**GIGAGEN, INC. (100.0%)  
407 Cabot Road  
South San Francisco, CA 94080, US**

72 Inventor/es:

**JOHNSON, DAVID, SCOTT y  
MEYER, EVERETT, HURTEAU**

74 Agente/Representante:

**ARIAS SANZ, Juan**

**ES 2 615 733 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

Métodos para el análisis masivo paralelo de ácidos nucleicos en células individuales

### ANTECEDENTES DE LA INVENCIÓN

#### Campo de la invención

- 5 La invención se refiere a los campos de la biología molecular y de diagnóstico molecular, y más específicamente a métodos para el análisis genético masivo paralelo de ácidos nucleicos en células individuales.

#### Descripción de la Técnica Relacionada

10 Determinados análisis genéticos cuantitativos de tejidos biológicos y organismos se realizan mejor a nivel de células individuales. Sin embargo, las células individuales sólo contienen picogramos de material genético. Métodos convencionales tales como la reacción en cadena de la polimerasa (PCR), la secuenciación de ARN (Mortazavi *et al.*, 2008 Nature Methods. 5:621-8), la secuenciación por inmunoprecipitación de cromatina (Johnson *et al.*, 2007 Science 316:1497-502) o la secuenciación del genoma completo (Lander *et al.*, 2001 Nature 409:860-921), requieren más material genético que se encuentra en una célula individual y se realiza generalmente con miles de millones de células. Estas técnicas proporcionan información genética útil al nivel de la población de células, pero tienen serias limitaciones para la comprensión de la biología a nivel de células individuales. Herramientas biológicas actuales también carecen de la capacidad para analizar las mediciones genéticas en muchas células individuales en paralelo.

15 Las técnicas de células individuales convencionales son lentas, tediosas y están limitadas a la cantidad de células que se pueden analizar a la vez. Por ejemplo, en el diagnóstico genético pre-implantación (PGD), una célula individual se retira de un embrión humano en estadio de división para el análisis de todo el genoma de enfermedades genéticas (Johnson *et al.*, 2010 Human Reproduction. 25:1066-75). Aplicaciones tales como el PGD requieren una tecnología de la biopsia guiada manualmente, que consume tiempo, y los estudios más amplios incluyen cientos de células individuales. En otro ejemplo, la recombinación genética entre loci de interés se puede medir en células de esperma individuales (Jiang *et al.*, 2005 Nucleic Acids Research 33: e91), pero un análisis manual de miles de espermatozoides individuales consumiría tiempo y sería poco práctico.

20 Un método establecido para el análisis de células individuales es la clasificación de células activadas por fluorescencia (FACS). Células individuales se diluyen en pocillos de reacción, y en los pocillos pueden llevarse a cabo diversas técnicas genéticas y de biología molecular, desde la amplificación de todo el genoma a ensayos por PCR de locus individuales. Sin embargo, debido a los límites físicos de paralelización utilizando pocillos de reacción, la FACS sólo es útil para el análisis de cientos de células individuales, en lugar de cientos de miles de células individuales.

25 Las células individuales también se pueden utilizar como compartimientos de reacción para llevar a cabo diversos análisis genéticos (Embleton *et al.*, 1992 Nucleic Acids Research 20:3831-37; Hviid, 2002 Clinical Chemistry 48:2115-2123; Patente de EE.UU. 5.830.663). Las células individuales pueden ser ordenadas en emulsiones acuosas en aceite de microgotitas, y los análisis moleculares se pueden realizar en las microgotitas (Johnston *et al.*, 1996 Science 271: 624-626; Brouzes *et al.*, 2009 PNAS 106: 14195-200; Kliss *et al.*, 2008 Anal Chem 80:8975-81; Zeng *et al.*, 2010 Anal Chem 82: 3183-90). Estos ensayos de células individuales se limitan a la PCR de células individuales en emulsiones, o a la PCR in situ en células individuales fijadas y permeabilizadas. Además de ello, cuando se analizan grandes poblaciones de células, es difícil rastrear cada uno de los productos genéticos a una célula individual o sub-poblaciones de células.

30 Por lo tanto, existe la necesidad de métodos para la caracterización genética de alto rendimiento, masivamente paralela, de células individuales y métodos para identificar la célula o sub-población de células que originó el material genético.

### SUMARIO DE LA INVENCIÓN

35 Se describe en esta memoria un método para analizar al menos dos secuencias de ácido nucleico en una célula individual contenida en una población de al menos 10.000 células. El método incluye proporcionar un primer conjunto de sondas de ácido nucleico, comprendiendo el primer conjunto una primera sonda que comprende una secuencia que es complementaria a una primera sub-secuencia de ácido nucleico diana, una segunda sonda que comprende una secuencia que es complementaria a una segunda sub-secuencia del primer ácido nucleico diana y una segunda secuencia que es complementaria a una secuencia exógena, comprendiendo una tercera sonda la secuencia exógena y una secuencia que es complementaria a una primera sub-secuencia de un segundo ácido nucleico diana, y comprendiendo una cuarta sonda una secuencia que es complementaria a una segunda sub-secuencia de la segunda secuencia de ácido nucleico diana.

El método incluye aislar las células individuales con al menos un conjunto de sondas de ácido nucleico; amplificar las primera y segunda secuencias de ácido nucleico diana de forma independiente, en el que la primera secuencia de ácido nucleico diana se amplifica utilizando la primera sonda y la segunda sonda, y en el que la segunda secuencia de ácido nucleico diana se amplifica utilizando la tercera sonda y la cuarta sonda; hibridar la secuencia exógena a su complemento; y amplificar la primera secuencia de ácido nucleico diana, la segunda secuencia de ácido nucleico diana y la secuencia exógena utilizando la primera y cuarta sondas, generando de este modo un complejo fusionado.

El método también proporciona realizar una reacción de secuenciación a granel para generar información de la secuencia de al menos 100.000 complejos fusionados de al menos 10.000 células dentro de la población de células, en el que la información de la secuencia es suficiente para co-localizar la primera secuencia de ácido nucleico diana y la segunda secuencia de ácido nucleico diana de una célula individual de la población de al menos 10.000 células.

En un aspecto, la célula individual es aislada en una microgotita de la emulsión. En otro aspecto, la célula individual es aislada en un recipiente de reacción.

En una realización, la etapa de amplificación incluye realizar una reacción en cadena de la polimerasa, en el que la etapa de amplificación comprende realizar una reacción en cadena de la polimerasa, y en el que la primera y tercera sondas son cebadores directos y la segunda y cuarta sondas son cebadores inversos para la reacción en cadena de la polimerasa. En otra realización, la etapa de amplificación incluye realizar una reacción en cadena de la polimerasa, en el que el primer y tercer cebadores de amplificación son cebadores directos y el segundo y cuarto cebadores de amplificación son cebadores inversos para la reacción en cadena de la polimerasa. En algunas realizaciones, la etapa de amplificación comprende realizar una reacción en cadena de la ligasa. La etapa de amplificación puede incluir realizar una reacción en cadena de la polimerasa, una reacción en cadena de la polimerasa con transcriptasa inversa, una reacción en cadena de la ligasa o una reacción en cadena de la ligasa seguida de una reacción en cadena de la polimerasa.

En otra realización, el complejo fusionado es circular. En algunas realizaciones, la primera o segunda secuencia de ácido nucleico diana es una secuencia de ARN. La primera o segunda secuencia de ácido nucleico diana también puede ser una secuencia de ADN. En una realización, la primera o segunda secuencia de ácido nucleico diana comprende una secuencia de receptor de células T. En otra realización, la primera secuencia de ácido nucleico diana, la segunda secuencia de ácido nucleico diana o ambas secuencias de ácido nucleico diana comprende una secuencia de inmunoglobulina. En otras realizaciones, el primer ácido nucleico diana comprende una secuencia de receptor de células T, y la segunda secuencia de ácido nucleico diana comprende una segunda molécula que se asocia con la función celular inmune. En otra realización, la primera secuencia de ácido nucleico diana comprende una secuencia de inmunoglobulina, y la segunda secuencia comprende una segunda molécula asociada con la función celular inmune. En una realización, la segunda molécula asociada con la función celular inmune se selecciona del grupo que consiste en: interleucina-2 (IL-2), interleucina-4 (IL-4), interferón gamma (IFN $\gamma$ ), interleucina-10 (IL-10), interleucina-1 (IL-1), la interleucina-13 (IL-13), interleucina-17 (IL-17), interleucina-18 (IL-18), factor de necrosis tumoral alfa (TNF $\alpha$ ), factor de necrosis tumoral beta (TNF $\beta$ ), factor de transcripción 21 de la caja T (TBX21), caja forkhead P3 (FOXP3), cúmulo de diferenciación 4 (CD4), cúmulo de diferenciación 8 (CD8), cúmulo de diferenciación 1d (CD1d), cúmulo de diferenciación 161 (CD161), cúmulo de diferenciación 3 (CD3), complejo principal de histocompatibilidad (MHC), cúmulo de diferenciación 19 (CD19), receptor de interleucina 7 (receptor de IL-17), cúmulo de diferenciación 10 (CD10), cúmulo de diferenciación 20 (CD20), cúmulo de diferenciación 22 (CD22), cúmulo de diferenciación 34 (CD34), cúmulo de diferenciación 27 (CD27), cúmulo de diferenciación 5 (CD5) y cúmulo de diferenciación 45 (CD45), cúmulo de diferenciación 38 (CD38), cúmulo de diferenciación 78 (CD78), receptor de interleucina-6, factor regulador de interferón 4 (IRF4), cúmulo de diferenciación 138 (CD138).

En una realización, el primer o segundo ácido nucleico diana incluye una secuencia de genes raros. En algunas realizaciones, la secuencia del gen raro está presente en menos de 5% de las células, menos de 1% de las células o menos de 0,1% de las células. En una realización, la secuencia del gen raro resulta de una mutación genética. En otra realización, la mutación genética es una mutación somática. En una realización, la mutación genética es una mutación en un gen seleccionado del grupo que consiste en el receptor del factor de crecimiento epidermal (EGFR), homólogo de fosfatasa y tensina (PTEN), proteína de tumor 53 (p53), MutS homólogo 2 (MSH2), neoplasia endocrina múltiple 1 (MEN1), poliposis adenomatosa coli (APC), receptor Fas (FASR), proteína retinoblastoma (Rb1), Janus quinasa 2 (JAK2), factor de transcripción 1 similar a (ETS) (ELK1), virus de la eritroblastosis aviar v-ets E26 homólogo de oncogén 1 (ETS1), cáncer de mama 1 (BRCA1), cáncer de mama 2 (BRCA2), receptor del factor de crecimiento de hepatocitos (MET), proto-oncogén ret (RET), leucemia eritroblástica viral V-erb-b2 homólogo de oncogén 2 (HER2), sarcoma de rata viral V-Ki-ras2 Kirsten homólogo de oncogén (KRAS), linfoma de células B 2 (BCL2), mielocitomatosis viral V-myc homólogo de oncogén (MYC), gen de neurofibromatosis de tipo 2 (NF2), mieloblastosis viral v-myb homólogo de oncogén (MYB) y homólogo mutS 6 (E. coli) (MSH6). En otras realizaciones, la mutación se asocia con una enfermedad, y en una realización, la enfermedad es cáncer. En algunas realizaciones, el cáncer es un cáncer seleccionado del grupo que consiste en carcinoma de pulmón, cáncer de pulmón de células no pequeñas, cáncer de pulmón de células pequeñas, cáncer de útero, cáncer de tiroides, carcinoma de mama, carcinoma de próstata, carcinoma de páncreas, carcinoma de colon, linfoma, linfoma de

5 Burkitt, linfoma de Hodgkin, leucemia mieloide, leucemia, sarcoma, blastoma, melanoma, seminoma, cáncer de cerebro, glioma, glioblastoma, astrocitoma cerebeloso, linfoma de células T cutáneo, cáncer gástrico, cáncer de hígado, ependimoma, cáncer de laringe, cáncer de cuello, cáncer de estómago, cáncer de riñón, cáncer de páncreas, cáncer de vejiga, cáncer de esófago, cáncer testicular, meduloblastoma, cáncer vaginal, cáncer de ovario, cáncer cervical, carcinoma de células basales, adenoma de la pituitaria, rhabdomyosarcoma o sarcoma de Kaposi.

Además, el método puede incluir, en determinadas realizaciones, fijar y permeabilizar las células antes de realizar la etapa de amplificación. El método también puede incluir lisar las células antes de realizar la etapa de amplificación y cuantificar la información de la secuencia generada a partir de la reacción de secuenciación a granel.

10 En algunas realizaciones, el método para analizar la célula individual incluye una célula individual contenida dentro de una población de al menos 25.000 células, al menos 50.000 células, al menos 75.000 células o al menos 100.000 células. En algunas realizaciones, la célula individual es una célula única con respecto a las células que quedan en la población. En otras realizaciones, la célula individual es un representante de una sub-población de células dentro de la población. La población puede ser considerada en algunas realizaciones como el número total de células analizadas en un método de la invención.

15 En una realización, la realización de la reacción de secuenciación a granel para generar información de la secuencia se lleva a cabo para al menos 1.000.000 de complejos fusionados de al menos 10.000 células dentro de la población de células.

20 El método incluye proporcionar un segundo conjunto de sondas de ácido nucleico, comprendiendo el segundo conjunto una quinta sonda que comprende una secuencia que es complementaria a una tercera sub-secuencia de ácido nucleico diana, una sexta sonda que comprende una secuencia que es complementaria a una segunda sub-secuencia de la tercera secuencia de ácido nucleico diana y una segunda secuencia que es complementaria a una segunda secuencia exógena, una séptima sonda que comprende la secuencia exógena y una secuencia que es complementaria a una primera sub-secuencia de una cuarta secuencia de ácido nucleico diana, y una octava sonda que comprende una secuencia que es complementaria a una segunda sub-secuencia de la cuarta secuencia de ácido nucleico diana.

25 El método también proporciona aislar las células individuales con el primer y segundo conjunto de sondas de ácido nucleico; amplificar las tercera y cuarta secuencias de ácidos nucleicos diana de forma independiente, en el que la tercera secuencia de ácido nucleico diana se amplifica utilizando la quinta sonda y la sexta sonda, y en el que la cuarta secuencia de ácido nucleico diana se amplifica utilizando la séptima sonda y la octava sonda; hibridar la secuencia exógena a su complemento; amplificar la tercera secuencia de ácido nucleico diana, la cuarta secuencia de ácido nucleico diana y la secuencia exógena utilizando las quinta y octava sondas, generando de este modo un complejo fusionado; y realizando una reacción de secuenciación a granel para generar información de la secuencia para al menos 100.000 complejos fusionados de al menos 10.000 células dentro de la población de células, en el que la información de la secuencia es suficiente para co-localizar la primera secuencia de ácido nucleico diana, la segunda secuencia de ácido nucleico diana, la tercera secuencia de ácido nucleico diana y la cuarta secuencia de ácido nucleico diana de una célula individual de la población de al menos 10.000 células.

En algunos aspectos, la primera secuencia de ácido nucleico diana y la tercera secuencia de ácido nucleico diana son las mismas. En otros aspectos, la primera secuencia de ácido nucleico diana y la tercera secuencia de ácido nucleico diana son diferentes.

40 En otras realizaciones, el método incluye proporcionar N conjuntos de sondas de ácido nucleico, en el que cada uno de los N conjuntos comprende una sonda  $I_1$  que comprende una secuencia que es complementaria a una primera sub-secuencia de ácido nucleico diana  $I_a$ , comprendiendo una sonda  $I_2$  una secuencia que es complementaria a una segunda sub-secuencia de ácido nucleico diana  $I_b$ , y una segunda secuencia que es complementaria a una secuencia exógena  $I$ , comprendiendo una sonda  $I_3$  la secuencia exógena  $I$  y una secuencia que es complementaria a una primera sub-secuencia de ácido nucleico diana  $I_b$ , y una sonda  $I_4$  que comprende una secuencia que es complementaria a una segunda sub-secuencia de ácido nucleico diana  $I_b$ , en donde  $I$  oscila entre 1 y N.

50 El método también incluye aislar las células individuales con los N conjuntos de sondas de ácido nucleico; amplificar para todos los valores de  $I$ , las secuencias de ácido nucleico diana  $I_a$  e  $I_b$  de forma independiente, en el que la secuencia de ácido nucleico diana  $I_a$  se amplifica utilizando la sonda  $I_1$  y la sonda  $I_2$  y la secuencia de ácido nucleico diana  $I_b$  se amplifica utilizando la sonda  $I_3$  y la sonda  $I_4$ ; hibridar la secuencia exógena  $I$  a su complemento; amplificar para cada una de las  $I$  la secuencia diana  $I_a$ , la secuencia diana  $I_b$  y la secuencia exógena  $I$  utilizando las sondas  $I_1$  e  $I_4$ , generando con ello N complejos fusionados; y realizar una reacción de secuenciación a granel para generar información de la secuencia de al menos 100.000 complejos fusionados de al menos 10.000 células dentro de la población de células, en el que la información de la secuencia es suficiente para co-localizar la N secuencia de ácido nucleico diana  $I_a$  y la secuencia de ácido nucleico diana  $I_b$  a una célula individual de la población de al menos 10.000 células.

En otras realizaciones, N es menor que o igual a 10, menor que o igual a 100, menor que o igual a 1000, menor que o igual a 10.000, menor que o igual a 100.000, o N representa todos los transcritos poliadenilados en una célula.

5 En algunas realizaciones, el método incluye introducir una secuencia de código de barras única que comprende al menos seis nucleótidos en cada una de la pluralidad de células individuales, en el que cada una de las secuencias de código de barras se selecciona de una agrupación de secuencias de código de barras con una diversidad mayor que 1000 veces en las secuencias. Para cada una de la pluralidad de células individuales, el método incluye proporcionar al menos un conjunto de sondas de ácido nucleico. El método incluye las etapas de analizar al menos dos secuencias de ácido nucleico en una célula individual contenida dentro de una población de al menos 10.000 células, que comprende aislar cada una de una pluralidad de células individuales de una población de al menos 10.000 células en una microgotita de la emulsión o un recipiente de reacción. El método incluye introducir una secuencia de código de barras única que comprende al menos seis nucleótidos en cada una de la pluralidad de células individuales, en el que cada una de las secuencias de código de barras se selecciona de una agrupación de secuencias de código de barras con una diversidad de más de 1000 veces en la secuencia.

15 Para cada una de la pluralidad de células individuales, el método proporciona al menos un conjunto de sondas de ácido nucleico, comprendiendo el conjunto una primera sonda que comprende una secuencia que es complementaria a una secuencia de ácido nucleico que está localizada en el extremo 5' de la secuencia de código de barras, una segunda sonda que comprende una secuencia que es complementaria a una secuencia de ácido nucleico que está localizada en el extremo 3' de la secuencia de código de barras y una segunda región de la secuencia que es complementaria a una secuencia exógena, no humana, comprendiendo una tercera sonda una secuencia que comprende la secuencia exógena, no humana, y una secuencia que es complementaria a una primera sub-secuencia de una segunda secuencia de ácido nucleico diana, y una cuarta sonda que comprende una secuencia que es complementaria a una segunda sub-secuencia de la segunda secuencia de ácido nucleico diana.

20 El método continúa amplificando la primera y segunda secuencias de ácido nucleico de forma independiente, en el que la primera secuencia de ácido nucleico diana se amplifica utilizando la primera sonda y la segunda sonda, y en el que la segunda secuencia de ácido nucleico diana se amplifica utilizando la tercera sonda y la cuarta sonda; hibridando la secuencia exógena a su complemento; amplificando la primera secuencia de ácido nucleico diana, la segunda secuencia de ácido nucleico diana y la secuencia exógena utilizando la primera y cuarta sondas; realizando la secuenciación a granel de los complejos fusionados; e identificando una célula individual para cada uno de los complejos fusionados en base a la secuencia de código de barras.

25 En algunas realizaciones, la secuencia de código de barras se fija a una perla o a una superficie sólida. La perla o la superficie sólida se pueden aislar en la microgotita de la emulsión o el recipiente de reacción.

30 En otras realizaciones, el método incluye introducir una secuencia de código de barras única que comprende condensar la microgotita de la emulsión o un recipiente de reacción que comprende la célula individual con la microgotita de la emulsión o un recipiente de reacción que comprende la secuencia de código de barras fijada a la perla o a la superficie sólida. La segunda secuencia de ácido nucleico diana puede ser complementaria a una secuencia de ARN. La segunda secuencia de ácido nucleico diana puede ser complementaria a una secuencia de ADN.

35 En determinadas realizaciones, la amplificación comprende realizar una reacción en cadena de la polimerasa, realizar una reacción en cadena de la ligasa o realizar una reacción en cadena de la ligasa seguida de una reacción en cadena de la polimerasa.

40 En una realización, la célula individual está contenida dentro de una población de al menos 25.000 células. En otras realizaciones, la célula individual está contenida dentro de una población de al menos 50.000 células. La célula individual puede estar contenida dentro de una población de al menos 75.000 células o dentro de una población de al menos 100.000 células.

45 En determinadas realizaciones, el método incluye también cuantificar los complejos fusionados. En otras realizaciones, los complejos fusionados son circulares.

50 En un aspecto, el método incluye proporcionar N conjuntos de sondas de ácido nucleico, en el que cada uno de los N conjuntos comprende una sonda  $I_1$  que comprende una secuencia que es complementaria a una primera sub-secuencia de una secuencia de código de barras, comprendiendo una sonda  $I_2$  una secuencia que es complementaria a una segunda sub-secuencia de la secuencia de código de barras, y una segunda secuencia que es complementaria a una secuencia exógena  $I$ , comprendiendo una sonda  $I_3$  la secuencia exógena  $I$  y una secuencia que es complementaria a una primera sub-secuencia de ácido nucleico diana  $I_b$ , y una sonda  $I_4$  que comprende una secuencia que es complementaria a una segunda sub-secuencia de ácido nucleico diana  $I_b$ , en donde  $I$  oscila entre 1 y N.

55 El método también proporciona aislar las células individuales con los N conjuntos de sondas de ácido nucleico; amplificar para todos los valores de  $I$ , la secuencia de código de barras y las secuencias de ácidos nucleicos diana  $I_b$

de forma independiente, en el que la secuencia de código de barras se amplifica utilizando la sonda I<sub>1</sub> y la sonda I<sub>2</sub> y la secuencia de ácido nucleico diana I<sub>b</sub> se amplifica utilizando la sonda I<sub>3</sub> y la sonda I<sub>4</sub>; hibridar la secuencia exógena I a su complemento; amplificar para cada uno de los I la secuencia de código de barras, la secuencia diana I<sub>b</sub> y la secuencia exógena I utilizando las sondas I<sub>1</sub> e I<sub>4</sub>, generando con ello N complejos fusionados; y realizando una  
 5 reacción de secuenciación a granel para generar información de la secuencia para al menos 100.000 complejos fusionados de al menos 10.000 células dentro de la población de células, en el que la información de la secuencia es suficiente para co-localizar la secuencia de código de barras y la secuencia de ácido nucleico diana I<sub>b</sub> a una célula individual de la población de al menos 10.000 células.

En otros aspectos, N es menor que o igual a 10, menor que o igual a 100, menor que o igual a 1000, menor que o igual a 100.000, o N representa a todos los transcritos poliadenilados en una célula. En algunas realizaciones, la  
 10 secuencia de código de barras es la misma secuencia para todos los N.

En otras realizaciones, la invención también proporciona un método para introducir dichas secuencias de código de barras en recipientes de reacción o microgotitas de la emulsión. El método incluye proporcionar una agrupación de secuencias de código de barras únicas, en el que cada una de las secuencias de código de barras está enlazada a un gen de resistencia a la selección. El método también incluye proporcionar una población de células individuales, transfectar la población de células individuales con la agrupación de secuencias de código de barras únicas, seleccionar células que contienen una secuencia de código de barras única y el gen de resistencia a la selección, y aislar cada una de las células seleccionadas en recipientes de reacción o microgotitas de la emulsión. El gen de resistencia a la selección puede codificar resistencia a gentamicina, neomicina, higromicina o puromicina, por  
 15 ejemplo.  
 20

#### BREVE DESCRIPCIÓN DE LAS DIVERSAS VISTAS DE LOS DIBUJOS

Estas y otras características, aspectos y ventajas de la presente invención se entenderán mejor con respecto a la siguiente descripción y dibujos que se acompañan, en que:

La Figura (FIG.) 1A muestra un ejemplo de enlace de secuencia en una célula individual por circularización multi-sonda intracelular de un complejo molecular, de acuerdo con una realización de la invención. Cada una de las sondas tiene una región de complementariedad con cada uno de los loci diana. El complejo incluye dos sondas de ácido nucleico (a y b) y dos ácidos nucleicos diana (c y d). La célula individual (e) puede estar contenida en un recipiente de reacción o una gotita de emulsión (j).  
 25

La FIG. 1B ilustra un ejemplo de enlace de secuencia en una célula individual (también en un recipiente de reacción o gotita de emulsión (j)) por circularización multi-sonda intracelular de un complejo, de acuerdo con una realización de la invención. Las dos sondas de ácidos nucleicos (a y b) se hibridan a las regiones complementarias de los dos ácidos nucleicos diana (c y d).  
 30

La FIG. 1C ilustra un ejemplo de la circularización de un complejo de enlace sonda-diana que se produce por amplificación, de acuerdo con una realización de la invención.

La FIG. 2 es un ejemplo de la amplificación de un complejo de enlace sonda-diana circularizado (a) utilizando una polimerasa (b), de acuerdo con una realización de la invención. En algunas realizaciones, una  $\phi$ -29 polimerasa se utiliza en una amplificación mediada por círculo rodante, y se generan copias (b y c) del complejo sonda-diana circularizado.  
 35

La FIG. 3 ilustra un ejemplo de la amplificación de un complejo de enlace sonda-diana circularizado (a) utilizando una polimerasa (b) y cebadores (c y d), de acuerdo con una realización de la invención. Los cebadores (c y d) se utilizan para amplificar la región del complejo sonda-diana circularizado que es complementario al ácido nucleico diana. Se generan múltiples copias (e) de un amplicón lineal de ácido polinucleico de doble cadena y se secuencian a granel.  
 40

La FIG. 4 ilustra un ejemplo de la amplificación de un complejo de enlace sonda-diana circularizado (a) en una célula individual (b), de acuerdo con una realización de la invención. La amplificación se produce por transformación en bacterias y posterior selección con antibióticos. El amplicón (a) contiene un gen resistente a antibióticos, y células (c) que se transforman con el amplicón se seleccionan en presencia de antibióticos. No se seleccionan las células sin el complejo sonda-diana circularizado (d).  
 45

La FIG. 5A muestra un ejemplo de un enlace de secuencia de células individuales por reacción en cadena de la polimerasa de extensión por solapamiento intracelular, de acuerdo con una realización de la invención. Un cebador directo (a) fija como objetivo un locus de un primer ácido nucleico diana (g). Un cebador inverso (b) fija como objetivo otro locus del primer ácido nucleico diana (g) y tiene una región de complementariedad (c) con una región (d) del cebador directo (e). El cebador directo (e) tiene una región de complementariedad con el segundo ácido nucleico diana (h) y el cebador inverso (f) fija como objetivo otra región del segundo ácido nucleico diana (h). Las etapas de la FIG. 5 se pueden realizar en un recipiente de reacción o una gotita de la emulsión.  
 50  
 55

- La FIG. 5B ilustra un ejemplo de la hibridación de las sondas (a, b, e y f) a ácidos nucleicos diana respectivos (g y h) de acuerdo con una realización de la invención.
- La FIG. 6A ilustra un ejemplo de las regiones complementarias (c) y (d) entre los amplicones (g) y (h), de acuerdo con una realización de la invención. La FIG. 6B muestra la amplificación de enlace de los amplicones (g) y (h) utilizando la polimerasa (e) para crear un amplicón principal enlazado (i). El producto final es un banco de "amplicones principales" que incluye los amplicones enlazados (g) y (h), que pueden ser secuenciados a granel. Las etapas de la FIG. 6 se pueden realizar en un recipiente de reacción o una gotita de emulsión.
- Las FIGs. 7A y 7B ilustran un ejemplo de enlace de secuencia de célula individual por reacción en cadena de la ligasa intracelular combinada con reacción en cadena de la polimerasa de extensión por solapamiento, de acuerdo con una realización de la invención.
- La FIG. 8A muestra un ejemplo de las regiones complementarias entre amplicones (a) y (d), de acuerdo con una realización de la invención. La FIG. 8B muestra la amplificación del enlace de los amplicones utilizando la polimerasa (e) para crear un amplicón principal enlazado. Las etapas de las FIGs. 7 y 8 se pueden realizar en un recipiente de reacción o una gotita de emulsión.
- La FIG. 9A muestra un ejemplo de un amplicón enlazado (f), de acuerdo con una realización de la invención. La FIG. 9B muestra el amplicón resultante producido a partir de las etapas mostradas en las FIGs. 8A y 8B. El producto final puede ser un banco de "amplicones principales" y se pueden secuenciar a granel.
- La FIG. 10 ilustra un ejemplo de los componentes requeridos para un enlace de la secuencia de células individuales por sondas-candado en combinación con la reacción en cadena de la polimerasa de extensión por solapamiento, de acuerdo con una realización de la invención.
- La FIG. 11 muestra las regiones complementarias entre una primera sonda-candado (a) y el primer ácido nucleico diana (c) y entre una segunda sonda-candado (b) y un segundo ácido nucleico diana (d) en una célula individual, de acuerdo con una realización de la invención.
- La FIG. 12 ilustra los amplicones circularizados (g) y (h) resultantes y los cebadores que se utilizan para amplificar los amplicones circularizados, de acuerdo con una realización de la invención.
- La FIG. 13 muestra un ejemplo de los amplicones resultantes de la amplificación de las sondas circulares (g) y (h), de acuerdo con una realización de la invención.
- La FIG. 14 muestra un ejemplo de amplificación por PCR de extensión por solapamiento de los amplicones utilizando una polimerasa (e), de acuerdo con una realización de la invención.
- La FIG. 15 ilustra un ejemplo de desconvolución del banco de plásmidos por reacción en cadena de la polimerasa de extremo de cola con código de barras (con código de barras en el extremo 5'), que es seguida por secuenciación a granel y la informática, de acuerdo con una realización de la invención. La secuencia de código de barras se puede rastrear a una posición de pocillo y placa, la secuencia de código de barras se puede rastrear luego a una secuencia de ácido nucleico y la secuencia de ácido nucleico se rastrea a un pocillo. Cada uno de los cebadores en (a) y (b) tiene una etiqueta con código de barras en el extremo 5'. Los ácidos nucleicos diana en (c) y (d) se amplifican utilizando los cebadores en (a) y (b). Las etapas se pueden realizar en recipientes cerrados o gotitas de emulsión tal como se muestra en (c) y (d).
- La FIG. 16 muestra un ejemplo de amplificación (e, f) de dos ácidos nucleicos diana (A y B), utilizando cebadores que incluyen secuencias de códigos de barras, de acuerdo con una realización de la invención. Los amplicones resultantes que incluyen las secuencias de códigos de barras se muestran en (g) y (h).
- La FIG. 17 muestra un ejemplo simplificado de rastreo de una secuencia de código de barras en un amplicón a una diana celular (A o B), y el rastreo de nuevo de la diana celular a un lugar físico (c, d) (p. ej., un pocillo), de acuerdo con una realización de la invención.
- La FIG. 18 ilustra el enlace molecular entre dos transcripciones (g y h) y una secuencia de código de barras molecular (k), de acuerdo con una realización de la invención.
- La FIG. 19 muestra un ejemplo de la amplificación de los ácidos nucleicos diana (g y h) utilizando los cebadores tal como se muestra, de acuerdo con una realización de la invención.
- La FIG. 20 muestra un ejemplo de los amplicones que resultan después de la amplificación de dos ácidos nucleicos diana y una secuencia de código de barras (k), de acuerdo con una realización de la invención.
- La FIG. 21 ilustra un amplicón fusionado que incluye secuencias de dos ácidos nucleicos diana (g y h) y una secuencia de código de barras (k) dentro de una gotita de emulsión o recipiente de reacción (j), de acuerdo con una

realización de la invención. El amplicón ("principal") fusionado puede ser aislado por emulsión inversa y secuenciado a granel.

La FIG. 22 es un ejemplo de enlace molecular entre dos transcripciones (g y h) y una secuencia de código de barras molecular (k) unidas a una perla (m), de acuerdo con una realización de la invención.

5 La FIG. 23 ilustra los cebadores directo e inverso que se utilizan en un enlace molecular entre dos transcripciones (g y h) y una secuencia de código de barras molecular (k) unida a una perla (m), de acuerdo con una realización de la invención.

10 La FIG. 24 muestra un ejemplo de los amplicones que resultan después de la amplificación de dos ácidos nucleicos diana y una secuencia de código de barras (k) unida a una perla (m), de acuerdo con una realización de la invención.

La FIG. 25 ilustra un amplicón fusionado que incluye secuencias de dos ácidos nucleicos diana (g y h) y una secuencia de código de barras (k), dentro de una gotita de emulsión o recipiente de reacción (j), de acuerdo con una realización de la invención. El amplicón ("principal") fusionado puede ser aislado por emulsión inversa y secuenciado a granel.

15 La FIG. 26 es un ejemplo de un enlace de la secuencia de células individuales por reacción en cadena de la ligasa en combinación con la reacción en cadena de la polimerasa de extensión por solapamiento, tal como se aplica a un método para el diagnóstico prenatal no invasivo, de acuerdo con una realización de la invención.

20 La FIG. 27 muestra un ejemplo de hibridación de cebadores y ácidos nucleicos diana en un enlace de la secuencia de células individuales por reacción en cadena de la ligasa en combinación con la reacción en cadena de la polimerasa de extensión por solapamiento, tal como se aplica a un método para el diagnóstico prenatal no invasivo, de acuerdo con una realización de la invención. El proceso se lleva a cabo en una gotita de emulsión o recipiente de reacción (k).

25 La FIG. 28 muestra un ejemplo de amplicones resultantes producidos en un enlace de la secuencia de células individuales por reacción en cadena de la ligasa en combinación con la reacción en cadena de la polimerasa de extensión por solapamiento, tal como se aplica a un método para el diagnóstico prenatal no invasivo, de acuerdo con una realización de la invención.

La FIG. 29 muestra la hibridación de regiones complementarias solapantes de los amplicones resultantes y la reacción en cadena de la polimerasa de extensión por solapamiento, tal como se aplica a un método para el diagnóstico prenatal no invasivo, de acuerdo con una realización de la invención.

30 La FIG. 30 muestra los amplicones que resultan de la reacción en cadena de la polimerasa de extensión por solapamiento, tal como se aplica a un método para el diagnóstico prenatal no invasivo, de acuerdo con una realización de la invención. El producto final es un banco de "amplicones principales", o loci enlazados, que luego pueden ser secuenciados a granel.

35 La FIG. 31 muestra un flujo de trabajo simplificado para la generación de alto rendimiento de bancos de repertorio TCR $\beta$ , de acuerdo con una realización de la invención.

## DESCRIPCIÓN DETALLADA DE LA INVENCION

40 En síntesis, y como se describe con mayor detalle más adelante, se describen en esta memoria métodos y sistemas para el análisis genético paralelo masivo de células individuales en gotitas de emulsión o recipientes de reacción. Loci genéticos de interés son fijados como objetivo en una célula individual utilizando sondas diseñadas especialmente y un complejo de fusión está formado por técnicas de enlace molecular y amplificación. Múltiples loci genéticos pueden ser fijados como objetivo, y muchos conjuntos de sondas se pueden multiplexar por PCR en un solo análisis, de manera que se analizan varios loci o incluso todo el transcriptoma o genoma.

45 La invención es útil para analizar la información genética en células individuales en un alto rendimiento, de forma paralela para una gran cantidad de células ( $10^4$  o más células). La invención también es útil para el rastreo de la información genética de una célula o población de células utilizando secuencias de código de barras únicas.

### Definiciones

Los términos y expresiones utilizados en las reivindicaciones y la memoria descriptiva se definen como se indica a continuación, a menos que se especifique lo contrario.

50 El término "célula" se refiere a una unidad básica funcional de los organismos vivos. Una célula incluye cualquier tipo de célula (procariota o eucariota) de un organismo vivo. Ejemplos incluyen, pero no se limitan a células de sangre mononucleares de mamífero, células de levadura o células bacterianas.

La expresión "reacción en cadena de la polimerasa" o PCR se refiere a una técnica de biología molecular para la amplificación de una secuencia de ADN de una sola copia a varios órdenes de magnitud (de miles a millones de copias). La PCR se basa en la ciclación térmica, lo que requiere ciclos de calentamiento y enfriamiento repetidos de la reacción para la fusión del ADN y la replicación enzimática del ADN. Los cebadores (fragmentos cortos de ADN) que contienen secuencias complementarias a la región diana de la secuencia de ADN y una ADN polimerasa son componentes clave para permitir una amplificación selectiva y repetida. A medida que progresa la PCR, el ADN generado es en sí utilizado como un molde para la replicación, poniendo en marcha una reacción en cadena en la que el molde de ADN se amplifica de manera exponencial. Se utiliza una ADN polimerasa termoestable, tal como Taq polimerasa. Las etapas de ciclación térmica son necesarias primero para separar físicamente las dos cadenas en una doble hélice de ADN a una alta temperatura en un proceso denominado fusión del ADN. A una temperatura inferior, cada una de las cadenas se utiliza entonces como molde en la síntesis de ADN por la ADN polimerasa para amplificar selectivamente el ADN diana. La selectividad de la PCR resulta del uso de cebadores que son complementarios a la región de ADN fijada como objetivo para la amplificación bajo condiciones específicas de ciclación térmica.

La expresión "reacción en cadena de la polimerasa con transcriptasa inversa" o RT-PCR se refiere a un tipo de reacción PCR utilizada para generar múltiples copias de una secuencia de ADN. En la RT-PCR, una cadena de ARN se transcribe primero inversamente en su complemento de ADN (ADN complementario o ADNc) utilizando la enzima transcriptasa inversa, y el ADNc resultante se amplifica utilizando técnicas tradicionales de PCR.

La expresión "reacción en cadena de la ligasa" o LCR se refiere a un tipo de amplificación de ADN en el que dos sondas de ADN son ligadas por una ADN ligasa, y se utiliza una ADN polimerasa para amplificar el producto de ligamiento resultante. Se utilizan métodos de PCR tradicionales para amplificar la secuencia de ADN ligada. La LCR proporciona una mayor especificidad en comparación con la PCR.

La expresión "gotita de emulsión" o "microgotita de emulsión" se refiere a una gotita que se forma cuando se combinan dos fluidos inmiscibles. Por ejemplo, una gotita acuosa se puede formar cuando un fluido acuoso se mezcla con un fluido no acuoso. En otro ejemplo, un fluido no acuoso puede ser añadido a un líquido acuoso para formar una gotita. Las gotitas se pueden formar por diversos métodos, incluyendo métodos realizados por dispositivos de microfluidos u otros métodos, tales como inyectando un fluido a otro fluido, empujando o extrayendo líquidos a través de un orificio o abertura, formando gotitas por la fuerza de cizallamiento, etc. Las gotitas de una emulsión pueden tener cualquier distribución uniforme o no uniforme. Cualquiera de las emulsiones descritas en esta memoria puede ser monodispersa (compuesta de gotitas de tamaño al menos generalmente uniforme), o puede ser polidispersa (compuesta de gotitas de diversos tamaños). Si es monodispersa, las gotitas de la emulsión pueden variar en volumen por una desviación estándar que es menor que aproximadamente más o menos 100%, 50%, 20%, 10%, 5%, 2% o 1% del volumen medio de las gotitas. Las gotitas generadas a partir de un orificio pueden ser monodispersas o polidispersas. Una emulsión puede tener cualquier composición adecuada. La emulsión se puede caracterizar por el compuesto líquido predominante o el tipo de compuesto líquido que se utiliza. Los compuestos líquidos predominantes en la emulsión pueden ser agua y aceite. "Aceite" es cualquier compuesto líquido o mezcla de compuestos líquidos que es inmiscible con el agua y que tiene un alto contenido en carbono. En algunos ejemplos, el aceite también puede tener un alto contenido de hidrógeno, flúor, silicio, oxígeno, o cualquier combinación de los mismos, entre otros. Por ejemplo, cualquiera de las emulsiones descritas en esta memoria puede ser una emulsión de agua en aceite (W/O) (es decir, gotitas acuosas en una fase continua de aceite). El aceite puede ser o incluir al menos un aceite de silicona, aceite mineral, aceite fluorocarbonado, aceite vegetal, o una combinación de los mismos, entre otros. Cualesquiera otros componentes adecuados pueden estar presentes en cualquiera de las fases de la emulsión tal como al menos un tensioactivo, reactivo, muestra (es decir, particiones de la misma), tampón, sal, elemento iónico, otros aditivos, etiqueta, partículas, o cualquier combinación de los mismos.

"Gotita" se refiere a un pequeño volumen de líquido, típicamente con una forma esférica o como un disco que llena el diámetro de un microcanal, encapsulado por un fluido inmiscible. El volumen de una gotita y/o el volumen medio de gotitas en una emulsión puede ser menos de aproximadamente un microlitro (es decir, una "microgotita") (o entre aproximadamente un microlitro y uno nanolitro o entre aproximadamente un microlitro y un picolitro), menos de aproximadamente un nanolitro (o entre aproximadamente un nanolitro y un picolitro), o menos de aproximadamente un picolitro (o entre aproximadamente un picolitro y un femtolitro), entre otros. Una gotita puede tener un diámetro (o un diámetro medio) de menos de aproximadamente 1000, 100 ó 10 micras, o de aproximadamente 1000 a 10 micrómetros, entre otros. Una gotita puede ser esférica o no esférica. En algunas realizaciones, la gotita tiene un volumen y un diámetro que es lo suficientemente grande para encapsular una célula.

La expresión "código de barras" se refiere a una secuencia de ácido nucleico que se utiliza para identificar una célula individual o una sub-población de células. En algunas realizaciones, una secuencia de código de barras se utiliza para identificar un organismo o una especie particular. Tal como se describe a continuación, las secuencias de código de barras se pueden introducir en una célula, enlazar por diversos métodos de amplificación a un ácido nucleico diana de interés y utilizar para rastrear el amplicón a la célula. Secuencias de código de barras pueden estar flanqueadas por secuencias universales que pueden ser utilizadas para amplificar bancos de códigos de barras

utilizando pares de cebadores universales. Las secuencias de códigos de barras pueden estar contenidas dentro de una molécula de doble cadena circular o lineal, o en una molécula lineal de cadena sencilla.

5 La expresión "secuenciación a granel" o "secuenciación de próxima generación" o "secuenciación masiva en paralelo" se refiere a cualquier tecnología de secuenciación de alto rendimiento que paraleliza el proceso de secuenciación de ADN. Por ejemplo, métodos de secuenciación a granel son típicamente capaces de producir más de un millón de amplicones de ácido polinucleico en un único ensayo. Las expresiones "secuenciación a granel", "secuenciación masiva en paralelo" y "secuenciación de próxima generación" se refieren sólo a métodos generales, no necesariamente a la adquisición de más de 1 millón de etiquetas de secuencias en una sola pasada. Cualquier método de secuenciación a granel se puede implementar en la invención tal como la química de terminador reversible (p. ej., Illumina), pirosecuenciación utilizando gotitas de emulsión polony (p. ej., Roche), la secuenciación semiconductora de iones (Ion Torrent), la secuenciación de una sola molécula (p. ej., Pacific Biosciences), secuenciación masiva en paralelo de la firma, etc.

15 El término "in situ" se refiere al examen de un fenómeno biológico en el entorno en el que se produce, p. ej., la práctica de la hibridación in situ se refiere a la hibridación de una sonda a una diana de ácido nucleico con la célula aún intacta.

El término "in vivo" se refiere a los procesos que se producen en un organismo vivo.

El término "mamífero", tal como se utiliza en esta memoria, incluye tanto seres humanos como no humanos e incluye, pero no se limita a, seres humanos, primates no humanos, caninos, felinos, murinos, bovinos, equinos y porcinos.

20 La expresión "células T" se refiere a un tipo de célula que juega un papel central en la respuesta inmune mediada por células. Células T pertenecen a un grupo de glóbulos blancos conocidos como linfocitos y se pueden distinguir de otros linfocitos, tales como células B y células asesinas naturales T (NKT) por la presencia de un receptor de células T (TCR) en la superficie de la célula. Las respuestas de células T son específicas para el antígeno y son activadas por antígenos extraños. Las células T son activadas para proliferar y diferenciarse en células efectoras cuando el antígeno extraño se expone en la superficie de las células presentadoras de antígeno en órganos linfoides periféricos. Las células T reconocen fragmentos de antígenos de proteínas que han sido degradados en parte dentro de la célula presentadora de antígeno. Existen dos clases principales de células T - células T citotóxicas y células T colaboradoras. Las células T efectoras citotóxicas matan directamente a las células que están infectadas con un virus o algún otro patógeno intracelular. Las células T efectoras colaboradoras ayudan a estimular las respuestas de otras células, principalmente macrófagos, células B y células T citotóxicas.

35 La expresión "célula B" se refiere a un tipo de linfocito que juega un papel importante en la respuesta inmune humoral (en oposición a la respuesta inmune mediada por células, que se rige por las células T). Las principales funciones de las células B son producir anticuerpos contra los antígenos, realizar el papel de las células presentadoras de antígeno (APCs) y, finalmente, convertirse en células B de memoria después de la activación por la interacción del antígeno. Las células B son un componente esencial del sistema inmune adaptativo.

Hay que señalar que, tal como se utiliza en la memoria descriptiva y las reivindicaciones adjuntas, las formas en singular "un", "una" y "el", "la" incluyen referentes en plural, a menos que el contexto indique claramente lo contrario.

## **Métodos de la invención**

### **I. Métodos de Análisis Molecular de Células Individuales Masivo Paralelo**

#### **A. Métodos de Microfluidos para Generar Gotitas de Emulsión de Células Individuales**

40 En algunas realizaciones, se utiliza un dispositivo de microfluidos para generar gotitas de emulsión de células individuales. El dispositivo de microfluidos expulsa células individuales en tampón de reacción acuoso en una mezcla de aceite hidrofóbica. El dispositivo puede crear miles de microgotitas de emulsión por minuto. Después de crear las microgotitas de emulsión, el dispositivo expulsa la mezcla de emulsión en una artesa. La mezcla puede ser pipeteada o recogida en un tubo de reacción estándar para la termociclación.

45 Dispositivos de microfluidos habituales para el análisis de células individuales se fabrican de forma rutinaria en los laboratorios académicos y comerciales (Kintses *et al.*, 2010 *Current Opinion in Chemical Biology* 14:548-555). Por ejemplo, pueden fabricarse chips de polidimetilsiloxano (PDMS), de plástico, vidrio o cuarzo. En algunas realizaciones, el fluido se mueve a través de los chips a través de la acción de una bomba de presión o jeringa. Las células individuales pueden incluso ser manipuladas en los chips de microfluidos programables utilizando un dispositivo de dielectroforesis habitual (Hunt *et al.*, 2008 *Lab Chip* 8:81-87). En una realización, se utiliza un chip de PDMS basado en presión que comprende una geometría de enfoque de flujo fabricada con tecnología litográfica suave (Dolomite Microfluidics (Royston, Reino Unido)) (Anna *et al.*, 2003 *Applied Physics Letters* 82:364-366). El diseño stock normalmente puede generar 10.000 microgotitas acuosas en aceite por segundo a intervalos de

tamaño de 10-150  $\mu\text{m}$  de diámetro. En algunas realizaciones, la fase hidrofóbica consistirá en aceite fluorado que contiene una sal de amonio de carboxi-perfluoropoliéter, que garantiza condiciones óptimas para la biología molecular y disminuye la probabilidad de coalescencia de las gotitas (Johnston *et al.*, 1996 *Science* 271:624-626). Para medir la periodicidad de la célula y el flujo de gotitas, las imágenes se registran a 50.000 fotogramas por segundo utilizando técnicas estándares tales como una cámara Phantom V7 o Fastec InLine (Abate *et al.*, 2009 *Lab Chip* 9:2628-31).

El sistema de microfluidos puede optimizar el tamaño de las microgotitas, la densidad de células de entrada, el diseño de los chips y los parámetros de carga de células de modo que más del 98% de las gotitas contiene una sola célula. Existen tres métodos comunes para la consecución de tales estadísticas: (i) la dilución extrema de la disolución de células; (ii) la selección fluorescente de gotitas que contienen células individuales; y (iii) la optimización de la periodicidad de la entrada de células. Para cada uno de los métodos, las métricas para el éxito incluyen: (i) la tasa de encapsulación (es decir, el número de gotitas que contienen exactamente una célula); (ii) el rendimiento (es decir, la fracción de la población original celular que termina en una gotita que contiene exactamente una célula); (iii) la tasa de multi-hit (es decir, la fracción de gotas que contienen más de una célula); (iv) la tasa negativa (es decir, la fracción de gotitas que no contienen células); y (v) la tasa de encapsulación por segundo (es decir, el número de gotitas que contienen células individuales formadas por segundo).

En algunas realizaciones, las emulsiones de células individuales se generan por dilución de células extremas. En condiciones desordenadas, la probabilidad de que una microgotita contenga  $k$  células viene dada por la distribución de Poisson:

$$f(k, \lambda) = \frac{\lambda^k e^{-\lambda}}{k!},$$

en que  $e$  es el logaritmo natural y el número esperado de ocurrencias en el intervalo es  $\lambda$ . Así, para  $P(k = 1) \approx 0,98$ , la disolución de células debe ser extremadamente diluida, de tal manera que  $\lambda \approx 0,04$  y sólo 3,84% de todas las gotitas contiene una célula individual.

En algunas realizaciones, se utiliza un simple chip de microfluidos con una unión formadora de gotitas, de manera que una corriente acuosa fluye a través de una boquilla cuadrada de 10  $\mu\text{m}$  y dispensa las mezclas acuosas en aceite de la emulsión en un depósito. La mezcla de emulsión puede ser pipeteada desde el depósito y termociclada en tubos de reacción estándares. Este método producirá previsiblemente altas tasas de encapsulación y bajas tasas multi-hit, pero una tasa de encapsulación baja por segundo. Un diseño que puede alcanzar el rendimiento de gotita llena de 1000 Hz es capaz de clasificar hasta  $10^6$  células en menos de 17 minutos.

También se pueden utilizar técnicas de fluorescencia para clasificar microgotitas con características de emisión particulares (Baroud *et al.*, 2007 *Lab Chip* 7:1029-1033; Kintses *et al.*, 2010 *Current Opinion in Chemical Biology* 14:548-555). En estos estudios, los métodos químicos se utilizan para teñir las células. En algunas realizaciones, la autofluorescencia se utiliza para seleccionar microemulsiones que contienen células. Un detector fluorescente reduce la tasa negativa que resulta de la dilución de células extrema. Un dispositivo microfluídico también puede estar equipado con un láser dirigido a una unión de clasificación "Y" aguas abajo de la unión de encapsulación de células. La unión Y tiene un canal de "mantenimiento" y un canal de "residuos". Se utiliza un tubo fotomultiplicador para recoger la fluorescencia de cada una de las gotitas a medida que pasan por el láser. La diferencia de tensión se calibra entre gotitas vacías y gotitas con al menos una célula. A continuación, cuando el dispositivo detecta una gotita que contiene al menos una célula, y los electrodos en la unión de clasificación Y crean un gradiente de campo por dielectroforesis (Hunt *et al.*, 2008 *Lab on a Chip* 8:81-87) y empujan gotitas que contienen células en el canal de mantenimiento. El dispositivo de microfluidos utiliza dilución celular extrema para controlar la tasa multi-hit y de clasificación de células fluorescentes para reducir la tasa negativa.

En algunas realizaciones, el flujo de células de entrada está alineado con la periodicidad de la formación de gotitas, de manera que más de un 98% de las gotitas contiene una sola célula (Edd *et al.*, 2008 *Lab Chip* 8:1262-1264; Abate *et al.*, 2009 *Lab Chip* 9:2628-31). En estos dispositivos de microfluidos, una suspensión de alta densidad de las células es forzada a pasar a través de un canal de alto aspecto-relación, de manera que el diámetro de la célula es una fracción grande de la anchura del canal. El chip está diseñado con un microcanal rectangular de  $27\mu\text{m} \times 52\mu\text{m}$  que hace fluir células en microgotitas a  $> 10\mu\text{L}/\text{min}$  (Edd *et al.*, 2008 *Lab Chip* 8:1262-1264). Se testa un cierto número de anchuras de canal de entrada y caudales para llegar a una solución óptima.

En algunas realizaciones, las células con diferente morfología pueden comportarse de forma diferente en la corriente de microcanales del dispositivo de microfluidos, confundiendo la optimización de la técnica cuando se aplica a muestras biológicas clínicas. Para abordar esta cuestión, se induce un gradiente de campo perpendicular al microcanal por dielectroforesis. La dielectroforesis tira de las células a un lado del microcanal, creando una ordenación en los canales que es independiente de la morfología de las células. Este método requiere la

optimización sustancial de la carga y el caudal y un chip y un dispositivo de diseño más complicado, por lo que este método puede ser necesario si las metodologías existentes fracasan en el funcionamiento para ciertos tipos de células.

5 Las mezclas de microgotitas de emulsión se pipetea desde la artesa en el dispositivo de microfluidos a un tubo de reacción para la termociclación. Después de termociclar las emulsiones, un cierto número de métodos podría conseguir la inversión de la emulsión para recuperar la fase acuosa de la reacción. Dos procesos de reversión directos que han sido utilizados por investigadores anteriores son la congelación instantánea en nitrógeno líquido durante 10 segundos (Kliss *et al.*, 2008 *Analytical Chem.* 80:8975-8981) y el paso a través de un filtro de malla de 15  $\mu\text{m}$  (Zeng *et al.*, 2010 *Analytical Chem* 82:3183-90). La inversión de la emulsión también se puede lograr utilizando reactivos comercialmente disponibles diseñados para este propósito (Brouzes *et al.*, 2009 *PNAS* 106:14195-200). El éxito de la inversión de la emulsión se evaluó mediante visualización de las fases acuosa e hidrofóbicas bajo un microscopio.

15 En algunas realizaciones, los métodos de la invención utilizan células individuales en recipientes de reacción, en lugar de gotitas de la emulsión. Ejemplos de recipientes de reacción de este tipo incluyen placas de 96 pocillos, tubos de 0,2 mL, tubos de 0,5 mL, tubos de 1,5 mL, placas de 384 pocillos, placas de 1536 pocillos, etc.

## **B. Métodos para el Enlace Molecular en Emulsiones de Células Individuales y Secuenciación Masiva en Paralelo**

### **1) Enlace Molecular utilizando la Reacción en Cadena de la Polimerasa (PCR)**

20 La PCR se utiliza para amplificar muchos tipos de secuencias, incluyendo pero no limitadas a SNPs, repeticiones cortas en tándem (STR), dominios variables de proteínas, regiones metiladas y regiones intergénicas. Los métodos para la PCR de extensión por solapamiento se utilizan para crear productos de amplicón de fusión de varios loci genómicos independientes en un solo tubo de reacción (Johnson *et al.*, 2005 *Genome Research* 15:1315-24; Patente de EE.UU. 7.749.697).

25 En algunas realizaciones, al menos dos secuencias diana de ácido nucleico (p. ej., primera y segunda secuencias de ácido nucleico diana, o primer y segundo loci) se eligen en la célula y se designan como loci diana. Se diseñan cebadores directos e inversos para cada una de las dos secuencias de ácido nucleico diana, y los cebadores se utilizan para amplificar las secuencias diana. Se generan amplicones "menores" mediante la amplificación de las dos secuencias de ácido nucleico diana por separado, y luego se fusionan mediante amplificación para crear un amplicón de fusión, también conocido como un amplicón "principal". En una realización, un amplicón "menor" es una secuencia de ácido nucleico amplificada a partir de loci genómicos diana, y un amplicón "principal" es un complejo de fusión generado a partir de secuencias amplificadas entre loci genómicos múltiples. Cebadores a modo de ejemplo que se pueden utilizar para generar amplicones menores y principales se enumeran en la Tabla 2. Estos cebadores se utilizan para la amplificación multiplexada de TCR $\beta$  de una célula individual y después enlace del TCR $\beta$  a las dianas efectoras inmunes IL-2, IL-4, INFG, TBX21, FOXP3 o TNFA. En una realización, SEQ ID NOs: 1-30 35 57 se agrupan junto con los cebadores para una única diana efectora inmune, p. ej., SEQ ID NOs: 68 y 69.

40 El método utiliza cebadores "internos" (es decir, el cebador inverso para el primer locus y el cebador directo para el segundo locus) que comprenden un dominio que se hibrida con un amplicón menor y un segundo dominio que se hibrida con un segundo amplicón menor. Cebadores "internos" son un reactivo limitante, de manera que durante la fase exponencial de la PCR, los cebadores internos se han agotado, conduciendo a que dominios solapantes en los amplicones menores se reasocian y creen amplicones principales.

45 Los cebadores de la PCR se diseñan contra dianas de interés utilizando parámetros estándares, es decir, la temperatura de fusión ( $T_m$ ) de aproximadamente 55-65°C, y con una longitud de 20-50 nucleótidos. Los cebadores se utilizan con condiciones estándares de la PCR, por ejemplo, Tris-HCl 1 mM pH 8,3, cloruro de potasio 5 mM, cloruro de magnesio 0,15 mM, cebadores 0,2-2  $\mu\text{M}$ , dNTPs 200  $\mu\text{M}$  y una ADN polimerasa termoestable. Están disponibles muchos kits comerciales para realizar la PCR tales como Platinum Taq (Life Technologies), AmpliTaq Gold (Life Technologies), Titanium Taq (Clontech), Phusion polymerase (Finnzymes), HotStartTaq Plus (Qiagen). Cualquier ADN polimerasa termoestable estándar se puede utilizar para esta etapa, tal como Taq polimerasa o el fragmento Stoffel.

50 En una realización, se utiliza un conjunto de sondas (o cebadores) de ácidos nucleicos para amplificar una primera secuencia de ácido nucleico diana y una segunda secuencia de ácido nucleico diana para formar un complejo de fusión. La primera sonda incluye una secuencia que es complementaria a una primera secuencia de ácido nucleico diana (p. ej., el extremo 5' de la primera secuencia de ácido nucleico diana). La segunda sonda incluye una secuencia que es complementaria a la primera secuencia de ácido nucleico diana (p. ej., el extremo 3' de la primera secuencia de ácido nucleico diana) y una segunda secuencia que es complementaria a una secuencia exógena. En algunas realizaciones, la secuencia exógena es una secuencia de ácido nucleico no humana y no es 55

complementaria a cualquiera de las secuencias de ácidos nucleicos diana. La primera y segunda sondas son el cebador directo y el cebador inverso para la primera secuencia de ácido nucleico diana.

5 La tercera sonda incluye una secuencia que es complementaria a la porción de la segunda sonda que es complementaria a la secuencia exógena y una secuencia que es complementaria a la segunda secuencia de ácido nucleico diana (p. ej., el extremo 5' de la segunda secuencia de ácido nucleico diana). La cuarta sonda incluye una secuencia que es complementaria a la segunda secuencia de ácido nucleico diana (p. ej., el extremo 3' de la segunda secuencia de ácido nucleico diana). La tercera sonda y la cuarta sonda son los cebadores directo e inverso para la segunda secuencia de ácido nucleico diana.

10 La segunda y tercera sonda también se denominan los cebadores "internos" de la reacción (es decir, el cebador inverso para el primer locus y el cebador directo para el segundo locus) y son limitantes de la concentración (p. ej., 0.01  $\mu\text{M}$  para los cebadores internos y 0.1  $\mu\text{M}$  para todos los otros cebadores). Esto impulsará la amplificación del amplicón principal preferentemente sobre los amplicones menores. La primera y cuarta sondas se denominan los cebadores "externos".

15 La primera y segunda secuencias de ácido nucleico se amplifican de forma independiente, de manera que la primera secuencia de ácido nucleico se amplifica utilizando la primera sonda y la segunda sonda, y la segunda secuencia de ácido nucleico se amplifica utilizando la tercera sonda y la cuarta sonda. A continuación, se genera un complejo de fusión mediante la hibridación de las regiones de secuencias complementarias de las primera y segunda secuencias de ácido nucleico amplificadas y amplificando las secuencias hibridadas utilizando la primera y cuarta sondas. Esto se denomina la amplificación de PCR de extensión por solapamiento.

20 Durante la amplificación de PCR de extensión por solapamiento, las regiones de secuencias complementarias de las secuencias de ácidos nucleicos primera y segunda amplificadas actúan como cebadores para la extensión en ambas cadenas y en cada dirección por moléculas de ADN polimerasa. En ciclos de PCR posteriores, los cebadores externos ceban la secuencia condensada completa, de modo que el complejo fusionado se duplica por la ADN polimerasa. Este método produce una pluralidad de complejos de fusión.

25 Las FIGs. 5-6 muestran un ejemplo del enlace de la secuencia de células individuales por reacción en cadena de la polimerasa de extensión por solapamiento intracelular, de acuerdo con una realización de la invención. En la FIG. 5A, un cebador directo (a) fija como objetivo un locus de un primer ácido nucleico diana (g). Un cebador inverso (b) fija como objetivo otro locus del primer ácido nucleico diana (g) y tiene una región de complementariedad (c) a una región (d) del cebador directo (e). El cebador directo (e) tiene una región de complementariedad con el segundo ácido nucleico diana (h) y el cebador inverso (f) fija como objetivo otra región del segundo ácido nucleico diana (h). La FIG. 5B ilustra un ejemplo de la hibridación de las sondas (a, b, e y f) a ácidos nucleicos diana respectivos (g y h), de acuerdo con una realización de la invención. La FIG. 6A ilustra un ejemplo de las regiones complementarias (c) y (d) entre los amplicones (g) y (h), de acuerdo con una realización de la invención. La FIG. 6B muestra la amplificación de enlace de los amplicones (g) y (h) utilizando la polimerasa (e) para crear un amplicón principal enlazado (i). El producto final es un banco de "amplicones principales" que incluyen los amplicones enlazados (g) y (h), que pueden ser secuenciados a granel. Las etapas de las FIGs. 5-6 se pueden realizar en un recipiente de reacción o una gotita de emulsión.

30 En algunas realizaciones, múltiples loci son fijados como objetivo en una célula individual, y muchos conjuntos de sondas se pueden multiplexar en un solo análisis, de tal manera que se analizan varios loci o incluso todo el transcriptoma o genoma. La PCR multiplex es una modificación de la PCR que utiliza múltiples conjuntos de cebadores dentro de una única mezcla de PCR para producir amplicones de diferentes tamaños que son específicos para diferentes secuencias de ADN. Al fijar como objetivo múltiples genes a la vez, se puede conseguir información adicional a partir de una única prueba que de otro modo requeriría varias veces los reactivos y más tiempo para llevarse a cabo. En una realización, 10-20 transcripciones diferentes están fijadas como objetivo en una sola célula y están enlazadas a un segundo ácido nucleico diana (p. ej., enlazadas a una región variable tal como una secuencia mutada del gen, un código de barras o una región variable inmune).

35 En una realización, las células individuales se encapsulan en microgotas de picolitros acuosas-en-aceite. Las gotitas permiten la compartimentación de las reacciones de tal manera que la biología molecular se puede realizar en millones de células individuales en paralelo. Microgotas acuosas-en-aceite monodispersas se pueden generar en dispositivos de microfluidos en intervalos de tamaño de 10-150  $\mu\text{m}$  de diámetro. Alternativamente, las gotitas pueden ser generadas mediante vórtice o por un TissueLyser (Qiagen). Dos realizaciones de disoluciones de aceite y acuosas para la creación de microgotitas de PCR son: (i) tampón de PCR que contiene 0,5  $\mu\text{g}/\mu\text{L}$  de albúmina de suero bovino (New England Biolabs) en combinación con una mezcla de aceite fluorocarbonado (3M), tensioactivo Krytox 157FSH (Dupont) y PicoSurf (Sphere Microfluidics); y (ii) tampón PCR con Tween 20 al 0,1% (Sigma) combinado con una mezcla de aceite mineral ligero (Sigma), EM90 (Evonik), y Triton X-100 (Sigma). Varios ensayos replicados que cuantifican 1 millón de los amplicones mediante secuenciación de la siguiente generación han demostrado que ambas químicas forman microgotitas monodispersas que son > 99,98% estables después de 40 ciclos de PCR. La PCR puede producirse en un tubo estándar de termociclación, una placa de 96 pocillos o una

placa de 384 pocillos, utilizando un termociclador estándar (Life Technologies). La PCR también puede producirse en chips de microfluidos calentados, o cualquier otro tipo de recipiente que pueda mantener la emulsión y la transferencia de calor.

5 Después de la termociclación y la PCR, el material amplificado debe ser recuperado de la emulsión. En una realización, el éter se utiliza para romper la emulsión, y luego se evapora el éter de la capa acuosa/éter para recuperar el ADN amplificado en disolución. Otros métodos incluyen la adición de un tensioactivo a la emulsión, congelación instantánea con nitrógeno líquido y centrifugación.

10 Una vez que los productos enlazados y amplificados se recuperan de la emulsión, existe un cierto número de métodos para preparar el producto para la secuenciación a granel. En una realización, el amplicón principal se aísla de los amplicones menores utilizando electroforesis en gel. Si el rendimiento no es suficiente, el amplicón principal se amplifica de nuevo utilizando la PCR y los dos cebadores externos. Este material entonces se puede secuenciar directamente mediante secuenciación a granel. En algunas realizaciones, los cebadores externos se utilizan para producir moléculas que pueden ser secuenciadas directamente. En otras realizaciones, se deben agregar adaptadores al amplicón principal antes de la secuenciación a granel. Una vez que se sintetiza el banco de  
15 secuenciación, se puede realizar la secuenciación a granel utilizando métodos estándares y sin modificación significativa.

### **2) Enlace Molecular Utilizando la Reacción en Cadena de la Polimerasa con Transcriptasa Inversa (RT-PCR)**

20 El método de la PCR de extensión por solapamiento se adapta a la RT-PCR de extensión por solapamiento individual en tubo, que amplifica el ADN a partir de las transcripciones de ARN. El método RT-PCR combina la síntesis de ADNc y PCR en tubos cerrados sin intercambio de tampón o adición de reactivos entre las etapas moleculares. Se utilizan enzimas termoestables de la transcriptasa inversa (RT) que soportan temperaturas superiores a 95°C, aunque la RT termoestable no es necesaria si la síntesis de ADNc de la primera cadena se produce antes de la amplificación por PCR. Por ejemplo, tanto ThermoScript RT (Lucigen) como GeneAmp Thermoestable rTth (Life Technologies) están diseñados y se utilizan en la PCR con transcriptasa inversa en un solo tubo. En una realización, se utiliza un conjunto de sondas (o cebadores) de ácidos nucleicos para amplificar una  
25 primera secuencia de ácido nucleico diana y una segunda secuencia de ácido nucleico diana para formar un complejo de fusión. La primera secuencia de ácido nucleico diana o la segunda secuencia de ácido nucleico diana es ARN.

30 La primera sonda incluye una secuencia que es complementaria a una primera secuencia de ácido nucleico diana (p. ej., el extremo 5' de la primera secuencia de ácido nucleico diana). La segunda sonda incluye una secuencia que es complementaria a la primera secuencia de ácido nucleico diana (p. ej., el extremo 3' de la primera secuencia de ácido nucleico diana) y una segunda secuencia que es complementaria a una secuencia exógena. En algunas realizaciones, la secuencia exógena es una secuencia de ácido nucleico no humana y no es complementaria a cualquiera de las secuencias de ácidos nucleicos diana. La primera y segunda sondas son el cebador directo y el  
35 cebador inverso para la primera secuencia de ácido nucleico diana.

La tercera sonda incluye una secuencia que es complementaria a la porción de la segunda sonda que es complementaria a la secuencia exógena y una secuencia que es complementaria a la segunda secuencia de ácido nucleico diana (p. ej., el extremo 5' de la segunda secuencia de ácido nucleico diana). La cuarta sonda incluye una secuencia que es complementaria a la segunda secuencia de ácido nucleico diana (p. ej., el extremo 3' de la  
40 segunda secuencia de ácido nucleico diana). La tercera sonda y la cuarta sonda son los cebadores directo e inverso para la segunda secuencia de ácido nucleico diana.

45 Las segunda y tercera sondas también se denominan los cebadores "internos" de la reacción (es decir, el cebador inverso para el primer locus y el cebador directo para el segundo locus) y son limitantes en la concentración (p. ej., 0.01µM para los cebadores internos y 0.1 µM para todos los otros cebadores). Esto impulsará la amplificación del amplicón principal preferentemente frente a los amplicones menores. La primera y cuarta sondas se denominan los cebadores "externos".

50 El método incluye amplificar mediante RT-PCR la primera y segunda secuencias de ácido nucleico de forma independiente, de manera que la primera secuencia de ácido nucleico se amplifica utilizando la primera sonda y la segunda sonda, y la segunda secuencia de ácido nucleico se amplifica utilizando la tercera sonda y la cuarta sonda. Utilizando la amplificación por PCR de extensión por solapamiento, se genera un complejo de fusión mediante la hibridación de las regiones de secuencias complementarias de las secuencias primera y segunda de ácido nucleico amplificadas y la amplificación de las secuencias hibridadas utilizando la primera y cuarta sondas. (Véanse las FIGs. 5-6).

### **3) Enlace Molecular Utilizando la Reacción en Cadena de la Ligasa**

55 La reacción en cadena de la ligasa (LCR) se utiliza para fijar como objetivo y amplificar loci genéticos de interés (Landegren *et al.*, 1988 *Science* 241:1077-1080. Benjamin *et al.*, 2003 *Methods in Molecular Biology* 226:135-149;

Patente de EE.UU. 6.235.472). En la reacción en cadena de la ligasa, dos sondas de ácido polinucleico fijan como objetivo un locus de ácido polinucleico de interés. Tras la hibridación, las dos sondas se ligan mediante una enzima ligasa. En contraposición con la PCR, la LCR amplifica tanto ARN como ADN, facilitando muchos tipos diferentes de análisis multiplexado. Otra ventaja notable de la reacción en cadena de la ligasa es la capacidad para la  
 5 amplificación específica para alelos. Mientras que la PCR amplifica ambos alelos para una variante particular, el proceso de ligamiento de la LCR es alelo-específico.

En algunas realizaciones, sondas de LCR se utilizan como un "interruptor" molecular. Por ejemplo, si millones de células individuales son rastreados para detectar una variante particular, sólo las células que incluyen esa variante producirán amplicones principales. La LCR se utiliza para realizar análisis genéticos sólo en células que contienen  
 10 una secuencia particular de interés. Células que carecen de la secuencia de interés no se amplifican sustancialmente y, por lo tanto, son silenciosas en la reacción. La LCR también se puede multiplexar de forma más eficiente que la PCR, utilizando cientos de sondas que fijan como objetivo cientos de loci genéticos en una microgotita de una célula individual o reacción intracelular.

En una realización, se formula una única mezcla de reacción por LCR/PCR de extensión por solapamiento de tampón de un solo tubo utilizando ADN y/o ARN, sondas de LCR, los cebadores de PCR Ampligase (Epicentre), una ADN polimerasa tal como el fragmento Stoffel (Life Technologies) y tampón de reacción (Tris-HCl 20 mM, KCl 25 mM, MgCl<sub>2</sub> 10 mM, NAD 0,5 mM, Triton X-100 al 0,01%). El método combina la LCR con la PCR de extensión por solapamiento para aprovechar los beneficios tanto de la LCR como de la PCR (FIGs. 7-9). Las sondas "internas" se añaden en 1/10 de la concentración de los otros oligonucleótidos en la reacción de tal manera que se convierten en un reactivo limitante en ciclos posteriores. Para la reasociación inicial y el ligamiento, las mezclas se pueden incubar durante 4 minutos a 20°C, durante 5 minutos a 95°C y durante 15 minutos a 60°C. Condiciones estándares de termociclación de PCR se utilizan para amplificar los amplicones menores y principales (95°C, 5 minutos; [95°C, 30 segundos; 60°C, 30 segundos; 72°C, 30 segundos] x 30 ciclos). El amplicón principal se amplifica adicionalmente mediante selección por tamaño de gel y otra ronda de amplificación utilizando sólo los cebadores externos. Las FIGs. 7A y 7B ilustran un ejemplo de enlace de secuencia célula individual por reacción en cadena de la ligasa intracelular combinada con reacción en cadena de la polimerasa de extensión por solapamiento, de acuerdo con una realización de la invención. Un cebador de LCR directo (a) fija como objetivo un locus de un primer ácido nucleico diana (g). Un cebador de LCR inverso (b) fija como objetivo otro locus del primer ácido nucleico diana (g) y tiene una región de complementariedad (c) a una región (d) del cebador directo (e). El cebador de LCR directo (e) tiene una  
 15 20 25 30 región de complementariedad con el segundo ácido nucleico diana (h) y el cebador de LCR inverso (f) fija como objetivo otra región del segundo ácido nucleico diana (h).

La FIG. 8A muestra otro ejemplo de las regiones complementarias entre amplicones (a) y (d), de acuerdo con una realización de la invención. La FIG. 8B muestra la amplificación del enlace de los amplicones utilizando la polimerasa (e) para crear un amplicón principal enlazado. Las etapas de las FIGs. 7 y 8 se pueden realizar en un  
 35 recipiente de reacción o una gotita de emulsión.

La FIG. 9A muestra un ejemplo de un amplicón enlazado (f) de acuerdo con una realización de la invención. La FIG. 9B muestra el amplicón resultante, producido a partir de las etapas mostradas en las FIGs. 8A y 8B. El producto final puede ser un banco de "amplicones principales" y se secuencian a granel.

En otra realización, el enlace de la secuencia de células individuales por reacción en cadena de la ligasa intracelular combinada con la reacción en cadena de la polimerasa de extensión por solapamiento se lleva a cabo con el siguiente conjunto de sondas: una primera sonda de LCR que comprende una secuencia que es complementaria a una primera sub-secuencia de ácido nucleico diana, una segunda sonda que comprende una secuencia que es complementaria a una segunda sub-secuencia del primer ácido nucleico diana y una segunda secuencia que es complementaria a una secuencia exógena, una tercera sonda que comprende la secuencia exógena y una  
 40 45 secuencia que es complementaria a una primera sub-secuencia de un segundo ácido nucleico diana, y una cuarta sonda que comprende una secuencia que es complementaria a una segunda sub-secuencia de la segunda secuencia de ácido nucleico diana. El método incluye aislar las células individuales con al menos un conjunto de sondas de ácido nucleico. La primera y segunda sondas se hibridan al primer ácido nucleico y se ligan mediante una enzima ligasa. De manera similar, la tercera y cuarta sondas se hibridan al segundo ácido nucleico diana y se ligan mediante una enzima ligasa. A continuación, las sondas ligadas para los primer y segundo ácidos nucleicos diana se hibridan a través de la región complementaria que comprende la secuencia exógena y la PCR de extensión por solapamiento se utiliza para generar un complejo fusionado. Los complejos fusionados pueden ser secuenciados a granel.

#### **4) Enlace Molecular Utilizando Sondas-Candado**

Una sonda-candado es una molécula de ADN o de ARN de cadena sencilla, circularizada, con complementariedad con una secuencia diana de interés (Hardenbol *et al.*, 2003 *Nature Biotechnology* 21: 673-678; Patente de EE.UU. 6.858.412). Después de la hibridación a las moléculas diana, una polimerasa llena el hueco entre los dos extremos de la sonda, y una ligasa completa la cadena de polinucleótidos para formar una molécula de polinucleótido

circularizada. La molécula circularizada se puede amplificar, a continuación, con la amplificación por desplazamiento múltiple (MDA). La MDA es un método de amplificación isotérmico que funciona mediante reasociación de polinucleótidos de cadena sencilla al molde, seguido de la síntesis de ADN por una enzima de alta fidelidad tal como una  $\phi$ -29 polimerasa. La PCR inversa también se puede utilizar para amplificar sólo las moléculas circularizadas, porque cebadores de PCR que amplifican las moléculas circularizadas no amplificarán las sondas de cadena sencilla (Patente de EE.UU. N° 6.858.412).

Una ventaja notable de las sondas-candado frente a la PCR es la capacidad para la amplificación específica para los alelos. Mientras que la PCR amplifica ambos alelos para una variante particular, el proceso de ligamiento de sondas-candado es alelo-específico. Al igual que con la LCR, las sondas-candado se utilizan como un "interruptor" molecular. Si millones de células individuales son examinadas para rastrear una variante particular, sólo las células que incluyan dicha variante producirán amplicones principales. Por lo tanto, las sondas-candado se utilizan para realizar el análisis genético sólo en células que contienen una secuencia particular de interés. Además, en determinadas realizaciones, las sondas-candado están altamente multiplexadas, fijando decenas de miles de tipos de sonda como objetivo decenas de miles de loci genéticos en una microgotita de célula individual o reacción intracelular (Véase la Patente de EE.UU. N° 6.858.412).

Sondas-candado se hibridan típicamente a las dianas mediante ciclación durante al menos 20 veces entre 95°C durante 5 min y 55°C durante 20 min (Baner *et al.*, 2003 *Nucleic Acids Research* 31:e103). Los huecos de nucleótidos individuales se rellenan con Stoffel polimerasa y ligasa tal como Tth ligasa o Ampligase (Epicentre). Las sondas circularizadas son entonces amplificadas utilizando PCR con cebadores universales. Cuando se multiplexan para la PCR de extensión por solapamiento, se utilizan dos conjuntos de cebadores universales, uno para cada tipo de sonda-candado. Los cebadores universales contienen regiones de secuencias de solapamiento, lo que permite una PCR de extensión por solapamiento estándar después de la captura de la secuencia inicial por las sondas-candado. (Véanse las FIGs. 10-14). Las sondas también pueden ser diseñadas para que contengan las secuencias de cebadores adecuadas para la secuenciación a granel, de modo que el banco se secuencia directamente después de la amplificación por PCR.

La FIG. 10 ilustra un ejemplo de los componentes requeridos para un enlace de la secuencia de células individuales mediante sondas-candado en combinación con la reacción en cadena de la polimerasa de extensión por solapamiento, de acuerdo con una realización de la invención. La FIG. 11 muestra las regiones de complementariedad entre una primera sonda-candado (a) y el primer ácido nucleico diana (c) y entre una segunda sonda-candado (b) y un segundo ácido nucleico diana (d) en una célula individual, de acuerdo con una realización de la invención. Los componentes de la reacción pueden estar contenidos en un recipiente de reacción física o en una gotita de emulsión (k). La primera sonda-candado (a) incluye dos regiones separadas que son complementarias al primer ácido nucleico diana (c). La segunda sonda-candado (b) incluye dos regiones separadas que son complementarias a un segundo ácido nucleico diana (d). Se utilizan una polimerasa y una ligasa (m) para amplificar y ligar el hueco entre regiones complementarias de las sondas-candado (a) y (b).

La FIG. 12 ilustra los amplicones circularizados (g) y (h) resultantes y los cebadores que se utilizan para amplificar los amplicones circularizados, de acuerdo con una realización de la invención. Se utilizan un cebador directo (a) y un cebador inverso (i) para amplificar el amplicón circular (g). Los cebadores directo e inverso (j) y (f) se utilizan para amplificar el amplicón circular (h). El cebador (i) tiene una región (b) que es complementaria a una región del amplicón (g) y una región (c) que es complementaria a la región (d) del cebador (j). El cebador (j) tiene una región (e) que es complementaria al amplicón (h) y una región (d) que es complementaria a la región (c) del cebador (i).

La FIG. 13 es un ejemplo de los amplicones resultantes de la amplificación de las sondas circulares (g) y (h) de acuerdo con una realización de la invención. En esta figura, la región (a) es complementaria al amplicón (g) y la región (b) es complementaria a la región (c). La región (d) es complementaria al amplicón (h) y la región (c) es complementaria a la región (b).

La FIG. 14 es un ejemplo de amplificación por PCR de extensión por solapamiento de los amplicones utilizando una polimerasa (e), de acuerdo con una realización de la invención. El amplicón (f) resultante incluye secuencias (a), (d) y las secuencias solapantes (b) y (c). El amplicón (f) resultante se puede utilizar para la secuenciación a granel. Las etapas se pueden realizar en un recipiente de reacción o una gotita de emulsión (g).

##### **5) Enlace Molecular Utilizando la Circularización de Multisondas**

En algunas realizaciones, se puede utilizar la circularización de multisondas. En la circularización de multisondas, dos sondas-candado fijan como objetivo dos loci genéticos. Después de la hibridación a las moléculas diana, una polimerasa llena el hueco entre los extremos de las dos sondas, y una ligasa completa las cadenas de polinucleótidos para formar una molécula de polinucleótidos circularizada. (Véanse las FIGs. 1A-1C). La molécula circularizada se puede amplificar a continuación con la amplificación de desplazamiento múltiple (MDA). La PCR inversa también se puede utilizar para amplificar sólo las moléculas circularizadas, porque cebadores de PCR que amplifican las moléculas circularizadas no amplificarán las sondas de cadena sencilla (Véanse las FIGs. 2-3).

En una realización, las sondas se hibridan a dianas ciclando al menos 20 veces entre 95°C durante 5 min y 55°C durante 20 min (Baner *et al.*, 2003 *Nucleic Acids Research* 31: e103). Los huecos de nucleótidos individuales se rellenan con una polimerasa y ligasa Stoffel. Las sondas circularizadas se amplifican mediante PCR con cebadores universales. Cuando se multiplexan para la PCR de extensión por solapamiento, se utilizan los dos conjuntos de cebadores universales, uno para cada tipo de sonda-candado. Los cebadores universales contienen regiones de secuencias de solapamiento, lo que permite la PCR de extensión por solapamiento estándar después de la captura de la secuencia inicial de las sondas-candado (FIGs. 2-3). Las sondas también pueden ser diseñadas para que contengan las secuencias de cebadores apropiadas para la secuenciación a granel, por lo que el banco se secuencia directamente después de la amplificación por PCR.

La FIG. 1 muestra un ejemplo de enlace de secuencia en una célula individual por circularización intra-celular de multisondas de un complejo molecular, de acuerdo con una realización de la invención. Cada una de las sondas tiene una región de complementariedad con cada uno de los loci diana. El complejo incluye dos sondas de ácido nucleico (a y b) y dos ácidos nucleicos diana (c y d). La célula individual (e) puede estar contenida en un recipiente de reacción o una gotita de emulsión (j). La FIG. 1A ilustra que la sonda de ácido nucleico (a) tiene una primera región (f) que es complementaria a una región en el ácido nucleico diana (c), y una segunda región (g) que es complementaria a una región en el ácido nucleico diana (d). La sonda de ácido nucleico (b) tiene una primera región (h) que es complementaria a una región en el ácido nucleico diana (c) y una segunda región (i) que es complementaria a una región en el ácido nucleico diana (d). La FIG. 1B ilustra un ejemplo de enlace de secuencia en una célula individual (también en un recipiente de reacción o gotita de emulsión (j)) por circularización intra-celular de multisondas de un complejo, de acuerdo con una realización de la invención. Las dos sondas de ácidos nucleicos (a y b) se hibridan a las regiones complementarias de los dos ácidos nucleicos diana (c y d). La FIG. 1C ilustra un ejemplo de la circularización de un complejo de enlace sonda-diana que se produce por amplificación, de acuerdo con una realización de la invención. En un ejemplo, se utiliza una amplificación mediada por círculo rodante por  $\phi$ -29 polimerasa para circularizar las regiones extremas (f) de las dos sondas de ácido nucleico (a) y (b).

La FIG. 2 muestra un ejemplo de amplificación de un complejo de enlace circularizado sonda-diana (a) utilizando una polimerasa (b), de acuerdo con una realización de la invención. En algunas realizaciones, una  $\phi$ -29 polimerasa se utiliza en una amplificación mediada por círculo rodante, y se generan copias (b y c) del complejo sonda-diana circularizado. Además, la FIG. 3 ilustra un ejemplo de la amplificación de un complejo de enlace sonda-diana (a) circularizado utilizando una polimerasa (b) y cebadores (c y d), de acuerdo con una realización de la invención. Los cebadores (c y d) se utilizan para amplificar la región del complejo sonda-diana circularizado que es complementaria al ácido nucleico diana. Se generan múltiples copias (e) de un amplicón lineal de ácido polinucleico de doble cadena y se secuencian a granel.

### **6) Métodos que Utilizan Ácidos Nucleicos de código de barras**

La secuenciación a granel requiere la destrucción de células o microgotitas de emulsión, de manera que todos los análisis de ácido polinucleico se agrupan en una sola mezcla de reacción. Típicamente, no es posible un rastreo de una diana particular de la secuencia a partir de los datos de secuenciación a granel para una célula en particular. Sin embargo, muchas aplicaciones requerirán el rastreo posterior de las secuencias de sus células individuales originales. Por ejemplo, un investigador puede desear analizar una población de células para los patrones de expresión de células individuales para dos transcripciones de ARN. La amplificación por PCR con transcriptasa inversa de extensión por solapamiento de dos dianas de transcripción de ARN, seguida de secuenciación a granel no es adecuada para este tipo de análisis debido a que todas las transcripciones se mezclan juntas, y las transcripciones de células de alta expresión son indistinguibles de las transcripciones de las células de baja expresión. Para abordar este problema, se utilizan códigos de barras de ácido polinucleico. Cada una de las microgotitas de emulsión de una célula individual o un recipiente de reacción físico contiene un solo código de barras ácido polinucleico clonal único. Este código de barras se enlaza entonces a los ácidos polinucleicos diana (es decir, transcripciones de ARN), y se utiliza para rastrear los amplicones principales a una célula individual (Véanse las FIGs. 18-25). Con un seguimiento posterior de cada una de las secuencias a una célula individual original es posible tabular los datos genéticos para cada una de las células individuales, que entonces permite la cuantificación de células individuales (es decir, los niveles de expresión génica de células individuales).

En una realización, el oligonucleótido enlazador de código de barras está muy diluido, de manera que menos de 1% de las microgotitas de emulsión en picolitros porta más de un código de barras enlazador. Esto permite el enlace de una célula individual a un solo código de barras. El oligonucleótido enlazador de código de barras se amplifica por PCR utilizando universalmente cebadores dentro de cada una de las gotitas, de manera que cada una de las gotitas contendrá millones de copias de sólo una secuencia de código de barras enlazador, y que el código de barras será único para esa gotita (FIGs. 18-21). La dilución sigue la estadística de Poisson, de modo que para  $P(k = 1) \approx 0,99$ , los códigos de barras de enlazador necesitan ser diluidos a  $\lambda \approx 0,01$ . El código de barras está entonces físicamente enlazado a la molécula diana por PCR de extensión por solapamiento. Los códigos de barras pueden ser producidos por un cierto número de métodos. En una realización, un banco de decámeros aleatorios se subclona en un vector plasmídico (p. ej., Life Technologies). Esto produce un banco de plásmidos mixtos con > 1 millón de códigos de barras de decámero únicos. A continuación, los plásmidos se transforman en bacterias y se recogen 3.840 clones.

Los clones son secuenciados por secuenciación capilar (Sequetech) y se archivan en materiales de glicerol en placas de 384 pocillos. Seguidamente, los clones se digieren en sitios de restricción en cualquier lado de las inserciones de decámero aleatorias para producir un fragmento de ~100 pb. Estos fragmentos son entonces biotinilados utilizando el fragmento de Klenow con procedimientos estándares. Un lavado entre las etapas moleculares se lleva a cabo con ayuda de tecnología de perlas Ampure (PerkinElmer). Los fragmentos biotinilados se fijan entonces a perlas de estreptavidina de 17 µm de diámetro (Life Technologies) en cada uno de los pocillos, produciendo 3.840 poblaciones clonales de perlas de código de barras. La Amplificación de ácidos nucleicos utilizando emulsiones de perlas se describe en la Patente de EE.UU. N° 7.842.457.

En una realización, el método proporciona perlas de código de barras fijadas a secuencias de ácidos nucleicos. Un banco de 15-meros al azar se subclona en un vector de plásmido (Life Technologies). Esto produce un banco de plásmidos mezclados con > 1 billón de códigos de barras de 15-meros únicos. Los fragmentos biotinilados se fijan a continuación a perlas de estreptavidina de 17 µm de diámetro (Life Technologies). La mezcla plásmidos-códigos de barras se diluye en una mezcla de PCR de manera que el 99% de las gotitas que contienen un plásmido contendrán sólo un único plásmido clonal. La mezcla de PCR contiene nucleótidos biotinilados, de manera que los códigos de barras amplificados se biotinilan. Entonces, perlas de estreptavidina se hacen fluir en esta mezcla de PCR para encapsular las perlas individuales en microgotitas. Típicamente, se encapsulan al menos 10 millones de perlas, y luego las mezclas de perlas/plásmido se termociclan para amplificar y biotinilar los códigos de barras. Las perlas de código de barras se recuperan entonces y se pueden utilizar en el método de código de barras de gotitas.

En otra realización, un dispositivo de microfluidos inyecta perlas recubiertas con oligonucleótidos de código de barras de enlazador clonales en las microgotitas de la emulsión de células individuales. Un dispositivo de este tipo permite la visualización de perlas individuales y células individuales en cada una de las gotas, eliminando el requisito de oligonucleótidos de códigos de barras de enlazador altamente diluidos. En esta realización, también se utiliza la PCR para amplificar el oligonucleótido de código de barras de enlazador, de forma que cada una de las gotitas contiene millones de copias de la misma secuencia de código de barras, pero cada uno de los códigos de barras sería único para una microgotita individual. El código de barras es entonces enlazado a la secuencia de ácido nucleico diana utilizando la PCR de extensión por solapamiento. Durante la amplificación por PCR de extensión por solapamiento, las regiones de secuencias complementarias de las primera y segunda secuencias de ácidos nucleicos amplificados actúan como cebadores para la extensión en ambas cadenas en cada dirección mediante moléculas de ADN polimerasa. En posteriores ciclos de la PCR, los cebadores externos ceban la secuencia condensada completa de tal modo que se duplica por la ADN polimerasa. Este método produce una pluralidad de complejos de fusión.

En otra realización, el método incluye las etapas de proporcionar una agrupación de secuencias de código de barras única, en donde cada una de las secuencias de código de barras está enlazada a un gen de resistencia a la selección, proporcionando una población de células individuales, transfectando la población de células individuales con la agrupación de secuencias de código de barras única, seleccionando células que comprenden una secuencia de código de barras única y el gen de resistencia a la selección, y aislando cada una de las células seleccionadas en recipientes de reacción o microgotitas de la emulsión. En algunas realizaciones, el gen de resistencia a la selección codifica la resistencia a gentamicina, neomicina, higromicina o puromicina. El gen de resistencia a la selección permite seleccionar células que han incorporado la secuencia de código de barras en la célula. Las células que carecen de plásmido también carecen del gen de resistencia a la selección y, por lo tanto, se exterminan en presencia de una sustancia química de selección de mamíferos tales como gentamicina, neomicina, higromicina o puromicina.

La FIG. 15 ilustra un ejemplo de desconvolución de un banco de plásmidos mediante reacción en cadena de la polimerasa con cola con código de barras (código de barras en el extremo 5'), que es seguido por secuenciación a granel e informática, de acuerdo con una realización de la invención. La secuencia de código de barras se puede rastrear de nuevo a una posición de pocillo y placa, la secuencia de código de barras se puede rastrear, a continuación, a una secuencia de ácido nucleico y la secuencia de ácido nucleico se rastrea de nuevo a un pocillo. Cada uno de los cebadores en (a) y (b) tiene una etiqueta con código de barras en el extremo 5'. Los ácidos nucleicos diana en (c) y (d) se amplifican utilizando los cebadores en (a) y (b). Las etapas se pueden realizar en recipientes cerrados o en gotitas de emulsión tal como se muestra en (c) y (d). La FIG. 16 también muestra un ejemplo de la amplificación (e, f) de dos ácidos nucleicos diana (A y B), utilizando cebadores que incluyen secuencias de códigos de barras, de acuerdo con una realización de la invención. Los amplicones resultantes que incluyen las secuencias de códigos de barras se muestran en (g) y (h). Además, la FIG. 17 ilustra un ejemplo simplificado de rastreo renovado de una secuencia de código de barras en un amplicón a una diana celular (A o B), y el rastreo renovado de la diana celular a un lugar físico (c, d) (p. ej., un pocillo), de acuerdo con una realización de la invención.

Además, la FIG. 18 ilustra los componentes para el enlace molecular entre dos transcripciones (g y h) y una secuencia de código de barras molecular (k), de acuerdo con una realización de la invención. Las dianas (g y h) pueden ser transcripciones de ARN, y la secuencia de código de barras molecular (k) está flanqueada por sitios de cebado universales. Sólo una copia del oligonucleótido de código de barras molecular está contenida en la gotita de

emulsión o recipiente de reacción (j), y cebadores de PCR universales amplifican el oligonucleótido para producir una pluralidad de ácidos polinucleicos de código de barras clonales. Un cebador directo (a) y cebador inverso (m) se utilizan para amplificar el ácido nucleico diana (g). Un cebador directo (n) y cebador inverso (f) se utilizan para amplificar el ácido nucleico diana (h). El cebador inverso (m) incluye una región (b) que es complementaria al ácido nucleico diana (g) y una región (c) que es complementaria a la región (d) en el cebador (n). El cebador (n) incluye una región (e) de complementariedad al ácido nucleico diana (h) y una región (d) de complementariedad a la región (c) del cebador (m). En algunas realizaciones, se pueden enlazar más de dos dianas, y las dianas también pueden ser ADN.

Además, la FIG. 19 muestra un ejemplo de la amplificación de los ácidos nucleicos diana (g y h) usando cebadores tal como se muestra, de acuerdo con una realización de la invención. El cebador directo (a) es complementario al ácido nucleico diana (g), y el cebador inverso (b) para el ácido nucleico diana (g) incluye una región (c) que es complementaria a la secuencia de código de barras (k). El cebador directo (e) y el cebador inverso (f) se utilizan para amplificar el ácido nucleico diana (h). El cebador directo (e) incluye una región (d) que es complementaria a la secuencia de código de barras (k).

En la FIG. 20 se muestran amplicones que resultan después de la amplificación de dos ácidos nucleicos diana y una secuencia de código de barras (k), de acuerdo con una realización de la invención. La FIG. 21 ilustra un amplicón fusionado que incluye secuencias de dos ácidos nucleicos diana (g y h) y una secuencia de código de barras (k) dentro de una gotita de emulsión o recipiente de reacción (j), de acuerdo con una realización de la invención. El amplicón ("principal") fusionado puede ser aislado por emulsión inversa y ser secuenciado a granel. En la FIG. 22, las dianas (g y h) pueden ser transcripciones de ARN, y la secuencia de código de barras molecular (k) está flanqueada por sitios de cebado universales. Sólo una copia de la secuencia de código de barras molecular (k) está contenida en la gotita de emulsión de una célula individual o recipiente de reacción (j), y cebadores de PCR universales amplifican el oligonucleótido para producir una pluralidad de ácidos polinucleicos de código de barras clonales. El cebador directo (a) y el cebador inverso (b) se utilizan para amplificar el ácido nucleico diana (g). El cebador directo (n) y el cebador inverso (f) se utilizan para amplificar el ácido nucleico diana (h). El cebador inverso (m) incluye una región (b) que es complementaria al ácido nucleico diana (g) y una región (c) que es complementaria a la región (d) en el cebador (n). El cebador (n) incluye una región (e) de complementariedad al ácido nucleico diana (h) y una región (d) de complementariedad a la región (c) de cebador (m). En algunas realizaciones, se pueden enlazar más de dos dianas, y las dianas también pueden ser ADN.

La FIG. 23 ilustra los cebadores directo e inverso que se utilizan en un enlace molecular entre dos transcripciones (g y h) y una secuencia de código de barras molecular (k) unidas a una perla (m), de acuerdo con una realización de la invención. El cebador directo (a) y el cebador inverso (b) se utilizan para amplificar el ácido nucleico diana (g). El cebador directo (n) y el cebador inverso (f) se utilizan para amplificar el ácido nucleico diana (h). El cebador inverso (m) incluye una región (b) que es complementaria al ácido nucleico diana (g) y una región (c) que es complementaria a la región (d) en el cebador (n). El cebador (n) incluye una región (e) de complementariedad al ácido nucleico diana (h) y una región (d) de la complementariedad a la región (c) del cebador (m). Los dos ácidos nucleicos diana son complementarios a una secuencia de ADN (1). La FIG. 24 es un ejemplo de amplicones que resultan después de la amplificación de dos ácidos nucleicos diana y una secuencia de código de barras (k) unida a una perla (m), de acuerdo con una realización de la invención. La FIG. 25 ilustra un amplicón fusionado que incluye secuencias de dos ácidos nucleicos diana (g y h) y una secuencia de código de barras (k), dentro de una gotita de emulsión o recipiente de reacción (j), de acuerdo con una realización de la invención. El amplicón ("principal") fusionado puede ser aislado por emulsión inversa y secuenciado a granel. Las FIGs. 24-25 ilustran un ejemplo de amplicones que resultan después de la amplificación de dos ácidos nucleicos diana y una secuencia de código de barras (k) unida a una perla (m), de acuerdo con una realización de la invención.

#### **7) Métodos que Utilizan la Amplificación de Combinación**

La fijación como objetivo y la amplificación de loci genéticos en células se puede realizar utilizando PCR, LCR, sondas-candado, RT-PCR o circularización multi-sonda. Se puede utilizar cualquier combinación de estos métodos para fijar como objetivo y amplificar diferentes loci. Por ejemplo, un enfoque de amplificación de la combinación se utiliza para amplificar un locus de ADN genómico y una transcripción de ARN. En una realización, una enzima transcriptasa inversa termoestable, tal como ThermoScript RT (Lucigen) o GeneAmp ThermoStable rTth (Life Technologies) se combina con una ADN polimerasa termoestable tal como el fragmento Stoffel o ADN Taq polimerasa. La termociclación puede inducir la síntesis de la primera cadena de ADNc de la diana de transcripción de ARN. Una vez que se ha sintetizado el ADNc a partir de la transcripción de ARN, se realiza una PCR de extensión por solapamiento utilizando el ADNc y las secuencias diana de ADN genómico.

#### **8) Método de secuenciación a Granel**

Existe un cierto número de nuevas metodologías comerciales para la secuenciación de ácido polinucleico. A estas tecnologías se las alude a menudo como "la secuenciación de generación siguiente", "secuenciación masiva en paralelo" o "secuenciación a granel". Estos términos se utilizan indistintamente para describir cualquier método de

secuenciación que es capaz de adquirir más de un millón de etiquetas de secuencias de ácido polinucleico en una sola operación. Típicamente, estos métodos funcionan al hacer mediciones altamente paralelizadas, es decir, el rastreo paralelizado de millones de clones de ADN sobre portaobjetos de vidrio. Los métodos para enlazar múltiples dianas de ácido polinucleico en células individuales se podrían utilizar en combinación con cualquier método de secuenciación a granel comercializado. Estos métodos incluyen la química de terminación reversible (Illumina), la pirosecuenciación utilizando gotitas de emulsión polony (Roche), la secuenciación de una sola molécula (Pacific Biosciences) y otros (IonTorrent, Halcyon, etc.).

Después del enlace molecular se llevan a cabo protocolos, y antes de la secuenciación a granel, es útil amplificar específicamente y purificar amplicones principales para reducir la secuenciación total requerida para obtener datos útiles. De lo contrario, muchos amplicones menores y otros tipos de secuencias de fondo no deseadas se secuenciarán de forma innecesaria. Esto se logra mediante la PCR utilizando solamente los cebadores externos y el analito de ácido nucleico obtenido a partir de las células lisadas, seguido de selección por tamaños utilizando un método tal como electroforesis en gel de agarosa. Otros métodos, tales como las columnas de exclusión por tamaños, electroforesis de microfluidos o filtros de microporos, se podrían utilizar para seleccionar las moléculas de tamaño adecuado.

En una realización, el método proporciona la etapa de realizar una reacción de secuenciación a granel para generar información de la secuencia de al menos 100.000 complejos fusionados de al menos 10.000 células dentro de una población de células. En otra realización, la reacción de secuenciación a granel genera información de la secuencia para al menos 75.000, 50.000 ó 25.000, ó 10.000 complejos fusionados de al menos 10.000 células dentro de una población de células.

Los complejos fusionados se pueden utilizar entonces para cuantificar el fenómeno biológico o clínico particular de interés. En el caso del análisis de células T o B funcionales, clonotipos particulares que expresan moléculas funcionales pueden analizarse determinando primero la secuencia de péptidos CDR3 del complejo fusionado, y luego tabulando los casos de ese péptido CDR3 enlazado a una molécula efectora particular. De esta manera la secuenciación a granel cuantifica la expansión clonal y la función biológica de cada uno de los clonotipos individuales. Cuando los cebadores fijan como objetivo múltiples moléculas efectoras y todas las posibles regiones variables se multiplexan en un único ensayo, se pueden separar clonotipos en compartimientos funcionales. En el caso de enlace entre códigos de barras y dianas de transcripción, se puede estratificar los datos de secuenciación a granel mediante código de barras y luego tabular los casos de un código de barras particular enlazado a una diana de transcripción. Cuando cebadores que fijan como objetivo múltiples transcripciones se multiplexan en un único ensayo, se pueden utilizar los códigos de barras para inferir patrones de expresión multigénicos para células individuales rastreadas de nuevo en cuanto a gotitas individuales. En el caso del enlace entre una secuencia mutante o variable y otras secuencias mutantes o variables, se pueden analizar los datos de secuenciación a granel para determinar la secuencia en cada uno de los locus en cada una de las moléculas en el banco de secuenciación a granel, y luego tabular los casos de cada tipo de secuencia. Por ejemplo, si se requiere una mutación en cada una de las dos dianas enlazadas para producir un fenotipo de la enfermedad, se puede utilizar la cuantificación del número de dianas enlazadas con dos mutaciones para detectar la enfermedad en un individuo.

### **C. Enlace Intracelular en Células Fijadas Seguido de Secuenciación Masiva Paralela**

Los métodos moleculares descritos en la sección B anterior se pueden realizar de forma intracelular en miles de millones de células fijas individuales (Embleton *et al.*, 1992, *Nucleic Acids Research* 20:3831-37; Hviid, 2002 *Clinical Chemistry* 48: 2115-2123; Patente de EE.UU N° 5.830.663). Las membranas celulares de las células sirven como compartimientos de reacción, permitiendo el enlace entre dos o más loci genéticos en miles a millones de células fijadas individuales analizadas en paralelo. El uso de células fijadas como compartimientos de reacción es más rentable que un chip microfluídico para hacer microgotitas de emulsión. Además, la heterogeneidad en el tamaño celular o la morfología en una población particular de células, es menos probable que interrumpa el método de células fijadas que el método de microgotitas de emulsión. Sin embargo, en algunos casos, la fuga de ácidos nucleicos a partir de las células puede provocar un ruido de fondo en el análisis genético molecular, por lo que se debe tener cuidado de lavar las células entre etapas moleculares y llevar a cabo un análisis riguroso de la calidad de los analitos. Por lo tanto, en una realización, células fijadas y permeabilizadas se encapsulan en microgotitas, y la amplificación se produce utilizando células fijadas y permeabilizadas en microgotitas en lugar de células lisadas en el interior de microgotitas.

#### **1) Enlace Molecular en Métodos de Células Fijadas**

El trabajo de los autores de la invención utilizando la amplificación del genoma entero (WGA) de células individuales y la PCR a partir de células individuales fijas ha demostrado que la fijación de células en glutaraldehído inhibe WGA, pero no la PCR. En cualquier realización de protocolos de enlace intracelulares, se debe tener cuidado de asegurar que la fijación y/o permeabilización no inhibe la amplificación molecular.

Para la fijación, se pueden utilizar reactivos tales como glutaraldehído, paraformaldehído, IntraStain (Dako), o reactivos similares. Para la permeabilización, se pueden utilizar reactivos tales como Triton X-100, Tween-20, IntraStain (Dako), o reactivos similares (Lippincott-Schwartz 2003 *Short Protocols in Cell Biology*; Celis 2005 *Cell Biology: A Laboratory Handbook*). Después de la fijación y/o permeabilización, las células se lavan múltiples veces en un tampón tal como solución salina tamponada con fosfato (PBS). Una vez que las células se fijan y/o se permeabilizan, tampones de reacción que contienen los cebadores/las sondas y enzimas se suministran al compartimiento intracelular sin maquinaria o métodos especiales.

Por ejemplo, cuando se utiliza la RT-PCR para amplificar los loci objetivos en células individuales, las células fijadas y permeabilizadas se sumergen en tampón de reacción y la primera cadena de ADNc se sintetiza intracelularmente a 55-70°C durante cuatro horas. Sin el lavado o intercambio de tampón, se podrían entonces utilizar las condiciones de termociclación de PCR de extensión por solapamiento estándar para amplificar y enlazar las dianas. Después de este procedimiento de amplificación, la mezcla se lava varias veces con PBS, y el sobrenadante se retiene para el análisis de control de calidad. Las membranas de las células resuspendidas son entonces interrumpidas utilizando tampón de lisis alcalina o disoluciones de proteinasa K (Johnson *et al.*, 2010 *Human Reproduction* 25:1066-75).

Después de la lisis de las células y antes de la secuenciación a granel, es útil amplificar específicamente los complejos enlazados para reducir la secuenciación global requerida para obtener datos útiles. Esto se logra mediante PCR utilizando solamente los cebadores externos y el analito de ácido nucleico obtenido a partir de las células lisadas, seguido de selección por tamaños utilizando un método tal como electroforesis en gel de agarosa.

## **II. Métodos de Desconvolución del Banco de Clones Agrupado**

Bancos altamente multiplexados de ácidos nucleicos se producen a menudo utilizando métodos paralelizados que no logran producir moléculas individuales a una molaridad optimizada para las aplicaciones de interés. El método en esta memoria proporciona la síntesis de paralelizado, la desconvolución y la re-multiplexación de bancos de ácidos polinucleicos. El método conserva las ventajas tanto de la síntesis paralelizada como de la optimización del clon individual. Estos bancos de ácidos polinucleicos se utilizan para una diversidad de aplicaciones, incluyendo pero no limitadas a, amplificación multiplexada de secuencias de ácidos nucleicos diana para la secuenciación y el análisis (FIGs. 15-17).

### **A. Método de Síntesis de Sonda-Candado**

#### **1) Método de Desconvolución de Agrupaciones de Pre-sonda**

En una realización, se genera una agrupación de miles de sondas-candado que fijan como objetivo polimorfismos de nucleótidos individuales, o SNP. Precursores de sondas de oligonucleótidos de ADN se sintetizan en agrupaciones (Atactic o NimbleGen). Cebadores universales se utilizan entonces para amplificar por PCR de ADN de doble cadena de la agrupación de oligonucleótidos (Porreca *et al.*, 2007 *Nature Methods* 4:931-36). A continuación, los extremos del banco de amplificación de doble cadena de PCR se digieren utilizando una enzima de restricción. Por ejemplo, se utiliza EcoP15I, que escinde 25 pares de bases del sitio de reconocimiento y separa los sitios de unión de PCR universales. EcoP15I es un ejemplo de una enzima que es adecuada para la subclonación, y productos no escindidos no afectan a etapas moleculares aguas abajo. El banco digerido se subclona en vectores de plásmidos de diseño personalizado que confieren resistencia a la ampicilina. Los plásmidos se transforman entonces en cultivos bacterianos bajo selección con un antibiótico.

La FIG. 4 ilustra un ejemplo de amplificación de un complejo de enlace circularizado sonda-diana (a) en una célula individual (b), de acuerdo con una realización de la invención. La amplificación se produce por transformación en bacterias y posterior selección con antibióticos. El amplicón (a) contiene un gen resistente a antibióticos y células (c) que se transforman con el amplicón se seleccionan en presencia de antibióticos. No se seleccionan las células sin el complejo sonda-diana (d) circularizado.

#### **2) Síntesis de sondas de cadena sencilla *en masa***

En algunas realizaciones, un material bacteriano que contiene un banco mixto de miles de clones, fijando cada uno como objetivo un SNP particular, se utiliza para la síntesis de sondas de cadena sencilla *en masa*. Por ejemplo, los cultivos bacterianos se extienden sobre placas de agar LB bajo selección con ampicilina, y luego se recogen colonias individuales. A continuación, la PCR con cebadores con código de barras se utiliza para amplificar la secuencia de la sonda y regiones de cebado flanqueantes universales. El resultado es un amplicón que contiene tanto la secuencia de la sonda como un código de barras que se puede rastrear de nuevo a un solo pocillo. En una realización, un código de barras molecular único indicará una posición particular del pocillo en una placa de 384 pocillos particular. Por ejemplo, el sistema podría tener 3.840 códigos de barras únicos que indican las posiciones de los pocillos y el número de placas para 3.840 PCRs en una de las diez placas de 384 pocillos. Para desconvular un banco de 10.000 complejos de clones, se realizan cuatro rondas de desconvolución utilizando el conjunto de 3.840 PCRs con código de barras, y sobremuestreando y detectando un total de 15.360 clones. Para cada una de las

rondas de desconvolución, los productos de PCR pueden entonces agruparse y secuenciarse utilizando cualquier método de secuenciación principal.

5 Con las secuencias de la sonda adaptadas a un código de barras, se puede entonces utilizar un algoritmo de desconvolución para desconvolutar el banco. Debido a que el código de barras se corresponde con la secuencia de inserción, se crea una tabla que coincide con la secuencia de código de barras para el pocillo y la placa originales y, en consecuencia, esto coincide con la secuencia de inserción para un pocillo. Los clones bacterianos pueden ser entonces almacenados como materiales de glicerol, y las secuencias de estos materiales pueden ser catalogadas en una base de datos y almacenadas a -80°C.

10 Para sintetizar sondas-candado de cadena sencilla de los materiales de glicerol de molde, se utiliza una derivación de la técnica de SMART (Krishnakumar *et al.*, 2008 *PNAS* 105:9296-9301). En una concentración alta, este método implica la digestión (i) de un ADN de doble cadena con una endonucleasa de restricción; (ii) la desfosforilación del "extremo pegajoso"; (iii) la digestión del segundo extremo del ADN de doble cadena con una segunda endonucleasa de restricción; y (iv) la digestión de la cadena desfosforilada de ADN utilizando una exonucleasa  $\lambda$ . En primer lugar, los clones deseados se recogen, y después se cultivan en placas de 384 pocillos. Después de la incubación durante la noche, se evalúa la densidad óptica de cada uno de los cultivos, y luego se igualan los materiales. Se agrupan 5  $\mu$ L de los cultivos bacterianos normalizados, y la agrupación de plásmido se purifica utilizando métodos estándares (Qiagen). A continuación, se utiliza un conjunto de cebadores de PCR universales para generar una agrupación de amplicones de PCR de doble cadena. La mezcla de la PCR resultante se somete entonces a digestión con una enzima de restricción, tal como HaeIII (NEB), seguido de desfosforilación con fosfatasa alcalina de camarón (SAP).  
15 Después de la desfosforilación, el analito se digiere con una enzima de restricción tal como BstUI (NEB). Este producto puede entonces ser digerido con exonucleasa  $\lambda$  (NEB), produciendo moléculas de ADN de cadena sencilla. Finalmente, el ADN de cadena sencilla (ADNs) se purifica a partir de cualquier ADN de doble cadena no digerido utilizando un kit comercial (Zymo Research). De este modo, se sintetizan en paralelo cientos de miles de sondas.

### **B. Desconvolución del Clon Celular**

25 Los métodos en la Sección II. A. también se pueden utilizar para desconvolutar bancos mixtos de células u organismos con diferentes características genéticas subyacentes. El objetivo es separar el banco mixto de clones en compartimientos de reacción, realizar PCR con código de barras, seguido por secuenciación principal en los clones, y a continuación mapeando los datos de secuencia de nuevo a los clones en los compartimientos de reacción. En un ejemplo, una población de células de mamífero se somete a mutagénesis y luego poblaciones clonales de células mutagenizadas se aíslan de la población mixta. En esta realización, las células mutagenizadas individuales están ordenadas en compartimientos de reacción, y después se realiza una PCR con código de barras fijada como objetivo o sondas-candado en loci genéticos de interés. Los datos de secuenciación a granel se utilizan para rastrear de nuevo los clones originales, y luego los materiales de clones físicos se utilizan para una investigación o uso adicional.

### **EJEMPLOS**

35 A continuación se presentan ejemplos de realizaciones específicas para llevar a cabo la presente invención. Los ejemplos se ofrecen para fines ilustrativos solamente, y no pretenden limitar el alcance de la presente invención de modo alguno. Se han hecho esfuerzos para asegurar la precisión con respecto a los números utilizados (p. ej., cantidades, temperaturas, etc.), pero, por supuesto, se deberían permitir algunos errores y desviaciones experimentales.

40 La práctica de la presente invención empleará, a menos que se indique lo contrario, métodos convencionales de la química de proteínas, bioquímica, técnicas de ADN recombinante y farmacología, dentro de la experiencia de la técnica. Tales técnicas se explican completamente en la bibliografía. Véase, p. ej., T.E. Creighton, *Proteins: Structures and Molecular Properties* (W.H. Freeman and Company, 1993); A.L. Lehninger, *Biochemistry* (Worth Publishers, Inc., edición actual); Sambrook, *et al.*, *Molecular Cloning: A Laboratory Manual* (2ª edición, 1989); *Methods in Enzymology* (S. Colowick y N. Kaplan eds., Academic Press, Inc.); *Remington's Pharmaceutical Sciences*, 18ª Edición (Easton, Pensilvania: Mack Publishing Company, 1990); Carey y Sundberg, *Advanced Organic Chemistry* 3ª Ed. (Plenum Press) Vols. A y B (1992).

### **Ejemplo 1: Métodos de Análisis de células T**

50 El sistema inmune responde a la enfermedad mediante la inducción de respuestas celulares. Casi toda la inmunología está involucrada en la detección de la expansión o contracción del clonotipo en respuesta a un antígeno y/o el análisis funcional de los clonotipos expandidos o contraídos. Descritos en este ejemplo son métodos que aprovechan la información contenida en la respuesta inmune para diagnosticar y tratar la enfermedad. Células activas y/o de la memoria son particularmente informativas, ya que estas células indican una respuesta inmune funcional a una enfermedad y, por lo tanto, tienen un alto contenido de información. Regiones de ADN variables y las transcripciones de ARN se analizaron en células individuales a partir de poblaciones de células inmunes y/o de la

memoria activadas, y luego se correlacionaron con la enfermedad. Estos perfiles se utilizaron para desarrollar diagnósticos no invasivos, diagnósticos de alto valor que informan de los regímenes de tratamiento y los nuevos agentes terapéuticos.

5 Células T incluyen receptores de células T (TCR) que reconocen antígenos y controlan la respuesta inmune. El receptor de células T se compone de dos subunidades:  $\alpha$  y  $\beta$  o  $\gamma$  y  $\delta$ . Los métodos actuales para examinar las células T por sus receptores de células T en su inmensa mayoría de secuencias de subunidades del receptor de células T a partir de poblaciones a granel que van desde unas pocas a millones de células. Esto resulta en un catálogo de secuencias de la subunidad ( $\alpha$  o  $\delta$ ) que son disociadas de la otra secuencia de la subunidad correspondiente que se encuentra en las células individuales ( $\beta$  o  $\gamma$ ). Esto da una información del nivel de la población sobre la diversidad de receptores de células T, pero no da una descripción de los receptores de células T individuales en células individuales por las dos subunidades ( $\alpha$  y  $\beta$  o  $\gamma$  y  $\delta$ ). Mediante el enlace de las secuencias en una célula individual utilizando los métodos en las Secciones I. A-C, los TCRs de células individuales en poblaciones mixtas se analizan con una resolución más fina, y esto permite un mapeado sin precedentes de la diversidad de células T humanas.

15 Las secuencias de subunidades TCR y moléculas de funcionalidad inmunes fueron enlazadas utilizando los métodos descritos en las secciones I. A-C. Este enfoque, denominado "secuenciación de células T funcional", se centró específicamente en las células T que puedan tener una función clínica o biológicamente relevante. Por ejemplo, la función inmune de una célula T se indica por la expresión de TCR clonales y moléculas de señalización tales como la interleucina-4 (IL-4). Células T naifs expresan TCR clonales, pero no expresan las moléculas de señalización tales como IL-4, y tienen diferentes funciones inmunes. El TCR se enlazó a la molécula de señalización que, a su vez, se enlazó al TCR de la función clínica. Los cebadores que amplifican el repertorio completo de TCR $\beta$  estaban enlazados a una única molécula de efector inmune tal como IL-4. Los cebadores que amplifican el repertorio completo de TCR $\beta$  estaban enlazados a docenas de moléculas de efector inmune, resultando en un fenotipo completo de células T para cada uno de los clonotipos de células T en el ensayo.

25 Ejemplos de moléculas que están asociadas con la función inmunológica y que están enlazadas a una secuencia de TCR incluyen, pero no se limitan a: interleucina-2 (IL-2), interleucina-4 (IL-4), interferón gamma (IFN $\gamma$ ), interleucina-10 (IL-10), interleucina-1 (IL-1), interleucina-13 (IL-13), interleucina-17 (IL-17), interleucina-18 (IL-18), factor de necrosis tumoral alfa (TNF $\alpha$ ), factor de necrosis tumoral beta (TNF $\beta$ ), factor de transcripción de la caja T 21 (TBX21), caja forkhead P3 (FOXP3), cúmulo de diferenciación 4 (CD4), cúmulo de diferenciación 8 (CD8), cúmulo de diferenciación 1d (CD1d), cúmulo de diferenciación 161 (CD161), cúmulo de diferenciación 3 (CD3) y factor de transcripción de la caja T TBX21 (T-BET).

35 La cadena  $\beta$  de TCR fue ligada a una molécula asociada con la función inmune. En otro método a modo de ejemplo, los TCR  $\alpha$  y  $\beta$ , o los TCR  $\gamma$  y  $\delta$ , o cualquiera de las subunidades individuales, estaban enlazadas a las moléculas de funcionalidad inmunes. Se utilizaron cebadores optimizados publicados para la amplificación de TCR genómico recombinado (Robins *et al.*, 2009 *Blood* 114:4099-107). Gran parte de la variabilidad peptídica del TCR estaba codificada en CDR3 $\beta$ , que se formó por recombinación entre los segmentos variables (V), de diversidad (D) y de unión (J) no contiguos en los loci de la cadena b (Wang *et al.*, 2010 *PNAS* 107:1518-1523). También se pueden utilizar cebadores de PCR que fijan como objetivo el locus CDR3 $\beta$ , previamente publicados (Robins *et al.*, 2009 *Blood*. 114:4099-107; Robins *et al.*, 2010 *Science Translational Med* 2:47ra64). Este conjunto de cuarenta y cinco cebadores directos y trece cebadores inversos amplifican la región CDR3 $\beta$  genómica recombinada de ~ 200 pares de bases para la amplificación multiplex del complemento completo de CDR3 $\beta$  de una muestra de células mononucleares de sangre periférica humana. La región CDR3 $\beta$  comienza con la segunda cisteína conservada en la región 3' del segmento V $\beta$  y termina con la fenilalanina conservada codificada por la región 5' del segmento J $\beta$  (Monod *et al.*, 2004 *Bioinformatics* 20:i379-i385). Por lo tanto, las secuencias amplificadas se tradujeron informáticamente para localizar la cisteína conservada, obtener la secuencia de péptidos intermedia y tabular los recuentos de cada uno de los clones único en la muestra.

45 Ejemplos de cebadores que pueden ser utilizados para la amplificación multiplex de secuencias de TCR y el enlace a diversas moléculas de efector inmunes se muestran en la Tabla 2. Estos cebadores se han utilizado, por ejemplo, con los métodos de la Sección I. A-C, para amplificar y enlazar secuencias TCR a diversas moléculas de efector inmunes.

### **Ejemplo 2: Protocolo de Alto Rendimiento para la Construcción del Banco de Repertorio de TCR $\beta$**

55 En una realización, se implementó un protocolo de alto rendimiento para la construcción de bancos de repertorio de TCR $\beta$  humano o de ratón. Los bancos fueron secuenciados directamente en la plataforma de secuenciación de generación siguiente GAIx (Illumina). Para las muestras humanas, se realizó una PCR multiplex utilizando un conjunto de 20 cebadores para amplificar los 50 segmentos V y 10 cebadores para amplificar los 13 segmentos J. Los bancos de cebadores generaron bancos que eran el complemento inverso de la secuencia de TCR $\beta$  nativa. Esto permitió la secuenciación desde el lado J de las construcciones sin manipulación adicional. Los cebadores también tenían colas con la misma secuencia que una parte del adaptador de banco TruSeq de Illumina. Los 30 cebadores

se agruparon en una sola PCR de 400  $\mu$ l, que contenía el ADN genómico de al menos  $5 \times 10^5$  células. Las reacciones fueron luego termocicladas durante no más de 25 ciclos, dependiendo del número de células de entrada. Después de la termociclación se utilizó una columna de PCR (Qiagen) para retirar los cebadores. A continuación, se realizó una segunda ronda de PCR, utilizando una parte alícuota del analito purificado de la primera ronda y un conjunto de cebadores universales. Los cebadores universales para la segunda ronda de PCR se reasociaron a las colas de los primeros cebadores, proporcionando productos finales de PCR que tenían la secuencia de adaptador de la secuenciación completa Illumina condensada a un banco de secuencias de TCR $\beta$ . Los cebadores universales también tenían etiquetas de código de barras, lo que permitió la multiplexación de docenas de muestras en una sola pista de secuenciación de siguiente generación. Finalmente, los bancos fueron purificados con gel de selección por tamaños y fueron cuantificados con un kit de PCR cuantitativa (Kapa Biosystems) antes de la secuenciación. Utilizando este protocolo se construyeron y secuenciaron más de 300 bancos de TCR $\beta$ .

La FIG. 31 muestra un flujo de trabajo simplificado para la generación de alto rendimiento de bancos de repertorio de TCR $\beta$ . La primera ronda utilizó un conjunto de 30 cebadores para amplificar el repertorio completo de TCR $\beta$  y fija regiones de cebado universales. La segunda ronda amplificó el repertorio con cebadores universales y añadió secuencias para la secuenciación de siguiente generación.

### **Ejemplo 3: Optimización del Protocolo Utilizando una Agrupación de 48 complejos de Clones del plásmido TCR $\beta$**

El verdadero contenido de cualquier repertorio de TCR $\beta$  particular no se conoce, por lo que un repertorio de TCR $\beta$  endógeno no puede servir como un patrón de oro para la optimización del protocolo. Una agrupación de 48 complejos de clones plasmídicos TCR $\beta$  de ratón fue diseñada para actuar como molde para la optimización del protocolo. En primer lugar, se realizó la amplificación multiplexada del repertorio de TCR $\beta$  de ratón tal como se describe en el Ejemplo 2. Los productos de PCR se subclonaron utilizando el vector TOPO-TA (Life Technologies), se transformaron post-ligamiento en células competentes TOP10 (Life Technologies) y se escogieron 48 colonias transformadas. A continuación, los clones se secuenciaron por secuenciación de Sanger para identificar las secuencias de clonotipo TCR $\beta$ . Todos los clones eran únicos, y representaban una amplia gama de posibles combinaciones V-J $\beta$ . Los plásmidos se mezclaron luego en un solo tubo, a través de tres órdenes de magnitud y con seis réplicas a cada una de las concentraciones.

La mezcla de 48 complejos se utilizó para optimizar el protocolo de amplificación de TCR $\beta$ . Se optimizaron la metodología de purificación después de la primera y segunda etapas de PCR, el número de ciclos en la primera PCR y la temperatura de reasociación en la primera PCR. Se utilizó la columna WA PCR o la escisión del gel para la tecnología de purificación. Debido al cebado falso erróneo, la primera ronda de PCR produjo múltiples bandas, además de una banda principal en el intervalo de tamaños diana de 150-200 pb. La escisión del gel separó el material no deseado, pero el proceso fue tedioso y resultó en una pérdida de hasta el 75% del material deseado. Protocolos con un menor número de primeros ciclos de amplificación por PCR producen típicamente una desviación de la amplificación menos grave, mientras que la desviación de la amplificación es típicamente distorsionada en los protocolos con > 30 ciclos. La temperatura de reasociación controla el rigor de los eventos de cebado, produciendo a temperaturas más bajas rendimientos más altos, pero menos especificidad.

68 bancos de Illumina se construyeron utilizando la mezcla de 48 plásmidos y variando los parámetros del protocolo tal como se describe anteriormente. Los bancos fueron secuenciados en una máquina de secuenciación de siguiente generación (Illumina) para obtener > 500k de etiquetas de 80 pb de extremos pareados para cada uno de los bancos. Para analizar los datos de secuenciación, cada una de las etiquetas de secuencia 2x80 pares de bases se alineó con las secuencias de los 48 clonotipos conocidos para obtener el mejor emparejamiento. Se contó el número de etiquetas alineadas a cada uno de los plásmidos para cada uno de los bancos y, a continuación, estos resultados se correlacionaron con las relaciones esperadas de los clones de plásmido de entrada. Se realizó un análisis de regresión lineal para adaptarse a cada uno de los conjuntos de datos (véase la Tabla 1: correlación del rendimiento,  $R^2$  de 1, y una pendiente de 1. El protocolo utilizó 15 ciclos de amplificación para la primera PCR, una temperatura de reasociación de 61°C, una purificación en columna por PCR después de la primera PCR, y purificación en gel después de la segunda PCR.

**Tabla 1.** Análisis de experimentos de optimización de protocolo piloto seleccionados.  $R^2$  y la pendiente se calcularon a partir de un análisis de regresión entre el recuento observado de las secuencias en cada uno de los bancos frente al recuento de entrada conocido. Las condiciones en la fila 3 (negrita) son un ejemplo de un protocolo optimizado.

Ciclos de 1ª PCR	Ta de 1ª PCR	Barrido en 1ª PCR	Barrido en 2ª PCR	R2	Pendiente
15	57	columna	gel	0,56	0,54
15	59	columna	gel	0,7	0,68
<b>15</b>	<b>61</b>	<b>columna</b>	<b>gel</b>	<b>0,72</b>	<b>0,71</b>
15	63	columna	gel	0,69	0,7
25	57	columna	gel	0,47	0,43
25	59	Columna	gel	0,44	0,4

25	61	Columna	gel	0,45	0,45
25	63	Columna	gel	0,41	0,39
35	57	Columna	gel	0,47	0,41
35	59	Columna	gel	0,43	0,37
35	61	Columna	gel	0,42	0,4
35	63	Columna	gel	0,41	0,4

**Ejemplo 4: Análisis de datos del Repertorio de TCRβ**

Debido a que el repertorio de TCRβ contiene tantos como 5x10<sup>6</sup> clonotipos, y las regiones CDR3 a menudo difieren en sólo unos pocos nucleótidos, una plataforma de análisis de encargo sofisticada era necesaria sólo para identificar los clones en el banco. Los métodos de tiempo de respuesta de alineación rápida, tales como BLAST (Altschul *et al.*, 1990), BLAT (Kent, 2002) y SOAP (Li *et al.*, 2008) eran inadecuados para la tarea en cuestión, ya que ocasionaban muchos emparejamientos falsos. Además de ello, los métodos de tiempo de respuesta de alta precisión tales como Smith-Waterman (Smith y Waterman, 1981) eran incómodamente lentos para este tipo de análisis. Por último, todos estos métodos requerirían un enorme banco de referencia (10<sup>15</sup> diversidad) de todas las posibles secuencias de CDR3 de nucleótidos, lo cual es una carga computacional.

Para hacer frente a estos problemas, se construyó un algoritmo que es más rápido que cualquier método actual en casi un orden de magnitud, y que tiene la misma precisión que los métodos de alineación estándares. Se genera una tabla de "palabras" de 4-8 nucleótidos que identifican de forma única los segmentos V y J de ratón o ser humano dentro de la región amplificada. La validez de cada emparejamiento se testa mediante la identificación de la distancia a y la secuencia de la segunda cisteína conservada. El emparejamiento fue aceptado como correcto sólo si tanto la distancia como la secuencia confirman el emparejamiento. Utilizando datos de los experimentos de secuenciación del repertorio de TCRβ de los autores de la invención, éstos identificaron típicamente ~ 99.98% de combinaciones V-Jβ sin ambigüedades. Las lecturas restantes fueron descartadas.

También emplearon otras dos etapas de control de calidad: (i) la región de CDR3 no debe contener ninguna secuencia de errores en forma de bases no exigidas; y (ii) la región de CDR3 está en el marco definido por la segunda cisteína conservada. Si se pasan todas las pruebas de calidad, el método identificó la secuencia codificadora de la proteína de la región de CDR3 dentro del marco de lectura conocido para ese gen particular. Este algoritmo aseguró velocidad, precisión y bajos índices de error. Se puede adaptar fácilmente para uso con otras familias de genes variables tales como TCRα o IgH.

Los autores de la invención realizaron un cierto número de experimentos para demostrar la utilidad de sus protocolos para la secuenciación de TCRβ profundo. Realizaron trasplantes de médula ósea de ratones en contextos genéticos coincidentes y no coincidentes. Para determinar el impacto sistémico de estos eventos de trasplante en los ratones, se examinaron los repertorios de células T del colon. Los clonotipos de TCRβ más comunes en colonos de trasplantes de médula ósea replicados no coincidentes estaban más estrechamente relacionados que los clonotipos más comunes de TCR en un colon de trasplante singénico, especialmente en la parte superior del 1% de los clones. Los perfiles de los colonos de control fueron casi idénticos en el 1% de los clonotipos. Estos datos indican que los protocolos descritos en esta memoria producen datos de calidad, cuantitativos de utilidad a los clientes de investigación.

**Ejemplo 5: Construcción de un Banco de Controles de Clones de TCRβ y Optimización de las Condiciones de PCR Utilizando el Banco de Controles**

Se llevan a cabo experimentos adicionales para construir un banco de 960 clones de TCRβ que contiene al menos un representante de cada uno de las 650 posibles combinaciones de V-Jβ humanas. Este conjunto de clones se utiliza para optimizaciones moleculares y estadísticas. Un banco de plásmidos de TCRβ humano se genera tal como se describe arriba en el Ejemplo 4. Se recogen aproximadamente 3.000 colonias transformantes y los clones se secuencian utilizando la secuenciación capilar estándar (p. ej., Sequetech). El emparejamiento de V-Jβ correspondiente a cada uno de los clones secuenciados se identifica tal como se describe arriba en el Ejemplo 4. El objetivo es obtener al menos un clon representativo para cada par de V-Jβ. Si la secuenciación encuentra que faltan algunos pares de V-Jβ, esos pares son rescatados mediante la creación de bancos de TCRβ utilizando sólo los cebadores de los pares de V-Jβ que faltan, se subclonan y secuencian. Después de varias rondas, los clones se identifican para cada posible par de V-Jβ. Estos plásmidos se mezclan en una sola mezcla de moldes, con 96 clones a cada concentración y 10 concentraciones diferentes a través de tres órdenes de magnitud.

**Ejemplo 6: Optimización de las Condiciones de la PCR Utilizando el Banco de Controles:**

Experimentos previos han demostrado que la primera amplificación por PCR provoca la mayor desviación de la amplificación. Se realizan experimentos adicionales utilizando la agrupación de 960 clones y secuenciación de siguiente generación para optimizar adicionalmente el número de ciclos de la primera PCR. Se generan

aproximadamente 60 bancos de TCR $\beta$  a partir de la mezcla de plásmidos, con cuatro réplicas para cada uno de los 15 ciclos de números entre 10 y 25. Las mezclas de los bancos se cuantifican y de cada uno de los bancos se obtienen ~ 4 millones de secuencias mediante un secuenciador GAllx de siguiente generación (Illumina). El emparejamiento de V-J $\beta$  correspondiente a cada uno de los clones secuenciados tal como se describe anteriormente en el Ejemplo 4, y los recuentos de etiquetas de las secuencias se cuentan para cada uno de los clones en cada conjunto de datos.

Trabajo previo ha demostrado que el contenido de GC puede afectar a la eficiencia de la amplificación (Markoulatos *et al.*, 2002). La inmensa diversidad de combinaciones de V(D)J $\beta$  da como resultado un surtido de contenidos y longitudes de GC. La desviación de la amplificación se testa después de la adición de diversos reactivos tales como betaína o cloruro de magnesio. Se generan aproximadamente 60 bancos de TCR $\beta$  a partir de la mezcla de plásmidos, con cuatro réplicas para cada uno de los 15 tampones diferentes. Las mezclas de los bancos se cuantifican y se obtienen ~ 4 millones de secuencias de cada uno de los bancos utilizando un secuenciador GAllx de siguiente generación (Illumina). Se identifica el emparejamiento V-J $\beta$  correspondiente a cada uno de los clones secuenciados tal como se describe arriba en el Ejemplo 4, y los recuentos de las etiquetas de secuencias se tabulan para cada uno de los clones en cada uno de los conjuntos de datos.

### **Ejemplo 7: Análisis de Células T y Seguimiento del Trasplante**

Los métodos de la invención se aplican al seguimiento inmune post-trasplante. Después de un trasplante alogénico (es decir, del riñón o del hígado), una respuesta de las células T de un huésped a los trasplantes se evalúa para controlar la salud del huésped y el injerto. El seguimiento molecular de la sangre o la orina es útil para detectar un rechazo agudo o crónico antes de que típicamente estuviera indicada una biopsia. Por ejemplo, la detección de alo-anticuerpos contra el antígeno leucocitario humano (HLA) se ha asociado con el rechazo crónico de aloinjerto (Terasaki y Ozawa, 2004 *American Journal of Transplantation* 4:438-43). Otros marcadores moleculares incluyen microglobulina b<sub>2</sub>, neopterinina y citocinas proinflamatorias en la orina y en la sangre (Sabek *et al.*, 2002 *Transplantation* 74:701-7; Tatapudi *et al.*, 2004 *Kidney International* 65: 2390; Matz *et al.*, 2006 *Kidney International* 69: 1683; Bestard *et al.*, 2010 *Current Opinion in Organ Transplantation* 15: 467-473). Sin embargo, ninguno de estos métodos ha sido ampliamente adoptado en la práctica clínica, tal vez debido a su baja especificidad y sensibilidad. El trabajo previo ha demostrado que las células T reguladoras (Treg) inducen la tolerancia del injerto al sub-regular células T colaboradoras (Th) (Graca *et al.*, 2002 *Journal of Experimental Medicine* 195:1641). Adicionalmente, el trasplante de células madre hematopoyéticas de donantes de emparejamiento erróneo de HLA en el receptor ha dado como resultado la tolerancia del trasplante renal no-inmunosupresivo hasta 5 años después del trasplante (Kawai *et al.*, 2008 *NEJM* 358:353-61).

Se diseñan cebadores que fijan como objetivo transcripciones de varios genes de funcionalidad inmune (descritos anteriormente), que producen construcciones de fusión de extensión por solapamiento con amplicones CDR3 $\beta$ . En una realización, estos cebadores se diseñan para amplificar específicamente ADNc que abarca uniones de corte y empalme de ARN y se hibrida a ADNc a partir de ARN mensajero procesado. Ejemplos de moléculas que están asociadas con la función inmunológica incluyen, pero no se limitan a, T-BET e IFN-g, que indican células T colaboradoras 1 (Th1); GATA3 e IL-4, que indican células T colaboradoras 2 (Th2); IL-17 que indica células T colaboradoras 17 (Th17); y FoxP3 e IL-10, que indican células T reguladoras (Treg). Tales moléculas de señalización son miembros de grandes familias de proteínas con una fuerte homología entre parálogos, lo cual puede dar lugar a la amplificación de fondo durante la PCR. Por consiguiente, se generan alineaciones de nucleótidos de todos los parálogos en cada una de las familias (es decir, todos los genes de interleucina) y se diseñan cebadores de la PCR de forma que abarquen exones y tengan la homología de secuencia más baja posible con otros genes en la familia.

El seguimiento de células T funcionales implica las siguientes etapas: (i) aislamiento de células mononucleares de la sangre periférica individuales en los reactores de microgotitas de emulsión; (ii) amplificación de extensión por solapamiento de complejos entre TCR $\beta$  y moléculas de funcionalidad inmunes en reactores de microgotitas; y (iii) la reversión de la emulsión, seguida de secuenciación a granel. El TCR $\beta$  y conjuntos de cebadores de funcionalidad inmunes se combinarán para producir construcciones de fusión de amplicón principal a partir de los amplicones menores. Los cebadores de extensión por solapamiento son una combinación de los cebadores de TCR $\beta$  inversos con aproximadamente la mitad de cada uno de los cebadores directos con moléculas de funcionalidad inmune, lo que resulta en un total de 91 cebadores de TCR $\beta$  de fusión inversa. Los cebadores de fusión entre el cebador directo para cada uno de los amplicones menores de funcionalidad inmune contienen aproximadamente la mitad de cada uno de los 13 cebadores de TCR $\beta$  inversa, para un total de 91 cebadores de TCR $\beta$  de funcionalidad inmune de fusión inversa. El resultado final es que el solapamiento entre cualquier par de TCR $\beta$  y amplicones menores de funcionalidad inmune tiene una temperatura de fusión de aproximadamente 55-65°C, de tal manera que cada uno de los amplicones menores actúa como un cebador para el amplicón emparejado. En las mezclas de reacción finales, los cebadores externos se diluyen a una concentración final de 0,1  $\mu$ M, y los cebadores internos se diluyen a 0,01  $\mu$ M, de manera que los cebadores internos son reactivos limitantes.

### **Ejemplo 8: Análisis de Células T y Diagnóstico de Tuberculosis Latente**

La tuberculosis latente (TB) es una epidemia mundial principal, que afecta a tanto como 2 billones de personas en todo el mundo. Actualmente no existe un test fiable para el diagnóstico clínico de la TB latente. Esta brecha tecnológica tiene consecuencias clínicas graves, ya que la TB reactivada es el único sello fiable de TB latente. Además de ello, los ensayos clínicos para vacunas y terapias carecen de biomarcadores para la TB latente y, por lo tanto, deben seguir las cohortes durante muchos años para demostrar la eficacia.

La principal vacuna actual contra la tuberculosis, el bacilo de Calmette-Guérin (BCG), es un profiláctico poco fiable. En un meta-análisis de docenas de estudios epidemiológicos, el efecto global del BCG fue del 50% contra las infecciones de TB, del 78% contra la TB pulmonar, del 64% contra la meningitis TB y del 71% contra la muerte debida a la infección de TB (Colditz *et al.*, 1994 *JAMA* 271:698-702). Adicionalmente, el rápido aumento de la TB resistente a múltiples fármacos ha aumentado la necesidad de nuevos enfoques de vacunas e inmunoterapias. Hasta el 90% de los individuos infectados, inmunocompetentes no progresan nunca a la enfermedad, lo que resulta en la enorme reserva mundial de TB latente (Kaufmann, 2005 *Trends in Immunology* 26:660-67).

Dado que la tuberculosis es un patógeno intracelular facultativo, la inmunidad está mediada casi en su totalidad a través de las células T. Células T colaboradoras 1 (Th1) que expresan el interferón-g suscitan la respuesta primaria de la TB, con una cierta intervención de células T colaboradoras 2 (Th2). Después de la respuesta primaria, las bacterias se vuelven latentes, controladas por células T reguladoras (Treg) y células T de la memoria (Tmem). Recientemente, once nuevos candidatos vacunales han iniciado ensayos clínicos (Kaufmann, 2005 *Trends in Immunology* 26:660-67). Estas vacunas son todas vacunas "post-exposición", es decir, fijan como objetivo las respuestas de células T a la TB latente y están destinadas a evitar la reactivación de la enfermedad. Debido al fracaso parcial de BCG para inducir una inmunidad total, el diseño racional y la validación de futuras vacunas de TB deben incluir un análisis sistemático de la respuesta inmune específica tanto para la TB como para las nuevas vacunas.

Durante décadas, el patrón de cuidado para el diagnóstico de la tuberculosis latente ha sido la prueba de la tuberculina en la piel (TST) (Pai *et al.*, 2004 *Lancet Infectious Disease* 4:761-76). Más recientemente, se han desarrollado dos ensayos comerciales *in vitro* de interferón-g: el ensayo QuantiFERON-TB y el ensayo T SPOT-TB. Estos ensayos miden la inmunidad mediada por células mediante la cuantificación de interferón-g liberado de las células T cuando se enfrentan a un cóctel de antígenos de la tuberculosis. Lamentablemente, ni el TST ni los ensayos más recientes de interferón-g son eficaces para distinguir TB latente de clara (Diel *et al.*, 2007 *American Journal of Respir Crit Care Med* 177:1164-70). Este es un problema importante, ya que los pacientes sin evidencia clínica de TB latente (es decir, visualización de granulomas) pero con TST positiva o prueba de interferón-g reciben típicamente terapia con isoniazida durante 6-9 meses, a pesar de que esta intervención empírica es innecesaria en los pacientes que han eliminado la infección primaria y puede causar complicaciones graves tal como insuficiencia hepática.

El trabajo previo ha demostrado que las respuestas de células T se utilizan para distinguir la TB latente de la activa (Schuck *et al.*, 2009 *PLoS One* 4: e5590). La premisa de este trabajo previo es que las células inmunes dirigidas contra antígenos de la TB aumentarán en la población de células T de memoria si la TB es latente, pero aumentarán en una fracción de células T colaboradoras si la TB es activa. La secuenciación de células T funcionales se utiliza para distinguir la TB latente de TB clara. El protocolo implica: (i) la captura de células T individuales en microgotitas de emulsión; (ii) la transcripción inversa y la amplificación de microgotitas en loci diana; (iii) la síntesis de microgotitas de complejos de fusión entre dos o más loci diana; y (iv) invertir emulsiones y secuenciar amplicones principales con la secuenciación a granel. La PCR específica para la secuencia se utiliza después de la RT-PCR de extensión por solapamiento para detectar la presencia de un biomarcador particular para la TB latente.

#### **Ejemplo 9: Análisis de Células T y Diagnóstico o Seguimiento de la Enfermedad**

De manera similar, el seguimiento de células T funcionales se utiliza para el diagnóstico y el seguimiento de casi cualquier enfermedad humana. Estas enfermedades incluyen, pero no se limitan a lupus eritematoso sistémico (SLE), alergia, enfermedades autoinmunes, trasplantes de corazón, trasplantes de hígado, trasplantes de médula ósea, trasplantes de pulmón, tumores sólidos, tumores líquidos, síndrome mielodisplásico (MDS), infección crónica, infección aguda, hepatitis, virus del papiloma humano (HPV), virus herpes simplex, citomegalovirus (CMV) y virus de la inmunodeficiencia humana (VIH). Dicho seguimiento incluye el diagnóstico y seguimiento individuales o el seguimiento de la población para los estudios epidemiológicos.

El seguimiento de células T se utiliza para fines de investigación utilizando cualquier sistema de modelo no humano, tales como el pez cebra, ratón, rata o conejo. El seguimiento de células T también se utiliza para fines de investigación utilizando cualquier sistema de modelo humano tal como líneas de células T primarias o líneas de células T inmortales.

#### **Ejemplo 10: Análisis de células B**

Los anticuerpos son producidos por secuencias recombinadas de inmunoglobulina genómica (Ig) en células de linaje B. Cadenas ligeras de inmunoglobulina se derivan de genes  $\kappa$  o  $\lambda$ . Los genes  $\lambda$  se componen de cuatro genes de región constante (C) y aproximadamente treinta genes de región variable (V). Por el contrario, los genes  $\kappa$  se componen de un gen de la región C y 250 genes de la región V. La familia de genes de la cadena pesada está compuesta de varios cientos de segmentos de genes V, quince segmentos de genes D y cuatro segmentos de genes de unión (J). La recombinación somática durante la diferenciación de células B elige al azar una combinación V-D-J de la cadena pesada y una combinación V-J en la cadena ligera  $\kappa$  o  $\lambda$ . Debido a que hay tantos segmentos de genes, son posibles millones de combinaciones únicas. Las regiones V también se someten a una hipermutación somática después de la recombinación, generando una diversidad adicional. A pesar de esta complejidad subyacente, es posible utilizar docenas de cebadores que fijan como objetivo secuencias conservadas para secuenciar el complemento completo de la cadena pesada y ligera en varias reacciones multiplexadas (van Dongen *et al.*, 2003 *Leukemia* 17:2257-2317).

Cualquiera de las subunidades de inmunoglobulina individuales está enlazada a las moléculas de funcionalidad inmune que indican la actividad o sub-poblaciones de células B. Una primera secuencia de ácido nucleico diana, una segunda secuencia de ácido nucleico diana o ambas secuencias de ácidos nucleicos diana pueden comprender una secuencia de inmunoglobulina. Alternativamente, la primera secuencia de ácido nucleico diana puede comprender una secuencia de inmunoglobulina, y la segunda secuencia puede comprender una segunda molécula asociada con la función celular inmune. Ejemplos de moléculas de marcador de células B funcionales incluyen, pero no se limitan a un complejo principal de histocompatibilidad (MHC), cúmulo de diferenciación 19 (CD 19), receptor de interleucina 7 (receptor de IL-17), cúmulo de diferenciación 10 (CD10), cúmulo de diferenciación 20 (CD20), cúmulo de diferenciación 22 (CD22), cúmulo de diferenciación 34 (CD34), cúmulo de diferenciación 27 (CD27), cúmulo de diferenciación 5 (CD5) y cúmulo de diferenciación 45 (CD45), cúmulo de diferenciación 38 (CD38), cúmulo de diferenciación 78 (CD78), receptor de interleucina-6, factor regulador de interferón 4 (IRF4) y cúmulo de diferenciación 138 (CD138). Una agrupación de cebadores que amplifica el complemento IgH completo de células B se combina con un solo par de cebadores marcadores de células B. Esto ensaya todos los clonotipos de células B en un grupo funcional particular tal como Bmem. Alternativamente, una agrupación de cebadores que amplifica el complemento completo de IgH de células B se combina con docenas de pares de cebadores de marcadores de células B. Este ensayo proporciona el fenotipo completo de cada uno de los clonotipos en la mezcla de células.

Se proporciona un método para enlazar IgH e Ig $\kappa$ . IgH e Ig $\lambda$  están enlazados en células individuales a moléculas de funcionalidad inmune que indican la actividad o sub-poblaciones de células B. La gran mayoría de la diversidad en el repertorio de células B está compuesta por las regiones V-D-J de regiones IgH y V-J de Ig $\kappa$  (Sandberg *et al.*, 2005 *Journal of Molecular Diagnostics*. 7:495-503; Boyd *et al.*, 2009 *Science Translational Med* 1:12ra23). Agrupaciones de cebadores previamente reseñadas (van Dongen *et al.*, 2003 *Leukemia* 17:2257-2317) se utilizan para amplificar estas regiones de IgH e Ig $\kappa$ . Cinco agrupaciones de cebadores en reacciones separadas se utilizan para amplificar el complemento de IgH e Ig $\kappa$  de un ser humano sano. El material amplificado se secuenció con la secuenciación a granel. Para analizar los resultados de la secuenciación a granel, se utiliza el algoritmo IgBLAST y la base de datos para determinar las uniones V-D y D-J de IgH y alinear las secuencias de IgH e Ig $\kappa$  a segmentos de genes de la línea germinal. En general, este método está más paralelizado que los métodos previamente reseñados para el análisis de Ig de células individuales (Patente de EE.UU. 7.749.697).

#### **Ejemplo 11: Análisis de células B y Descubrimiento de Fármacos**

Los productos terapéuticos de anticuerpos se utilizan cada vez más por las compañías farmacéuticas para el tratamiento de enfermedades incurables tales como el cáncer (Carter 2006 *Nature Reviews Immunology* 6:343-357). Sin embargo, el proceso de descubrimiento de fármacos de anticuerpos es caro y tedioso, requiriendo la identificación de un antígeno, y a continuación, el aislamiento y la producción de anticuerpos monoclonales con actividad contra el antígeno. Los individuos que han estado expuestos a una enfermedad producen anticuerpos contra antígenos asociados con esa enfermedad, por lo que es posible explotar los repertorios inmunológicos de los pacientes para anticuerpos que podrían utilizarse para el desarrollo de productos farmacéuticos. Sin embargo, un anticuerpo monoclonal funcional requiere tanto de inmunoglobulinas de cadena pesada como ligera. La PCR de extensión de solapamiento y/o la RT-PCR de extensión por solapamiento en microgotitas de emulsión de células individuales se utiliza para capturar secuencias de anticuerpos funcionales a partir de repertorios de células B del paciente. En síntesis, el método implica las siguientes etapas: (i) aislamiento de células B individuales en microrreactores acuosos-en-aceite utilizando un dispositivo de microfluidos; (ii) el enlace molecular entre amplicones de inmunoglobulina de cadena pesada y ligera (IgH e Ig $\kappa$ ) dentro de los microrreactores de células individuales; y (iii) la inversión de las emulsiones, seguido por secuenciación a granel de las secuencias de ácido polinucleico enlazadas. Esto produce emparejamientos de cadena pesada y ligera de millones de células B individuales analizadas en paralelo, que se explotan como agentes terapéuticos potenciales.

Las secuencias de los cebadores de fusión para la PCR de extensión por solapamiento y la RT-PCR de extensión por solapamiento son idénticas a los cebadores IgH e Ig $\kappa$  independientes, excepto que determinados cebadores contienen secuencias adicionales de polinucleótidos para la extensión por solapamiento: (i) el cebador directo del locus IgH tiene una secuencia de 10-20 nt al azar sin complementariedad con cualquiera de las dianas; (ii) el

cebador inverso de los loci IgH tiene una secuencia de 10-20 nt con complementariedad con el cebador directo de Igk, y (iii) el cebador directo de Igk tiene complementariedad con los cebadores inversos para el locus IgH. En las mezclas de reacción finales, los cebadores externos se diluyen a una concentración final de 0,1  $\mu$ M, y los cebadores internos se diluyen a 0,01  $\mu$ M, de tal manera que los cebadores internos serán un reactivo limitante. Esto impulsa la formación del amplicón principal.

#### **Ejemplo 12: Análisis de Células B y Seguimiento de la Inmunidad**

Células B de memoria humores (Bmem) ayudan a los sistemas inmunes de mamíferos a conservar determinados tipos de inmunidad. Después de la exposición a un antígeno y de la expansión de células productoras de anticuerpos, las células Bmem sobreviven durante muchos años y contribuyen a la respuesta inmune secundaria tras la re-introducción de un antígeno. Esta inmunidad se mide típicamente en un ensayo *in vitro* celular o basado en anticuerpos. En algunos casos, es beneficioso detectar la inmunidad mediante la amplificación, el enlace y la detección de las regiones variables de IgH y de inmunoglobulina de cadena ligera en células B individuales. Un método de este tipo es más específico y sensible que los métodos actuales. La secuenciación del repertorio de células B masiva en paralelo se utiliza tal como se describe en el Ejemplo 13 para el rastreo de células Bmem que contienen un determinado emparejamiento de cadena pesada y ligera que es indicativo de la inmunidad. En otro método a modo de ejemplo, se combina un emparejamiento de cadena pesada y ligera de células individuales con la secuenciación de células B funcional, es decir, desarrollando cebadores de RT-PCR de extensión por solapamiento que fijan como objetivo transcripciones de ARN que están sobre-representadas en células Bmem (es decir, CD27). Mediante la combinación de la amplificación de inmunoglobulina ligera y pesada con la expresión génica de transcripciones de la función inmune de células Bmem o plasmáticas, se evitan la clasificación de células por FACS u otros métodos tediosos.

#### **Ejemplo 13: Análisis de Células B y Diagnóstico y Seguimiento de Enfermedades**

El seguimiento de células B se utiliza para el diagnóstico y el seguimiento de casi cualquier enfermedad humana. Estas enfermedades incluyen, pero no se limitan a lupus eritematoso sistémico (SLE), alergia, enfermedades autoinmunes, trasplantes de corazón, trasplantes de hígado, trasplantes de médula ósea, trasplantes de pulmón, tumores sólidos, tumores líquidos, síndrome mielodisplásico (MDS), infección crónica, infección aguda, hepatitis, virus del papiloma humano (VPH), virus herpes simplex (HSV), citomegalovirus (CMV) y virus de la inmunodeficiencia humana (VIH). Dicho seguimiento podría incluir el diagnóstico y seguimiento individuales o el seguimiento de la población para los estudios epidemiológicos.

El seguimiento de células B también se utiliza para fines de investigación utilizando cualquier sistema de modelo no humano tal como el pez cebra, ratón, rata o conejo. El seguimiento de células B se utiliza con fines de investigación utilizando cualquier sistema modelo humano, tal como líneas de células B primarias o líneas de células B inmortales.

#### **Ejemplo 14: Métodos para el Diagnóstico Prenatal No Invasivo**

En ausencia de un diagnóstico prenatal, aproximadamente el 2% de los bebés tienen discapacidades físicas o mentales graves, aproximadamente el 3,3% de los bebés tienen algún tipo de malformación congénita y aproximadamente el 0,5% tiene una anomalía cromosómica fenotípicamente significativa. Métodos clínicos actuales para el diagnóstico prenatal son invasivos y presentan riesgos significativos para el feto, limitando su uso a pacientes de edad materna avanzada. Se necesitan tecnologías no invasivas y precisas para el diagnóstico genético prenatal en el primer trimestre. La mayoría de los actuales métodos preclínicos para el diagnóstico prenatal no invasivo capturan y diagnostican células fetales circulantes. Estos métodos se basan en proteínas de la superficie celular y/o en la morfología celular para enriquecer poblaciones particulares de células fetales. Tales enfoques defectuosos no han logrado llegar a la clínica a pesar de décadas de intensa investigación y desarrollo.

El aislamiento de glóbulos rojos fetales nucleados en circulación (FNRBCs) de la sangre materna es un enfoque para el diagnóstico prenatal no invasivo. Glóbulos rojos nucleados son uno de los primeros tipos de células hematopoyéticas producidas durante el desarrollo fetal. Estas células atraviesan la placenta y son detectables a bajas concentraciones en la sangre materna durante el primer trimestre (Ganshirt *et al.*, 1994 *Lancet* 343:1038-9). Otra característica atractiva de los FNRBCs es su corta vida en comparación con otros tipos de células fetales circulantes (Pearson, 1967, *Journal of Pediatrics* 70:166-71), haciendo poco probable que persistan en la sangre materna de embarazos anteriores.

La escasez de células fetales circulantes, estimada en una célula fetal por cada  $10^5$ - $10^9$  células maternas (Price *et al.*, 1991 *Am J Obstet Gynecol* 165:1731-7; Ganshirt-Ahlert *et al.*, 1994 *Clin Genet* 38:38-43), hace necesario el uso de métodos de enriquecimiento de células fetales sensibles y específicos antes del diagnóstico. Métodos de enriquecimiento ampliamente adoptados incluyen combinaciones de centrifugación en gradiente de densidad (Samura *et al.*, 2000 *Prenat Diagn* 20:281-6), la clasificación de células activadas por fluorescencia (FACS) y la clasificación celular magnética (MACS) (Busch *et al.*, 1994 *Ann NY Acad Sci* 731:144-6). A pesar del desarrollo de estos métodos, ninguno ha sido comercializado.

### **i. Métodos para el Diagnóstico Prenatal No Invasivo de Trastornos de Genes Individuales**

La LCR o sondas-candado se utilizan para capturar y amplificar alelos específicos paternos de una manera específica para el alelo y para llevar a cabo la PCR de extensión por solapamiento para detectar alelos de la enfermedad (FIGs. 26-30). El método implica las siguientes etapas: (i) genotipado parental para encontrar polimorfismos específicos paternos; (ii) aislamiento de células mononucleares individuales de la sangre materna en microgotitas de emulsión; (iii) amplificación de la enfermedad y loci de "enlazador" específicos paternos mediante un protocolo de LCR/PCR modificado en reactores de microgotitas de emulsión; (iv) la amplificación de extensión por solapamiento de complejos entre la enfermedad y loci de enlazador en reactores de microgotitas; (v) recuperación de los complejos enlazados por inversión de la emulsión; y (vi) secuenciación masiva en paralelo. Los datos de secuenciación masiva en paralelo se analizan para cuantificar los casos de genotipos enlazados. Solamente los reactores de microgotitas que contienen células fetales individuales proporcionan complejos enlazados entre el locus de la enfermedad y el alelo específico paternal. Ambos alelos se amplifican a partir de la célula fetal, proporcionando al médico el estado como un vehículo, homocigoto normal u homocigoto afectado.

Se diseñan sondas de LCR para fijar como objetivo un locus asociado con una enfermedad y un locus de SNP enlazador. Las sondas de LCR son de 20-30 nucleótidos de largo y tienen temperaturas de fusión ( $T_m$ ) de aproximadamente 55-65°C. Los nucleótidos 5' se fosforilan y las sondas se diseñan para minimizar la auto-complementariedad de la sonda, así como la complementariedad entre sondas. Además de las regiones de complementariedad a loci diana, tres de las sondas incluyen secuencias de polinucleótidos que permiten la amplificación después de ligamiento: (i) la sonda 5' para el locus de la enfermedad tiene una secuencia al azar de 10-20 nt sin complementariedad con cualquiera de los locus diana; (ii) la sonda 3' para el locus de la enfermedad tiene una secuencia de 10-20 nucleótidos con complementariedad al extremo 5' del locus SNP de enlazador; y (iii) la sonda 5' para el locus SNP de enlazador tiene complementariedad con el extremo 3' del locus de la enfermedad (FIGs. 26-30).

Para cada par de locus de la enfermedad y del enlazador, se formula una mezcla de reacción utilizando ADN genómico de líneas celulares, las sondas de LCR, los cebadores de la PCR, Ampligase (Epicentre), fragmento de ADN polimerasa Stoffel (Life Technologies) y tampón de reacción (según Hardenbol *et al.*, 2005; Tris-HCl 20 mM, KCl 25 mM,  $MgCl_2$  10 mM, NAD 0,5 mM, Triton X-100 al 0,01%). Las sondas "interiores" se añaden a 1/10 de la concentración de los otros oligonucleótidos en la reacción. Para la reasociación inicial y el ligamiento, las mezclas se incuban durante 4 minutos a 20°C, durante 5 minutos a 95°C y durante 15 minutos a 60°C. Después, las condiciones de termociclación por PCR se utilizan para amplificar los amplicones menores y principales (p. ej., 95°C, 5 minutos; [95°C, 30 segundos; 60°C, 30 segundos; 72°C, 30 segundos] x 30 ciclos).

Después de la secuenciación a granel de los amplicones principales, la enfermedad y los alelos no afectados se analizan para diagnosticar el feto como homocigoto normal, heterocigoto portador u homocigoto afectado. En los heterocigotos portadores, los amplicones principales enlazados al alelo específico paternal comprende aproximadamente 50% de alelos de la enfermedad y 50% de alelos normales. De manera similar, en los homocigotos portadores, los amplicones principales enlazados a los alelos específicos paternos comprenden casi el 100% de los alelos de la enfermedad. Este método se puede extender más allá de mutaciones de un solo nucleótido para encontrar patrones de expresión génica específicos para alelos paternos y/o análisis multiplexado de muchas mutaciones de la línea germinal en las células fetales circulantes.

Ejemplos de genes que a menudo están mutados y son de interés en el diagnóstico prenatal incluyen, pero no se limitan a receptor transmembrana de la fibrosis quística (CFTR), aspartoacilasa (ASPA), anemia de Fanconi, grupo de complementación C (FANCC), glucosa-6-fosfatasa (G6CP), glucocerebrosidasa (GBA), hexosaminidasa A (HEXA), hemoglobina beta (HBB), frataxina (FXN), receptor de lipoproteínas de baja densidad (LDLR) y proteína de unión 2 a metil CpG (MECP2).

Por ejemplo, en la FIG. 26 se ilustra el enlace de la secuencia de células individuales mediante reacción en cadena de la ligasa en combinación con la reacción en cadena de la polimerasa de extensión por solapamiento tal como se aplica a un método para el diagnóstico prenatal no invasivo. El ácido nucleico diana (g) es un alelo específico paternal, el ácido nucleico diana (h) es un primer alelo de la enfermedad y el ácido nucleico diana (i) es un segundo alelo de la enfermedad. En particular, los dos alelos (h) e (i) se amplifican en cualquier célula (j) que contiene la variante específica paternal y no se producen amplicones principales en células que carecen de la variante de nucleótidos específica paternal. El cebador (a) es una sonda de LCR directa y el cebador (b) es una sonda de LCR inversa para la amplificación de ácido nucleico diana (g). El cebador (e) es un cebador de PCR directo y el cebador (f) es un cebador de PCR inverso para los dos alelos de la enfermedad (h) e (i). El cebador directo que fija como objetivo el locus de la enfermedad tiene una región de complementariedad con la sonda inversa que fija como objetivo la variante de nucleótidos específica paternal. El proceso puede llevarse a cabo en una gotita de emulsión o contenedor de reacción (k). La FIG. 27 también muestra un ejemplo de hibridación de los cebadores y los ácidos nucleicos diana en un enlace de secuencia de célula individual mediante reacción en cadena de la ligasa en combinación con la reacción en cadena de la polimerasa de extensión por solapamiento, tal como se aplica a un

método para el diagnóstico prenatal no invasivo, de acuerdo con una realización de la invención. El proceso se lleva a cabo en una gotita de emulsión o recipiente de reacción (k).

Además, la FIG. 28 muestra un ejemplo de amplicones resultantes producidos en un enlace de secuencia de célula individual mediante reacción en cadena de la ligasa en combinación con la reacción en cadena de la polimerasa de extensión por solapamiento tal como se aplica a un método para el diagnóstico prenatal no invasivo, de acuerdo con una realización de la invención. La FIG. 29 muestra la hibridación de la superposición de las regiones complementarias de los amplicones resultantes y la reacción en cadena de la polimerasa de extensión por solapamiento tal como se aplica a un método para el diagnóstico prenatal no invasivo, de acuerdo con una realización de la invención. La FIG. 30 ilustra los amplicones resultantes que se producen a partir de la reacción en cadena de la polimerasa de extensión por solapamiento tal como se aplica a un método para el diagnóstico prenatal no invasivo. El producto final es un banco de "amplicones principales", o loci enlazados, que luego pueden ser secuenciados a granel.

## **ii. Métodos para el Cariotipado Molecular Prenatal No Invasivo**

Métodos para la detección de enfermedades genéticas están adaptados para el cariotipado molecular prenatal no invasivo. Un método de este tipo implica las siguientes etapas: (i) genotipado parental para encontrar polimorfismos específicos paternos; (ii) aislamiento de células mononucleares individuales de la sangre materna de microgotas de emulsión; (iii) amplificación de la enfermedad y loci de "enlazador" específicos paternos mediante un protocolo LCR/PCR modificado en reactores de microgotitas de emulsión; (iv) amplificación de extensión por solapamiento de complejos entre decenas de miles a cientos de miles de sondas cromosómicas y loci de enlazador en reactores de microgotitas; (v) recuperación de los complejos enlazados por inversión de emulsión; y (vi) secuenciación masiva en paralelo. Los datos de secuenciación masiva en paralelo se analizan para cuantificar los casos de genotipos enlazados. Solamente los reactores de microgotitas que contienen células fetales individuales proporcionan complejos enlazados entre las sondas cromosómicas y el alelo específico paterno. Las sondas cromosómicas se utilizan para cuantificar el número de cromosomas o segmentos de cromosomas presentes en las células fetales y, por asociación, en el feto. El número de copias de cromosomas se cuantifica mediante la comparación de los recuentos de secuencia de un cromosoma desconocido a recuentos de la secuencia de un cromosoma de referencia conocido dentro de un único experimento, o mediante la búsqueda de desequilibrio alélico (Johnson *et al.*, 2010 Human Reproduction 25:1066-1075). Este método también se utiliza para detectar una diversidad de trastornos cromosómicos, incluyendo aneuploidía, trastornos cromosómicos estructurales desequilibrados, microdeleciones, microinserciones y otros tipos de trastornos congénitos. Ejemplos de trastornos de interés incluyen trisomía 13, trisomía 18 y trisomía 21.

### **Ejemplo 15: Métodos de Diagnóstico No Invasivo del Cáncer**

La comunidad médica ha buscado durante mucho tiempo el diagnóstico y seguimiento no invasivo de pacientes con cáncer, y ya existe un método aprobado por la FDA (CellSearch, Veridex) para la cuantificación de células tumorales circulantes para pacientes con cáncer de próstata y de mama. Métodos no invasivos para el diagnóstico pueden permitir la estadificación molecular de tumores antes de la biopsia, la cual puede reducir el costo y conducir a mejores resultados clínicos. Después del tratamiento, se utilizan métodos no invasivos para evaluar el éxito del régimen de tratamiento sin la necesidad de una re-biopsia invasiva y cara. Existe un consenso general entre los médicos de que los métodos no invasivos para la caracterización de los tumores beneficiarían en gran medida a los pacientes y aumentarían la probabilidad de resultados favorables.

La PCR de extensión por solapamiento de células individuales, la LCR, sondas-candado y/o RT-PCR se utilizan para analizar específicamente sólo a las células tumorales en poblaciones de células heterogéneas tales como líquido cefalorraquídeo (CSF) o la sangre (FIGs. 18-25). A diferencia de los métodos actuales, este enfoque evita por completo las complejidades provocadas por las diferencias en marcadores de la superficie celular y la morfología. Tales métodos son particularmente útiles en los cánceres en los que una biopsia es invasiva y costosa, y las decisiones de tratamiento tales como las decisiones de terapia farmacológica, se beneficiarían de análisis molecular del tumor. La tecnología se utiliza para cualquier tipo de tumor o cualquier tipo de problema genético o combinación de problemas genéticos en los tumores.

Los métodos descritos anteriormente en las Secciones I y II también se utilizan para detectar un gen o SNP asociado con el cáncer. La PCR de extensión por solapamiento de células individuales, la LCR, sondas-candado y/o la RT-PCR se utilizaron para amplificar un primer ácido nucleico o un segundo ácido nucleico que está asociado con el cáncer. El primer ácido nucleico diana incluye una mutación somática rara y la segunda diana es una transcripción de genes asociada con el cáncer. Alternativamente, una secuencia es un código de barras molecular y la segunda secuencia es una secuencia de mutación rara o una transcripción de genes asociada con el cáncer. En cualquiera de las alternativas, niveles más altos de multiplexación producen patrones de expresión de células individuales de 10, 100, 1000, 10.000 transcripciones o incluso todas las transcripciones en la célula. Niveles más altos de multiplexación también pueden producir perfiles de mutación de genes enteros, o muchos genes enteros, o incluso todo el genoma. La secuencia del gen rara está presente en menos de 5% de las células, menos de 1% de las

células o menos de 0,1% de las células. La secuencia del gen rara resulta de una mutación genética. La mutación genética puede ser una mutación somática. La mutación genética puede ser una mutación en un gen seleccionado del grupo que consiste en: receptor del factor de crecimiento epidermal (EGFR), homólogo de fosfatasa y tensina (PTEN), proteína de tumor 53 (p53), MutS homólogo 2 (MSH2), neoplasia endocrina múltiple 1 (MEN1), poliposis adenomatosa coli (APC), receptor Fas (FASR), proteína retinoblastoma (Rb1), Janus quinasa 2 (JAK2), factor de transcripción 1 similar a (ETS) (ELK1), virus de la eritroblastosis aviar v-ets E26 homólogo de oncogén 1 (ETS1), cáncer de mama 1 (BRCA1), cáncer de mama 2 (BRCA2), receptor del factor de crecimiento de hepatocitos (MET), protocolo-oncogén ret (RET), leucemia eritroblástica viral V-erb-b2 homólogo de oncogén 2 (HER2), sarcoma de rata viral V-Ki-ras2 Kirsten homólogo de oncogén (KRAS), linfoma de células B 2 (BCL2), mielocitomatosis viral V-myc homólogo de oncogén (MYC), gen de neurofibromatosis de tipo 2 (NF2), mieloblastosis viral v-myb homólogo de oncogén (MYB) y homólogo mutS 6 (E. coli) (MSH6). La transcripción asociada al cáncer es un gen seleccionado del grupo que consiste en molécula de adhesión celular epidermal (EpCAM), leucemia eritroblástica viral V-erb-b2 homólogo de oncogén 2 (HER2), receptor de estrógeno (ER), transductor de señales y activador de la transcripción 3 (STAT3), proteínas de unión al potenciador CCAAT (C/EBP), antígeno prostático específico (PSA), receptor de andrógenos (AR), receptor de progesterona (PR), Jun B (JUNB), proteína Rab-31 relacionada con Ras (RAB31), proteína de respuesta de crecimiento temprano 1 (EGR1), linfoma de células B 2 (BCL2), proteína C-ets-1 (ETS1), osteosarcoma murino viral FBJ homólogo de oncogén (c-Fos) y factor de crecimiento 1 similar a la insulina (IGF-1). Transductor de señal y activador de la transcripción 2 (STAT2) (Irgon *et al.*, 2010 *BMC Cancer* 10:319).

Las transcripciones asociadas al cáncer pueden multiplexarse para producir una señal de 10, 100, 1000, 10.000 transcripciones, o la totalidad de las transcripciones en la célula, que es analizada por secuenciación de siguiente generación para identificar una mutación. La mutación se asocia con el cáncer. El cáncer se selecciona del grupo que consiste en carcinoma de pulmón, cáncer de pulmón de células no pequeñas, cáncer de pulmón de células pequeñas, cáncer de útero, cáncer de tiroides, carcinoma de mama, carcinoma de próstata, carcinoma de páncreas, carcinoma de colon, linfoma, linfoma de Burkitt, linfoma de Hodgkin, leucemia mieloide, leucemia, sarcoma, blastoma, melanoma, seminoma, cáncer de cerebro, glioma, glioblastoma, astrocitoma cerebeloso, linfoma de células T cutáneo, cáncer gástrico, cáncer de hígado, endimoma, cáncer de laringe, cáncer de cuello, cáncer de estómago, cáncer de riñón, cáncer de páncreas, cáncer de vejiga, cáncer de esófago, cáncer testicular, meduloblastoma, cáncer vaginal, cáncer de ovario, cáncer cervical, carcinoma de células basales, adenoma de la pituitaria, rhabdomyosarcoma y sarcoma de Kaposi.

Los métodos en este Ejemplo se pueden aplicar en un ensayo que utiliza mezclas de células de mamífero intactas para detectar células cancerosas. Las células de carcinoma de pulmón de células no pequeñas CRL-5908 (ATCC) se utilizan como un modelo de cáncer y las células de Jurkat se utilizan como un sustituto de linfocitos primarios. CRL-5908 tiene una mutación puntual EGFR en L858R, y expresa EpCAM. Jurkat no expresa EpCAM (Landolin *et al.*, 2010). Se crean mezclas de células en seis relaciones CRL-5908:Jurkat de 0% y 1%. Las células se encapsulan a partir de las mezclas con perlas en una mezcla de lisis y después se fusionan con una corriente que contiene una mezcla de RT-PCR utilizando los métodos descritos anteriormente. Las células se diluyen de manera que la distribución de células sigue la estadística de Poisson con  $\lambda = 1,5$ , y ~44% de las gotitas con células que tienen múltiples células. Utilizando este método, se genera > 1 millón de gotitas en cada uno de seis experimentos de réplica para cada una de las mezclas de células. Una cámara de alta velocidad se utiliza para obtener tasas de encapsulación de perlas y células. Los amplicones principales se purifican por electroforesis en gel y se secuencian por secuenciación de siguiente generación para obtener al menos 10 millones de etiquetas de secuencia para cada uno de los bancos.

La detección de células cancerosas en estas mezclas de células requiere un marco analítico especial. La secuenciación genera recuentos de EGFR mutado y EpCAM enlazados a cada uno de los códigos de barras y los códigos de barras se rastrean de nuevo a las células. Si cada una de las gotitas contiene sólo una célula individual, a continuación, estos recuentos se utilizan para cuantificar directamente el porcentaje de CRL-5908 en la mezcla de células. Sin embargo, puede existir un cierto número arbitrario de células encapsuladas en gotitas de acuerdo con una distribución de Poisson, dando como resultado muchas gotitas con múltiples células.

Por lo tanto, para este tipo de análisis se utiliza un algoritmo que calcula el número de células cancerosas en una muestra dando recuentos de marcadores de cáncer, tales como EGFR mutado o EpCAM y estadísticas para la encapsulación de células Poisson  $\lambda$ . Para testar la validez de este algoritmo y para estimar los límites de detección que impone la encapsulación de múltiples células por gotita, se simula el proceso de encapsulación, y se determina la relación de la expresión de marcadores de cáncer en las células cancerosas con las células normales. Se supone una distribución de Poisson para la tasa de encapsulación de células, los niveles de expresión se distribuyen de forma normal logarítmica a lo largo de un fondo fijo, y la relación señal-ruido (SNR) se define como la relación entre el nivel de expresión medio al fondo medio. Esta simulación indica una tasa de error de < 1% en un escenario en donde ~44% de gotitas que contienen células tendrá múltiples células ( $\lambda = 1,5$ ) y SNR = 10.

**Ejemplo 16: Análisis No Invasivo de Expresión Génica en Glioblastoma Multiforme**

Determinados genes se co-expresan específicamente sólo en células tumorales circulantes, por lo que el enlace de dos transcripciones específicas de tumor en la misma célula es un método potencialmente poderoso para la detección de células tumorales circulantes en la sangre periférica o CSF. El método permite la estadificación molecular no invasiva de glioblastoma multiforme (GBM). GBM es el tipo más común de tumor cerebral maligno primario, con una incidencia de 16.000 nuevos casos por año en los Estados Unidos. Después de la caracterización por resonancia magnética (MRI) y el tratamiento clínico, la caracterización molecular de las biopsias se realiza a menudo para guiar a los regímenes de tratamiento. Existe un creciente consenso de que categorías moleculares distintas de tumores deben ser sometidas a distintos regímenes de tratamiento específicos (Mischel *et al.*, 2003 *Cáncer Biol Ther* 2:242-247). La investigación anterior del GBM ha indicado que un mal pronóstico es indicado mediante la co-expresión de los genes C/EBP $\beta$  y STAT3 (Carro *et al.*, 2010 *Nature* 463:318-26). Estas transcripciones no se co-expresan en los tejidos normales. Sin embargo, las biopsias de GBM son altamente invasivas y costosas, por lo que existe una demanda clínica de los métodos mínimamente invasivos para la estadificación molecular.

El método implica las siguientes etapas: (i) aislamiento de células mononucleares a partir de CSF (Spriggs 1954; *Journal of Clinical Pathology* 7:122) con la tecnología de microgotitas de emulsión; (ii) reacción en cadena de la polimerasa con transcriptasa inversa que fija como objetivo C/EBP $\beta$ , STAT3 y una secuencia de código de barras de enlazador único para cada una de las microgotitas; (iii) la amplificación de extensión por solapamiento de complejos entre C/EBP $\beta$ , STAT3 y la secuencia de enlazador; (iv) la recuperación de complejos enlazados por inversión de la emulsión; y (v) la cuantificación digital de los complejos de fusión utilizando secuenciación de siguiente generación. Sólo reactores de microgotitas que contienen células tumorales que co-expresan C/EBP $\beta$  y STAT3 producen grandes números de complejos enlazados completos. Aunque la secuenciación de siguiente generación agrupa todos los analitos de todas las células, las secuencias de códigos de barras de enlazador permiten el rastreo de la expresión génica de células individuales. El resultado final es la cuantificación digital de múltiples transcripciones enlazadas que se remontan a millones de células individuales analizadas en paralelo.

El método también proporciona la síntesis de ADNc y PCR en microgotitas de emulsión sin intercambio de tampón ni adición de reactivos entre las etapas moleculares. Se utilizan enzimas termoestables de la transcriptasa inversa (RT) que soportan temperaturas > 95°C, tales como ThermoScript RT (Lucigen) y GeneAmp Thermostable rTth (Life Technologies). Además de las regiones de cebador que fijan como objetivo C/EBP $\beta$  y STAT3 (FIGs. 18-25), tres de los cebadores en el conjunto incluyen secuencias de polinucleótidos que permiten la amplificación de un complejo de fusión: (i) el cebador 5' del locus C/EBP $\beta$  tiene una secuencia al azar de 10-20 nt sin complementariedad con cualquiera de los locus diana; (ii) el cebador 3' del locus C/EBP $\beta$  tiene una secuencia de 10-20 nt con complementariedad con el extremo 5' del oligonucleótido de código de barras de enlazador; (iii) la sonda 5' del locus STAT3 tiene complementariedad con el extremo 3' del enlazador. Otros dos oligonucleótidos actúan como cebadores directo e inverso de la PCR para amplificar específicamente el oligonucleótido de código de barras de enlazador. Los cebadores "internos" de los loci STAT3 y C/EBP $\beta$  (es decir, el cebador inverso para C/EBP $\beta$  y el cebador directo para STAT3) están a una concentración límite, es decir, 0,01  $\mu$ M para los cebadores internos y 0,1  $\mu$ M para todos los otros cebadores. Esto impulsa la amplificación del amplicón principal preferentemente frente a los amplicones menores.

Después de la inversión de la emulsión, los amplicones principales se someten a secuenciación a granel. El código de barras está enlazado a las secuencias de C/EBP $\beta$  y STAT3, y se utilizan para rastrear los amplicones principales a una célula individual (FIGs. 18-25). Con el rastreo de cada una de las secuencias a una célula individual original, es posible tabular datos genéticos para cada una de las células individuales, que permite entonces una cuantificación de la transcripción de células individuales, es decir, niveles de expresión de genes de células individuales que se traducen a un diagnóstico clínicamente procesable.

#### **Ejemplo 17: Cariotipado Molecular**

A menudo, los cambios cromosómicos estructurales tales como la pérdida de heterocigosidad (LOH) o la ganancia de cromosomas completos o segmentos de los mismos, conducirá a la progresión de un tumor (Parsons *et al.*, 2008 *Science* 321:1807-1812). Los médicos a menudo examinan el cariotipo de un tumor para formular un régimen de pronóstico y tratamiento. Los métodos arriba esbozados están adaptados para analizar tanto la expresión del gen como para detectar anomalías cromosómicas para cualquier tipo de tumor en una única reacción multiplexada.

Una secuencia de cáncer mutante está enlazada a sondas para determinar el número de copias de cromosomas o aberraciones cromosómicas estructurales. Un método de este tipo implica las siguientes etapas: (i) el aislamiento de células mononucleares individuales de sangre en microgotitas de emulsión; (ii) la amplificación de sondas cromosómicas y loci de "enlazador" de mutaciones del cáncer por un protocolo de LCR/PCR modificado en reactores de microgotitas de emulsión; (iii) la amplificación de extensión por solapamiento de complejos entre sondas cromosómicas y loci de enlazador mutantes en reactores de microgotitas; (iv) la recuperación de los complejos enlazados por inversión de la emulsión; y (v) la secuenciación masiva en paralelo.

Los datos de secuenciación masiva en paralelo se analizan para cuantificar los casos de genotipos enlazados. Solamente los reactores de microgotitas que contienen células con mutaciones del cáncer proporcionan complejos enlazados entre las sondas cromosómicas y la secuencia específica del cáncer. Las sondas cromosómicas se utilizan para cuantificar el número de cromosomas o segmentos de cromosomas presentes en las células cancerosas circulantes y, por asociación, el tumor. El número de copias de cromosomas se cuantifica mediante la comparación de recuentos de secuencias de un cromosoma desconocido con los recuentos de secuencias de un cromosoma de referencia conocida dentro de un único experimento, o mediante la búsqueda de un desequilibrio alélico (Johnson *et al.*, 2010 Human Reproduction 25:1066-75). Este método también se utiliza para detectar una diversidad de trastornos cromosómicos, incluyendo la aneuploidía, trastornos cromosómicos estructurales desequilibrados, microdeleciones, microinserciones y otros tipos de trastornos congénitos. Las sondas cromosómicas están ligadas a una secuencia de código de barras en lugar de una mutación del cáncer, de manera que la secuenciación cromosómica masiva en paralelo mide los trastornos cromosómicos en todas las células en el ensayo en lugar de sólo las células que albergan una mutación particular.

#### **Ejemplo 18: Mutaciones de Células Somáticas**

A menudo, las mutaciones de células somáticas, es decir, en los genes promotores de tumores tales como *p53*, *p16* y/o *EGFR*, contribuyen a la progresión del cáncer (Parsons *et al.*, 2008 *Science* 321:1807-1812). Los médicos a menudo analizan los tumores en cuanto a mutaciones de células somáticas conocidas para formular un régimen de pronóstico y tratamiento. En particular, las mutaciones de células somáticas son a menudo indicativas de la progresión a etapas más agresivas de un tumor. Los métodos descritos anteriormente están adaptados para analizar la expresión de genes, las mutaciones de células somáticas y/o cambios cromosómicos para cualquier tipo de tumor en las reacciones de microgotitas de emulsión multiplexadas en millones de células individuales en paralelo. Si se conocen las mutaciones de células somáticas, no es necesario un código de barras molecular, porque la LCR específica para el alelo o sondas-candado se utilizan para amplificar específicamente los amplicones principales sólo en las células que albergan la mutación de células somáticas.

Cualquier combinación de expresión génica, cariotipado molecular y análisis de la mutación de células somáticas se lleva a cabo en las células tumorales individuales en poblaciones de células heterogéneas. Por ejemplo, la LCR o sondas-candado se utilizan para afectar a la captura del locus específica para el alelo y la amplificación del amplicón principal sólo en células con una mutación de células somáticas particular. Este método es una alternativa al método de código de barras molecular descrito anteriormente, al menos en la sección B.6), consiguientemente un análisis genético específico para las células tumorales en un fondo mixto altamente heterogéneo de células. La amplificación de mutación de células somáticas específica para el alelo está ligada a las transcripciones de ARN asociadas a los resultados de la enfermedad y/o sondas para la cuantificación de la pérdida de heterocigosidad (LOH) o duplicaciones regionales en el cromosoma. El método se utiliza para analizar la co-expresión de dos o más secuencias de microARN en las células individuales, o la co-expresión de un microARN con otra transcripción, una secuencia de ADN metilado o la mutación de células somáticas.

#### **Ejemplo 19: Métodos de Análisis de Poblaciones de Células Quiméricas**

Determinadas aplicaciones requieren un análisis multiplexado de poblaciones de células que son quimeras entre dos organismos. Por ejemplo, después de un trasplante de células madre hematopoyéticas (HSC), las células T y B del huésped son quiméricas entre el huésped y el injerto. La amplificación por PCR en una población de células quiméricas de un locus genético variable combinada con algún tipo de locus genético funcional, tal como una transcripción de ARN, permite el análisis del locus genético funcional de una manera específica para el individuo.

Los métodos descritos en esta memoria se aplican al trasplante de células madre hematopoyéticas no mieloablativas. Los médicos carecen de herramientas poderosas para el seguimiento de los pacientes después de los trasplantes alogénicos no mieloablativos de células madre hematopoyéticas (HSC) (Pollack *et al.*, 2009 *American Journal of Clinical Oncology*. 32:618-28). Después del trasplante no mieloablativo, el sistema inmune del huésped es una quimera entre el huésped y las células T del injerto. La quimera es un tejido mal caracterizado, y la inestabilidad quimérica se asocia con un pronóstico deficiente. El equilibrio donante-receptor de reconstitución inmune parece influir en un cierto número de resultados inmunológicos de trasplante, incluido el efecto de injerto frente a tumor (GVT), la enfermedad de injerto frente a huésped (GVHD) y la susceptibilidad a la infección. Las células T parecen jugar un papel importante en la mediación de cada uno de estos procesos a través de la inmunidad adaptativa y reconocimiento de antígeno del receptor de células T (TCR). Actualmente, los médicos carecen de herramientas para el seguimiento de la identidad de las células T del huésped y el injerto después de trasplante de HSC. Tales métodos se utilizan para supervisar directamente GVT, GVHD y la respuesta a las infecciones (Kristt *et al.*, 2007 *Bone Marrow Transplantation* 39: 255-68).

Un método se utiliza para el seguimiento de las poblaciones de células T quiméricas. Para el análisis de quimerismo de células T, TCR $\beta$  y polimorfismos de un solo nucleótido específicos para huésped e injerto (SNPs) están enlazados por PCR de extensión por solapamiento o RT-PCR de extensión por solapamiento en microgotitas de células individuales. Este método implica las siguientes etapas: (i) determinación del genotipo para encontrar SNPs

específicos para el injerto y el huésped; (ii) el aislamiento post-trasplante de células individuales a partir de sangre del huésped en microgotitas de emulsión; (iii) la PCR de extensión por solapamiento de complejos de fusión entre SNPs y TCR $\beta$  en reactores de microgotitas; y (iv) la recuperación y secuenciación de los complejos de enlace SNP-TCR $\beta$  por la inversión de la emulsión. El resultado es un banco de secuencias de TCR $\beta$  con enlace al huésped o injerto. Las secuencias de TCR $\beta$  se correlacionan con los resultados clínicos a lo largo del tiempo.

Tipos similares de análisis se llevan a cabo utilizando LCR, circularización multi-sonda o sondas-candado, o cualquier combinación de los mismos. También, otros tipos de secuencias variantes tales como los STRs, también son útiles para indicar la fuente de células en una población quimérica de las células. El análisis de quimerismo de células T es adaptable a aplicaciones tales como el análisis de células B o cualquier otra sub-población de células mononucleares en sangre. Adicionalmente, el método se puede combinar con la secuenciación de células T funcionales para indicar la actividad inmune de clones particulares de células T.

Existen muchas aplicaciones para el análisis de la población de células quiméricas fuera del campo de la medicina. Por ejemplo, un investigador puede crear organismos quiméricos tales como moscas de la fruta, ratones o ratas, que son quimeras entre múltiples individuos con diferentes antecedentes genéticos, o incluso entre múltiples especies. Poblaciones de células quiméricas para las transcripciones de ARN, metilación de ADN, mutaciones de células somáticas, presencia de un gen recombinante o una región de ADN variable también son susceptibles de análisis con este método. Por lo tanto, métodos para el análisis de poblaciones de células T y B quiméricas están adaptados a otros organismos y otros tipos de poblaciones de células. Adicionalmente, estos métodos se utilizan para la terapia celular alogénica o autóloga. Actualmente los médicos carecen de herramientas poderosas para el seguimiento de los pacientes después de haber introducido las células inmunes ya sea de un donante o como se han recogido previamente del paciente. Células T, células B o células NK son monitoreadas para establecer las características y la eficacia de la terapia.

#### **Ejemplo 20: Métodos para el Análisis de la Secuencia Reguladora de Genes**

Variantes de ADN regulador tienen un impacto en los niveles de expresión de las transcripciones de ARN (Brown *et al.*, 2007 Science 317:1557-60). Rastros funcionales de variantes de regulación requieren mucho tiempo y son costosas. En un método a modo de ejemplo, el método incluye mutagenizar células, capturar células individuales en microgotitas acuosas-en-aceite y luego fusionar un locus regulador supuesto amplificado con transcripciones de ARN del gen cercano. De esta manera, las mutaciones en las secuencias reguladoras podrían enlazarse con los niveles de expresión génica.

A menudo, un investigador desea entender cómo las secuencias de ácido nucleico genómico afectan a la expresión de las transcripciones. Debido a que muchos nucleótidos afectan a la expresión del gen, estos experimentos son tediosos. Idealmente, un investigador quiere analizar la expresión génica cuantitativa al nivel de células individuales en función de las secuencias reguladoras mutagenizadas. Las secuencias reguladoras sospechosas se mutagenizan para crear un banco de secuencias reguladoras variables. Entonces, se utiliza una combinación de PCR de extensión por solapamiento y de RT-PCR de extensión por solapamiento en microgotitas de emulsión de células individuales para enlazar la secuencia de ADN reguladora a niveles de transcripción de ARN. De esta manera, el efecto de la mutagénesis de la secuencia reguladora en los niveles de transcripción de ARN se mide en células individuales.

#### **Ejemplo 21: Métodos para el Haplotipado Molecular**

Muchos tipos de análisis genéticos tales como PGD o estudios de asociación de genoma completo se benefician del enlace del haplotipo de varios loci genéticos en el ADN derivado de una célula de esperma individual. En un método a modo de ejemplo, células de esperma individuales son capturadas en microgotitas acuosas-en-aceite, y luego varios loci genéticos variables se fusionan en las células, tales como SNPs o STRs. Esto permite un ajuste de fase molecular masivamente paralelo.

Se proporciona un método para el ajuste de fase de dos loci. Millones de haplotipos de esperma individual se analizan en paralelo. El método implica las siguientes etapas: (i) el aislamiento de las células de esperma individuales que utilizan la tecnología de microgotitas de emulsión; (ii) amplificación de dos variantes genéticas por PCR en reactores de microgotitas; (iii) la amplificación por PCR de extensión por solapamiento de complejos de fusión entre las variantes en los reactores de microgotitas; y (iv) la recuperación de complejos enlazados por inversión de la emulsión. El resultado es un banco de haplotipos ajustados en fase, que luego se secuencian utilizando la secuenciación de siguiente generación.

Un método alternativo para el ajuste de fase alterno de múltiples loci se proporciona en este párrafo. En algunos casos, el ajuste de fases de sólo dos loci no es adecuado para la mejora de los estudios de asociación de todo el genoma u otros tipos de análisis. En tales situaciones, se utilizan métodos moleculares como la LCR o sondas-candado, que permiten una multiplexación mayor de la sonda, como una alternativa. Adicionalmente, una diversidad

de pares de cebadores de PCR se fijan a las perlas, de manera que miles de pares de cebadores de la PCR se distribuyen a microgotitas de emulsión que contienen una única perla y una célula de esperma individual.

**Ejemplo 22: Métodos de Detección de la Evolución Molecular Dirigida**

5 Algunos tipos de aplicaciones industriales requieren enzimas y/o cepas biológicas para optimizar biosistemas modificados por ingeniería. Por ejemplo, las enzimas que degradan un tipo particular de residuos industriales no se pueden encontrar en la naturaleza, pero la evolución *in vitro* de enzimas existentes podría resultar en una enzima optimizada. Muchos de tales procesos se benefician de análisis genético molecular de múltiples loci en millones de células individuales analizadas en paralelo.

Evolución de Levaduras

10 La evolución industrial *in vitro* a menudo implica la mutagénesis de células seguida por el crecimiento en medios selectivos. En un método a modo de ejemplo, células de levadura se someten a mutagénesis y se cultivan en medios especiales que contienen xilosa como fuente de alimentación primaria. Las células de levadura individuales son capturadas en microgotitas acuosas-en-aceite, y luego se secuencian varios genes de la ruta metabólica. Al menos una compañía (Microbiogen, Sydney, AUS) está desarrollando cepas de levadura para el crecimiento en xilosa, pero está utilizando métodos de rastreo lentos y tradicionales.

Otras aplicaciones

20 Muchos grupos están investigando actualmente métodos para mejorar las cepas naturales de algas y bacterias con el fin de la producción de biocombustibles. Los métodos para la determinación del genotipo enlazado y/o el análisis de la expresión génica de células individuales se utilizan para activar la evolución *in vitro* de estos organismos con el fin de la producción de biocombustibles u otros tipos de energía.

**Ejemplo 23: Otras Aplicaciones y Métodos Derivados**

Agricultura

25 Todos los métodos clínicos descritos anteriormente (p. ej., secuenciación de células T y secuenciación de células B) son aplicables a los animales. Estos animales incluyen, pero no se limitan a vacas, cerdos, pollos o salmón, etc. En particular, el ganado y otros animales agrícolas padecen enfermedades infecciosas que resultan en considerables dificultades económicas. Los métodos descritos en esta memoria son adaptables para mejorar el seguimiento y la detección de enfermedades infecciosas en un entorno agrícola.

Metagenómica

30 La metagenómica es un método de estudio de la diversidad genética en los ecosistemas en los que muestras ambientales se secuencian directamente. En un método a modo de ejemplo, células tales como algas en muestras del medio ambiente tales como el agua de mar se separan en microgotitas de emulsión de células individuales, y después se analizan para al menos dos loci genéticos. Por ejemplo, un investigador puede estar interesado en encontrar una especie particular de alga que expresa una forma particular de clorofila y pertenece a una especie de alga particular. El genotipado por LCR se utiliza para amplificar amplicones principales sólo de células de algas de una especie particular que alberga esa forma particular de clorofila. Un experto en la técnica también puede apreciar que un método de este tipo también es útil para probar la diversidad de clorofila en una clase, especie o género particular de algas mediante el enlace de LCR específica de la especie o sondas-candado con la amplificación por PCR de exones de clorofila. Las algas y la clorofila son sólo un ejemplo específico; las células son de cualquier especie, y hay muchos tipos de loci genéticos diana, incluidas las transcripciones de ARN, variantes genómicas y ADN mitocondrial.

Detección de la metilación del ADN

45 La metilación del ADN es un tipo de modificador epigenético que ayuda a las células a controlar la transcripción de ARN y otros procesos celulares (Brunner *et al.*, 2009 *Genome Research* 19: 1044-56). Por ejemplo, los linfocitos de la sangre pueden sufrir metilación del ADN aberrante, conduciendo a tumores líquidos. Los métodos descritos anteriormente son útiles para analizar la metilación del ADN en células individuales (p. ej., loci de metilación del ADN múltiple en células individuales, o al menos un locus de metilación del ADN con una diana de transcripción de ARN o diana de la secuencia de ADN). La metilación del ADN se analiza mediante enzimas de restricción específicas para la metilación, la conversión de bisulfito o precipitación con anti-metilcitosina. La mayor parte de estos análisis requerirían múltiples entradas de tampones de reacción si se utiliza un chip de microfluidos para crear microgotitas de emulsión. Por ejemplo, la realización de la conversión de bisulfito requiere un tampón que no es apropiado para 50 la PCR, LCR, RT-PCR o sondas-candado. En un método a modo de ejemplo, las células individuales se encapsulan en microgotitas de emulsión utilizando un tampón de conversión de bisulfito estándar. Luego, después de la conversión de bisulfito, las microgotitas se fusionan con un segundo tampón acuoso. Este segundo tampón diluye el

tampón de conversión de bisulfito, permitiendo PCR, LCR, RT-PCR o métodos de sonda-candado. Enfoques similares son útiles para anti-metilcitosina o digestión de restricción sensible a la metilación.

#### Inmunoprecipitación de la cromatina

- 5 La inmunoprecipitación de la cromatina es un método en el que el ADN se reticula a las proteínas en los núcleos celulares (Johnson *et al.*, 2007 *Science* 316:1497-502). Un anticuerpo dirigido contra una proteína de unión de ADN de interés se utiliza entonces para precipitar específicamente los complejos de ADN-proteína, y luego el ADN se secuencia o analiza con un chip de ADN. Los métodos de enlace molecular descritos anteriormente se utilizan para analizar múltiples loci de unión a ADN-proteína en células individuales, o al menos un locus de unión ADN-proteína con una diana de transcripción de ARN o diana de secuencia de ADN. La mayor parte de estos análisis requieren múltiples entradas de tampones de reacción si se utiliza un chip de microfluidos para crear microgotitas de emulsión.
- 10 Por ejemplo, la realización de inmunoprecipitación de la cromatina requiere un tampón que no es apropiado para la PCR, LCR, RT-PCR o sondas-candado. En un método a modo de ejemplo, las células individuales se encapsulan en microgotitas de emulsión utilizando un tampón de inmunoprecipitación estándar. Luego, después de la precipitación, las microgotitas se fusionan con un segundo tampón acuoso. Este segundo tampón diluye el tampón de precipitación, permitiendo PCR, LCR, RT-PCR, o métodos de sonda-candado.
- 15

**TABLA 2: LISTADO INFORMAL DE SECUENCIAS**

SEQ ID N°	NOMBRE	REGIÓN DE FIJACIÓN COMO OBJETIVO DEL GENOMA	SECUENCIA DE SOLAPAMIENTO	DESCRIPCIÓN
SEQ ID NO: 1	TRBV2.F	TCAAATTTCACTCTGAAGATCCGGTCCACAA		directo externo
SEQ ID NO: 2	TRBV3-1.F	GCTCACTTAAATCTTCACATCAATTCCCTGG		directo externo
SEQ ID NO: 3	TRBV4-1.F	CTTAAACCTTCACCTACACGCCCTGC		directo externo
SEQ ID NO: 4	TRBV4-2_4-3.F	CTTATTCCTTCACCTACACACCCTGC		directo externo
SEQ ID NO: 5	TRBV5-1.F	GCTCTGAGATGAATGTGAGCACCTTG		directo externo
SEQ ID NO: 6	TRBV5-3.F	GCTCTGAGATGAATGTGAGTGCCTTG		directo externo
SEQ ID NO: 7	TRBV 5-4_5-5_5-6_5-7_5-8.F	GCTCTGAGCTGAATGTGAACGCCTTG		directo externo
SEQ ID NO: 8	TRBV6-1.F	TCGCTCAGGCTGGAGTCGGCTG		directo externo
SEQ ID NO: 9	TRBV6-2_6-3.F	GCTGGGGTTGGAGTCGGCTG		directo externo
SEQ ID NO: 10	TRBV6-4.F	CCCTCACGTTGGCGTCTGCTG		directo externo
SEQ ID NO: 11	TRBV6-5.F	GCTCAGGCTGCTGTCCGGCTG		directo externo
SEQ ID NO: 12	TRBV6-6.F	CGCTCAGGCTGGAGTTGGCTG		directo externo
SEQ ID NO: 13	TRBV6-7.F	CCCCTCAAGCTGGAGTCAGCTG		directo externo
SEQ ID NO: 14	TRBV6-8.F	CACTCAGGCTGGTGTCCGGCTG		directo externo
SEQ ID NO: 15	TRBV6-9.F	CGCTCAGGCTGGAGTCAGCTG		directo externo
SEQ ID NO: 16	TRBV7-1.F	CCACTCTGAAGTTCCAGCGCACAC		directo externo
SEQ ID NO: 17	TRBV7-2.F	CACTCTGACGATCCAGCGCACAC		directo externo
SEQ ID NO: 18	TRBV7-3.F	CTCTACTCTGAAGATCCAGCGCACAG		directo externo
SEQ ID NO: 19	TRBV7-4.F	CCACTCTGAAGATCCAGCGCACAG		directo externo
SEQ ID NO: 20	TRBV7-6.F	CACTCTGACGATCCAGCGCACAG		directo externo
SEQ ID NO: 21	TRBV7-7.F	CCACTCTGACGATCCAGCGCACAG		directo externo
SEQ ID NO: 22	TRBV7-8.F	CCACTCTGAAGATCCAGCGCACAC		directo externo
SEQ ID NO: 23	TRBV7-9.F	CACCTTGGAGATCCAGCGCACAG		directo externo

ES 2 615 733 T3

SEQ ID NO: 24	TRBV9.F	GCACTCTGAACTAAACCTGAGCTCTCTG		directo externo
SEQ ID NO: 25	TRBV10-1.F	CCCCTCACTCTGGAGTCTGCTG		directo externo
SEQ ID NO: 26	TRBV10-2.F	CCCCCTCACTCTGGAGTCAGCTA		directo externo
SEQ ID NO: 27	TRBV10-3.F	CCTCCTCACTCTGGAGTCCGCTA		directo externo
SEQ ID NO: 28	TRBV11- 1_11-3.F	CCACTCTCAAGATCCAGCCTGCAG		directo externo
SEQ ID NO: 29	TRBV11-2.F	CTCCACTCTCAAGATCCAGCCTGCAA		directo externo
SEQ ID NO: 30	TRBV12- 3_12-4_12- 5.F	CCACTCTGAAGATCCAGCCCTCAG		directo externo
SEQ ID NO: 31	TRBV13.F	CATTCTGAACTGAACATGAGCTCCTTGG		directo externo
SEQ ID NO: 32	TRBV14.F	CTACTCTGAAGGTGCAGCCTGCAG		directo externo
SEQ ID NO: 33	TRBV15.F	GATAACTTCCAATCCAGGAGGCCGAACA		directo externo
SEQ ID NO: 34	TRBV16.F	CTGTAGCCTTGAGATCCAGGCTACGA		directo externo
SEQ ID NO: 35	TRBV17.F	CTTCCACGCTGAAGATCCATCCCG		directo externo
SEQ ID NO: 36	TRBV 18.F	GCATCCTGAGGATCCAGCAGGTAG		directo externo
SEQ ID NO: 37	TRBV19.F	CCTCTCACTGTGACATCGGCC		directo externo
SEQ ID NO: 38	TRBV20-1.F	CTTGTCCACTCTGACAGTGACCAGTG		directo externo
SEQ ID NO: 39	TRBV23-1.F	CAGCCTGGCAATCCTGTCCTCAG		directo externo
SEQ ID NO: 40	TRBV24-1.F	CTCCCTGTCCCTAGAGTCTGCCAT		directo externo
SEQ ID NO: 41	TRBV25-1.F	CCCTGACCCTGGAGTCTGCCA		directo externo
SEQ ID NO: 42	TRBV27.F	CCCTGATCCTGGAGTCGCCA		directo externo
SEQ ID NO: 43	TRBV28.F	CTCCCTGATTCTGGAGTCCGCCA		directo externo
SEQ ID NO: 44	TRBV29-1.F	CTAACATTCTCAACTCTGACTGTGAGCAACA		directo externo
SEQ ID NO: 45	TRBV30.F	CGGCAGTTCATCCTGAGTTCTAAGAAGC		directo externo
SEQ ID NO: 46	TRBJ1-1.R	TTACCTACAACCTGTGAGTCTGGTGCCTTGTC CAAA	GCTCATCTGGC ATAATTCTCCT	inverso interno
SEQ ID NO: 47	TRBJ1-2.R	ACCTACAACGGTTAACCTGGTCCCCGAACC GAA	GCTCATCTGGC ATAATTCTCCT	inverso interno
SEQ ID NO: 48	TRBJ1-3.R	ACCTACAACAGTGAGCCAACCTCCCTCTCCA AA	GCTCATCTGGC ATAATTCTCCT	inverso interno
SEQ ID NO: 49	TRBJ1-4.R	CCAAGACAGAGAGCTGGGTTCCACTGCCAA A	GCTCATCTGGC ATAATTCTCCT	inverso interno
SEQ ID NO: 50	TRBJ1-6.R	CTGTCACAGTGAGCCTGGTCCCGTTCCCAA A	GCTCATCTGGC ATAATTCTCCT	inverso interno
SEQ ID NO: 51	TRBJ2-1.R	CGGTGAGCCGTGTCCCTGGCCCGAA	GCTCATCTGGC ATAATTCTCCT	inverso interno
SEQ ID NO: 52	TRBJ2-2.R	CCAGTACGGTCAGCCTAGAGCCTTCTCCAA A	GCTCATCTGGC ATAATTCTCCT	inverso interno
SEQ ID NO: 53	TRBJ2-3.R	ACTGTCAGCCGGGTGCCTGGGCCAAA	GCTCATCTGGC ATAATTCTCCT	inverso interno
SEQ ID NO:	TRBJ2-4.R	AGAGCCGGGTCCCGGCCCGAA	GCTCATCTGGC	inverso interno

ES 2 615 733 T3

54			ATAATTCTCCT	
SEQ ID NO: 55	TRBJ2-5.R	GGAGCCGCGTGCCTGGCCCCGAA	GTCATCTGGC ATAATTCTCCT	inverso interno
SEQ ID NO: 56	TRBJ2-6.R	GTCAGCCTGCTGCCGGCCCCGAA	GTCATCTGGC ATAATTCTCCT	inverso interno
SEQ ID NO: 57	TRBJ2-7.R	GTGAGCCTGGTGCCCGCCCCGAA	GTCATCTGGC ATAATTCTCCT	inverso interno
SEQ ID NO: 58	IL2.F	TCACCAGGATGCTCACATTTAAGT	AGGAGAATTAT GCCAGATGAGC	directo interno
SEQ ID NO: 59	IL2.F	GAGGTTTGAGTTCTTCTTAGACTGA		inverso externo
SEQ ID NO: 60	IL4.F	CCACGGACACAAGTGCATA	AGGAGAATTAT GCCAGATGAGC	directo interno
SEQ ID NO: 61	IL4.R	CCCTGCAGAAGGTTTCCTTCT		inverso externo
SEQ ID NO: 62	INFG.F	TCAGCTCTGCATCGTTTTGG	AGGAGAATTAT GCCAGATGAGC	directo interno
SEQ ID NO: 63	INFG.R	GTTCCATTATCCGCTACATCTGAA		inverso externo
SEQ ID NO: 64	TNFA.F	GCCCAGGCAGTCAGATCATC	AGGAGAATTAT GCCAGATGAGC	directo interno
SEQ ID NO: 65	TNFA.R	GGGTTTGCTACAACATGGGCT		inverso externo
SEQ ID NO: 66	FOXP3.F	AACAGCACATTCCCAGAGTTCCT	AGGAGAATTAT GCCAGATGAGC	directo interno
SEQ ID NO: 67	FOXP3.R	CATTGAGTGTCGCTGCTTCT		inverso externo
SEQ ID NO: 68	TBX21.F	GTCCAACAATGTGACCCAGAT	AGGAGAATTAT GCCAGATGAGC	directo interno
SEQ ID NO: 69	TBX21.R	GCTGGTACTTATGGAGGGACTG		inverso externo
SEQ ID NO: 70	TBX21.F	AGCTGACTCACGCCGTCC	AGGAGAATTAT GCCAGATGAGC	directo interno
SEQ ID NO: 71	TBX21.F	CACAGAAACCCTCGACAAGCC		inverso externo
SEQ ID NO: 72	IL2.F	CTGGAATAGCCAATACTGATTACCTG	AGGAGAATTAT GCCAGATGAGC	directo interno
SEQ ID NO: 73	IL2.R	CATGAATTTTATACCTTAGGAGACGG		inverso externo

**REIVINDICACIONES**

1. Un método para analizar al menos dos secuencias de ácido nucleico en una célula individual contenida dentro de una población de al menos 10.000 células, en el que la célula individual se aísla en una microgotita de emulsión, comprendiendo dicho método:
- 5 proporcionar un primer conjunto de sondas de ácido nucleico, comprendiendo el primer conjunto una primera sonda que comprende una secuencia que es complementaria a una primera sub-secuencia de ácido nucleico diana, comprendiendo una segunda sonda una secuencia que es complementaria a una segunda sub-secuencia del primer ácido nucleico diana y una segunda secuencia que es complementaria a una secuencia exógena, comprendiendo una tercera sonda la secuencia exógena y una secuencia que es complementaria a una primera sub-secuencia de un segundo ácido nucleico diana, y una cuarta sonda que comprende una secuencia que es complementaria a una segunda sub-secuencia de la segunda secuencia de ácido nucleico diana, en donde el primer ácido nucleico diana o el segundo ácido nucleico diana comprende una secuencia de receptor de células T o una secuencia de inmunoglobulina;
- aislar las células individuales con al menos un conjunto de sondas de ácido nucleico;
- 15 amplificar la primera y segunda secuencias de ácido nucleico diana de manera independiente, en donde la primera secuencia de ácido nucleico diana se amplifica utilizando la primera sonda y la segunda sonda, y en donde la segunda secuencia de ácido nucleico diana se amplifica utilizando la tercera sonda y la cuarta sonda;
- hibridar la secuencia exógena a su complemento;
- 20 amplificar la primera secuencia de ácido nucleico diana, la segunda secuencia de ácido nucleico diana y la secuencia exógena, utilizando las primera y cuarta sondas, generando con ello un complejo fusionado; y
- realizar una reacción de secuenciación a granel para generar información de la secuencia para al menos 100.000 complejos fusionados a partir de al menos 10.000 células dentro de la población de células, en donde la información de la secuencia es suficiente para co-localizar la primera secuencia de ácido nucleico diana y la segunda secuencia de ácido nucleico diana en una célula individual de la población de al menos 10.000 células.
- 25 2. El método de la reivindicación 1, en el que la etapa de amplificación comprende llevar a cabo
- (a) una reacción en cadena de la polimerasa y en donde la primera y tercera sondas son cebadores directos y la segunda y cuarta sondas son cebadores inversos para la reacción en cadena de la polimerasa;
- (b) una reacción en cadena de la ligasa; o
- (c) una reacción en cadena de la polimerasa, una reacción en cadena de la polimerasa con transcriptasa inversa o una reacción en cadena de la ligasa, seguida de una reacción en cadena de la polimerasa.
- 30 3. El método de la reivindicación 1 ó 2, en el que el complejo fusionado es circular.
4. El método según una de las reivindicaciones 1-3, en el que la primera o la segunda secuencia de ácido nucleico diana es una secuencia de ARN o una secuencia de ADN.
5. El método según una de las reivindicaciones 1-4, en el que
- 35 (a) la primera o segunda secuencia de ácido nucleico diana comprende una secuencia del receptor de células T, en donde preferiblemente la primera secuencia de ácido nucleico diana, la segunda secuencia de ácido nucleico diana o las dos secuencias de ácido nucleico diana comprenden una secuencia de inmunoglobulina;
- (b) el primer ácido nucleico diana comprende una secuencia del receptor de células T y la segunda secuencia de ácido nucleico diana comprende una segunda molécula que está asociada con la función de células inmunes;
- 40 (c) la primera secuencia de ácido nucleico diana comprende una secuencia de inmunoglobulina, y la segunda secuencia comprende una segunda molécula asociada con la función de células inmunes; o
- (d) el primero o segundo ácido nucleico diana comprende una secuencia de genes rara.
- 45 6. El método de la reivindicación 5(b) en el que la segunda molécula se selecciona del grupo que consiste en: interleucina-2 (IL-2), interleucina-4 (IL-4), interferón gamma (IFN $\gamma$ ), interleucina-10 (IL-10), interleucina-1 (IL-1), interleucina-13 (IL-13), interleucina-17 (IL-17), interleucina-18 (IL-18), factor de necrosis tumoral alfa (TNF $\alpha$ ), factor de necrosis tumoral beta (TNF $\beta$ ), factor de transcripción 21 de la caja T (TBX21), caja forkhead P3 (FOXP3), cúmulo

de diferenciación 4 (CD4), cúmulo de diferenciación 8 (CD8), cúmulo de diferenciación 1d (CD1d), cúmulo de diferenciación 161 (CD161), cúmulo de diferenciación 3 (CD3), complejo principal de histocompatibilidad (MHC), cúmulo de diferenciación 19 (CD19), receptor de interleucina 7 (receptor de IL-7), cúmulo de diferenciación 10 (CD10), cúmulo de diferenciación 20 (CD20), cúmulo de diferenciación 22 (CD22), cúmulo de diferenciación 34 (CD34), cúmulo de diferenciación 27 (CD27), cúmulo de diferenciación 5 (CD5) y cúmulo de diferenciación 45 (CD45), cúmulo de diferenciación 38 (CD38), cúmulo de diferenciación 78 (CD78), receptor de interleucina-6, factor regulador de interferón 4 (IRF4) y cúmulo de diferenciación 138 (CD138).

7. El método de la reivindicación 5(c), en el que la segunda molécula se selecciona del grupo que consiste en: interleucina-2 (IL-2), interleucina-4 (IL-4), interferón gamma (IFN $\gamma$ ), interleucina-10 (IL-10), interleucina-1 (IL-1), interleucina-13 (IL-13), interleucina-17 (IL-17), interleucina-18 (IL-18), factor de necrosis tumoral alfa (TNF $\alpha$ ), factor de necrosis tumoral beta (TNF $\beta$ ), factor de transcripción 21 de la caja T (TBX21), caja forkhead P3 (FOXP3), cúmulo de diferenciación 4 (CD4), cúmulo de diferenciación 8 (CD8), cúmulo de diferenciación 1d (CD1d), cúmulo de diferenciación 161 (CD161), cúmulo de diferenciación 3 (CD3), complejo principal de histocompatibilidad (MHC), cúmulo de diferenciación 19 (CD19), receptor de interleucina 7 (receptor de IL-7), cúmulo de diferenciación 10 (CD10), cúmulo de diferenciación 20 (CD20), cúmulo de diferenciación 22 (CD22), cúmulo de diferenciación 34 (CD34), cúmulo de diferenciación 27 (CD27), cúmulo de diferenciación 5 (CD5) y cúmulo de diferenciación 45 (CD45), cúmulo de diferenciación 38 (CD38), cúmulo de diferenciación 78 (CD78), receptor de interleucina-6, factor regulador de interferón 4 (IRF4) y cúmulo de diferenciación 138 (CD138).

8. El método de la reivindicación 5(d), en el que

(a) la secuencia de genes rara está presente en menos de 5% de las células, preferiblemente en menos de 1% de las células, y más preferiblemente en menos de 0,1% de las células; o

(b) la secuencia de genes rara resulta de una mutación genética, en donde preferiblemente

(ba) la mutación es una mutación somática;

(bb) la mutación está asociada con una enfermedad; o

(bc) la mutación genética es una mutación en un gen seleccionado del grupo que consiste en el receptor del factor de crecimiento epidermal (EGFR), homólogo de fosfatasa y tensina (PTEN), proteína de tumor 53 (p53), MutS homólogo 2 (MSH2), neoplasia endocrina múltiple 1 (MEN1), poliposis adenomatosa coli (APC), receptor Fas (FASR), proteína retinoblastoma (Rb1), Janus quinasa 2 (JAK2), factor de transcripción 1 similar a (ETS) (ELK1), virus de la eritroblastosis aviar v-ets E26 homólogo de oncogén 1 (ETS1), cáncer de mama 1 (BRCA1), cáncer de mama 2 (BRCA2), receptor del factor de crecimiento de hepatocitos (MET), proto-oncogén ret (RET), leucemia eritroblástica viral V-erb-b2 homólogo de oncogén 2 (HER2), sarcoma de rata viral V-Ki-ras2 Kirsten homólogo de oncogén (KRAS), linfoma de células B 2 (BCL2), mielocitomatosis viral V-myc homólogo de oncogén (MYC), gen de neurofibromatosis de tipo 2 (NF2), mieloblastosis viral v-myb homólogo de oncogén (MYB) y homólogo mutS 6 (E. coli) (MSH6);

siendo la enfermedad preferiblemente cáncer, seleccionándose dicho cáncer preferiblemente del grupo que consiste en carcinoma de pulmón, cáncer de pulmón de células no pequeñas, cáncer de pulmón de células pequeñas, cáncer de útero, cáncer de tiroides, carcinoma de mama, carcinoma de próstata, carcinoma de páncreas, carcinoma de colon, linfoma, linfoma de Burkitt, linfoma de Hodgkin, leucemia mieloide, leucemia, sarcoma, blastoma, melanoma, seminoma, cáncer de cerebro, glioma, glioblastoma, astrocitoma cerebeloso, linfoma de células T cutáneo, cáncer gástrico, cáncer de hígado, ependimoma, cáncer de laringe, cáncer de cuello, cáncer de estómago, cáncer de riñón, cáncer de páncreas, cáncer de vejiga, cáncer de esófago, cáncer testicular, meduloblastoma, cáncer vaginal, cáncer de ovario, cáncer de cuello de útero, carcinoma de células basales, adenoma de la pituitaria,

rabdomiosarcoma o sarcoma de Kaposi.

9. El método según una de las reivindicaciones 1-8, que comprende, además, fijar y permeabilizar las células antes de realizar la etapa de amplificación.

10. El método según una de las reivindicaciones 1-9, que comprende, además, lisar las células antes de realizar la etapa de amplificación.

11. El método según una de las reivindicaciones 1-10, que comprende, además, cuantificar la información de la secuencia generada a partir de la reacción de secuenciación a granel.

12. El método según una de las reivindicaciones 1-11, en el que la realización de la reacción de secuenciación a granel para generar información de la secuencia se lleva a cabo para al menos 1.000.000 de complejos fusionados de al menos 10.000 células dentro de la población de células.

13. El método de la reivindicación 1, que comprende, además:

- 5 (A) proporcionar un segundo conjunto de sondas de ácido nucleico, comprendiendo el segundo conjunto una quinta sonda que comprende una secuencia que es complementaria a una tercera sub-secuencia de ácido nucleico diana, comprendiendo una sexta sonda una secuencia que es complementaria a una segunda sub-secuencia de la tercera secuencia de ácido nucleico diana y una segunda secuencia que es complementaria a una segunda secuencia exógena, comprendiendo una séptima sonda la segunda secuencia exógena y una secuencia que es complementaria a una primera sub-secuencia de una cuarta secuencia de ácido nucleico diana, y comprendiendo una octava sonda una secuencia que es complementaria a una segunda sub-secuencia de la cuarta secuencia de ácido nucleico diana;
- 10 aislar las células individuales con los primer y segundo conjuntos de sondas de ácido nucleico;
- amplificar la tercera y cuarta secuencias de ácido nucleico diana de manera independiente, en el que la tercera secuencia de ácido nucleico diana se amplifica utilizando la quinta sonda y la sexta sonda, y en el que la cuarta secuencia de ácido nucleico diana se amplifica utilizando la séptima sonda y la octava sonda;
- 15 hibridar la segunda secuencia exógena a su complemento;
- amplificar la tercera secuencia de ácido nucleico diana, la cuarta secuencia de ácido nucleico diana y la segunda secuencia exógena, utilizando las quinta y octava sondas, generando con ello un complejo fusionado; y
- 20 realizar una reacción de secuenciación a granel para generar información de la secuencia para al menos 100.000 complejos fusionados a partir de al menos 10.000 células dentro de la población de células, en donde la información de la secuencia es suficiente para co-localizar la primera secuencia de ácido nucleico diana, la segunda secuencia de ácido nucleico diana, la tercera secuencia de ácido nucleico diana y la cuarta secuencia de ácido nucleico diana en una célula individual de la población de al menos 10.000 células; o
- 25 (B) proporcionar N conjuntos de sondas de ácido nucleico, en donde cada uno de los N conjuntos comprende una sonda  $I_1$  que comprende una secuencia que es complementaria a una primera sub-secuencia de ácido nucleico diana  $I_a$ , comprendiendo una sonda  $I_2$  una secuencia que es complementaria a una segunda sub-secuencia de ácido nucleico diana  $I_a$  y una segunda secuencia que es complementaria a una secuencia exógena  $I$ , comprendiendo una sonda  $I_3$  la secuencia exógena  $I$  y una secuencia que es complementaria a una primera sub-secuencia de ácido nucleico diana  $I_b$  y comprendiendo una sonda  $I_4$  una secuencia que es complementaria a una segunda sub-secuencia de ácido nucleico diana  $I_b$ , en donde  $I$  oscila entre 1 y N;
- 30 aislar las células individuales con los N conjuntos de sondas de ácido nucleico;
- 35 amplificar para todos los valores de  $I$  las secuencias de ácido nucleico diana  $I_a$  e  $I_b$ , independientemente, en donde la secuencia de ácido nucleico diana  $I_a$  se amplifica utilizando la sonda  $I_1$  y la sonda  $I_2$  y la secuencia de ácido nucleico diana  $I_b$  se amplifica utilizando la sonda  $I_3$  y la sonda  $I_4$ ;
- 40 hibridar la secuencia exógena  $I$  a su complemento;
- amplificar para cada una de las  $I$ , la secuencia diana  $I_a$ , la secuencia diana  $I_b$  y la secuencia exógena  $I$  utilizando las sondas  $I_1$  e  $I_4$ , generando con ello N complejos fusionados; y
- 45 realizar una reacción de secuenciación a granel para generar información de la secuencia para al menos 100.000 complejos fusionados a partir de al menos 10.000 células dentro de la población de células, en donde la información de la secuencia es suficiente para co-localizar las N secuencias de ácido nucleico diana  $I_a$  y las secuencias de ácido nucleico diana  $I_b$  en una célula individual de la población de al menos 10.000 células.

14. Un método para analizar al menos dos secuencias de ácido nucleico en una célula individual contenida dentro de una población de al menos 10.000 células, que comprende:

- 50 aislar cada una de una pluralidad de células individuales de una población de al menos 10.000 células en una microgotita de emulsión,

- introducir una secuencia de código de barras exógena, única, que comprende al menos seis nucleótidos en cada una de las microgotitas de emulsión, en donde cada una de las secuencias de código de barras se selecciona de una agrupación de secuencias de código de barras con una diversidad de la secuencia mayor que 1000 veces, en donde cada una de dichas secuencias de código de barras exógenas, únicas, coincide con una célula individual de dicha pluralidad de células individuales;
- 5 para cada una de la pluralidad de células individuales,
- proporcionar al menos un conjunto de sondas de ácido nucleico, comprendiendo el conjunto una primera sonda que comprende una secuencia que es complementaria a una secuencia de ácido nucleico que está localizada en el extremo 5' de la secuencia de código de barras exógena, única, comprendiendo una segunda sonda una secuencia que es complementaria a una secuencia de ácido nucleico que está localizada en el extremo 3' de la secuencia de código de barras exógena, única, y una segunda región de secuencia que es complementaria a una secuencia exógena, no humana, comprendiendo una tercera sonda una secuencia que comprende la secuencia exógena, no humana, y una secuencia que es complementaria a una primera sub-secuencia de una segunda secuencia de ácido nucleico diana, y comprendiendo una cuarta sonda una secuencia que es complementaria a una segunda sub-secuencia de la segunda secuencia de ácido nucleico diana, y en el que la segunda secuencia de ácido nucleico diana comprende una secuencia del receptor de células T o una secuencia de inmunoglobulina;
- 10 15
- amplificar la primera y segunda secuencias de ácido nucleico independientemente, en donde la secuencia de código de barras exógena se amplifica utilizando la primera sonda y la segunda sondas, y en donde la segunda secuencia de ácido nucleico diana se amplifica utilizando la tercera sonda y la cuarta sonda;
- 20 hibridar la secuencia exógena, no humana, a su complemento;
- amplificar la secuencia de código de barras, la segunda secuencia de ácido nucleico diana y la secuencia exógena, no humana, utilizando las primera y cuarta sondas;
- realizar la secuenciación a granel de los complejos fusionados; e
- 25 identificar una célula individual para cada uno de los complejos fusionados basado en la secuencia del código de barras.
15. El método de la reivindicación 14, en el que la secuencia de código de barras se fija a una perla o a una superficie sólida, en donde preferiblemente la perla o la superficie sólida está aislada en la microgotita de emulsión o el recipiente de reacción, en el que preferiblemente introducir una secuencia de código de barras única comprende condensar la microgotita de emulsión o un recipiente de reacción que comprende la célula individual con la microgotita de emulsión o un recipiente de reacción que comprende la secuencia de código de barras fijada a la perla o a la superficie sólida.
- 30
16. El método de la reivindicación 14 ó 15, en el que la segunda secuencia de ácido nucleico diana es complementaria a una secuencia de ARN o es complementaria a una secuencia de ADN.
17. El método según una de las reivindicaciones 14-16, en el que la amplificación comprende
- 35 (a) realizar una reacción en cadena de la polimerasa;
- (b) realizar una reacción en cadena de la ligasa; o
- (c) realizar una reacción en cadena de la ligasa seguida de una reacción en cadena de la polimerasa.
18. El método de la reivindicación 1 ó 14, en el que la célula individual está contenida dentro una población de al menos 25.000 células, al menos 50.000 células, al menos 75.000 células o al menos 100.000 células.
- 40 19. El método de la reivindicación 14, que comprende, además, cuantificar los complejos fusionados.
20. El método según una de las reivindicaciones 14-19, en el que los complejos fusionados son circulares.
21. El método de la reivindicación 14, que comprende, además:
- proporcionar N conjuntos de sondas de ácido nucleico, en donde cada uno de los N conjuntos comprende una sonda  $l_1$  que comprende una secuencia que es complementaria a una primera sub-secuencia de una secuencia de código de barras, comprendiendo una sonda  $l_2$  una secuencia que es complementaria a una segunda sub-secuencia de la secuencia de código de barras y una segunda secuencia que es complementaria a una secuencia exógena I, comprendiendo una sonda  $l_3$  la secuencia exógena I y una secuencia que es complementaria a una primera sub-secuencia de ácido nucleico diana  $l_b$  y comprendiendo una sonda  $l_4$  una secuencia que es complementaria a una segunda sub-secuencia de ácido nucleico diana  $l_b$ , en donde I oscila entre 1 y N;
- 45

aislar las células individuales con los N conjuntos de sondas de ácido nucleico;

amplificar para todos los valores de I la secuencia de código de barras y las secuencias de ácido nucleico diana I<sub>b</sub> independientemente, en donde la secuencia de código de barras se amplifica utilizando la sonda I<sub>1</sub> y la sonda I<sub>2</sub> y la secuencia de ácido nucleico diana I<sub>b</sub> se amplifica utilizando la sonda I<sub>3</sub> y la sonda I<sub>4</sub>;

5 hibridar la secuencia exógena I a su complemento;

amplificar para cada uno de las I, la secuencia de código de barras, la secuencia diana I<sub>b</sub> y la secuencia exógena I utilizando las sondas I<sub>1</sub> e I<sub>4</sub>, generando con ello N complejos fusionados; y

10 realizar una reacción de secuenciación a granel para generar información de la secuencia para al menos 100.000 complejos fusionados a partir de al menos 10.000 células dentro de la población de células, en donde la información de la secuencia es suficiente para co-localizar la secuencia de código de barras y la secuencia de ácido nucleico diana I<sub>b</sub> en una célula individual de la población de al menos 10.000 células.

22. El método de la reivindicación 13(B) o 21, en el que N es menor que o igual a 10, menor que o igual a 100, menor que o igual a 1000, menor que o igual a 10.000, menor que o igual a 100.000 o N representa la totalidad de las transcripciones poliadeniladas en una célula.

15 23. El método de la reivindicación 21 ó 22, en el que la secuencia del código de barras es la misma secuencia para todos los N.

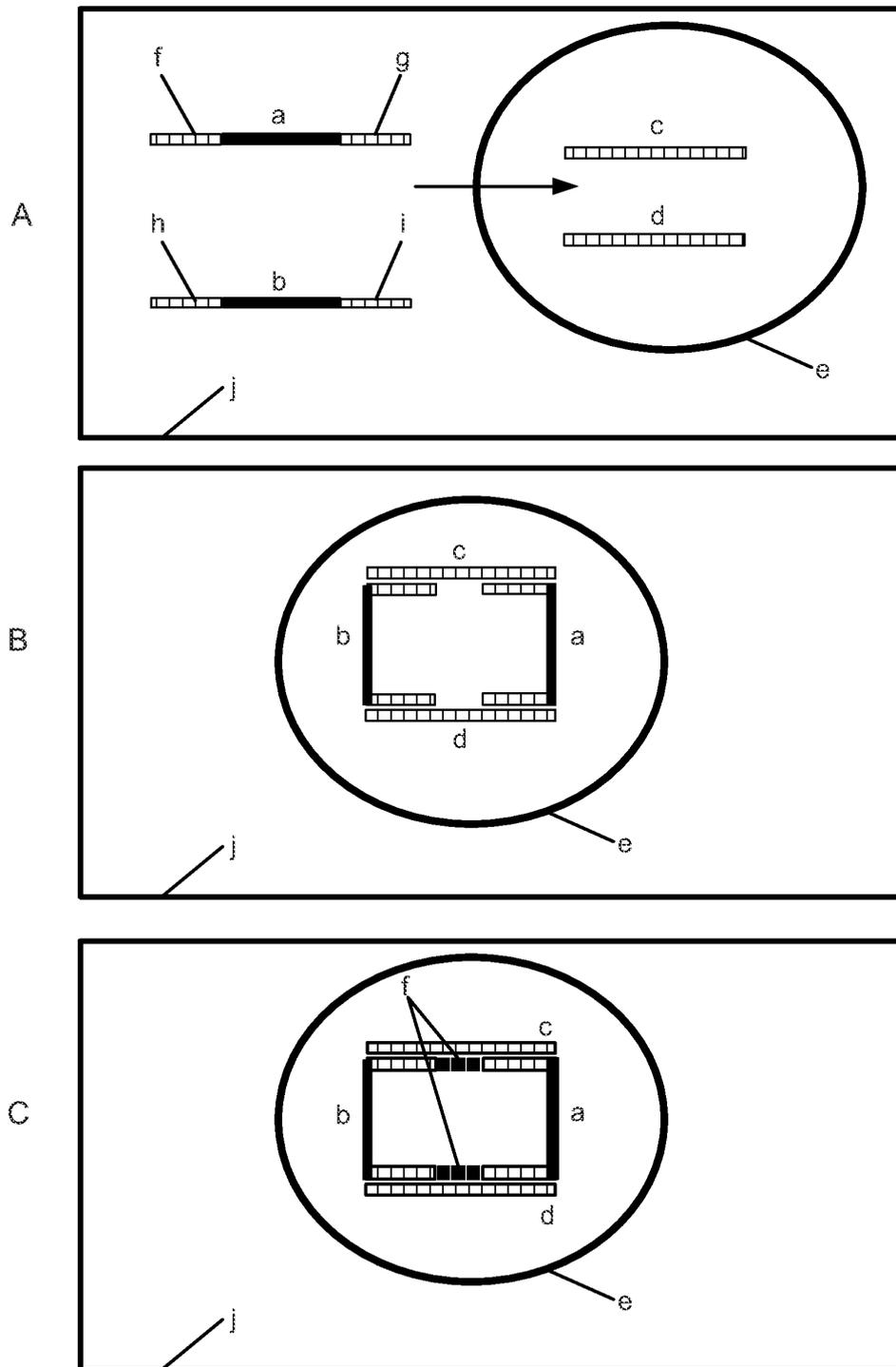


FIG. 1

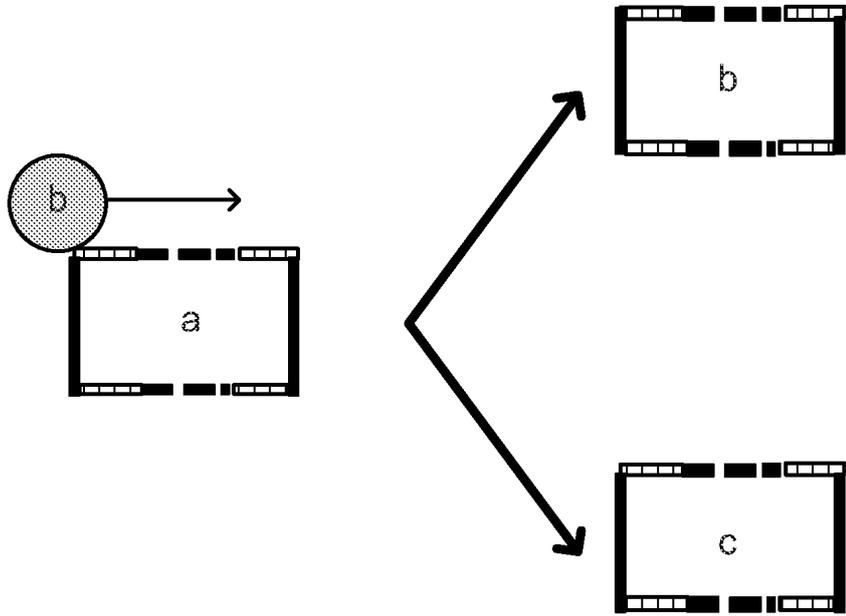


FIG. 2

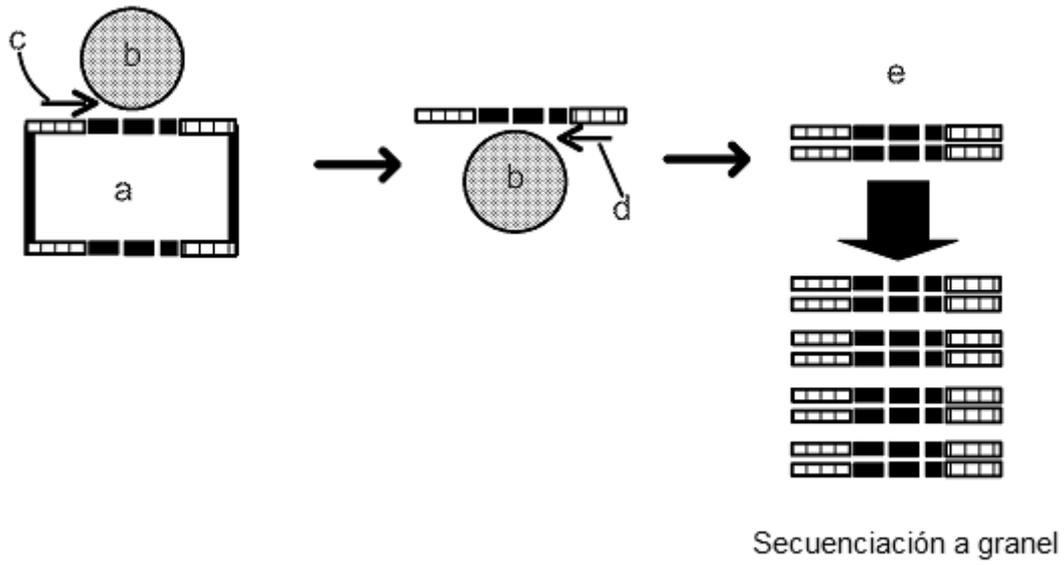


FIG. 3

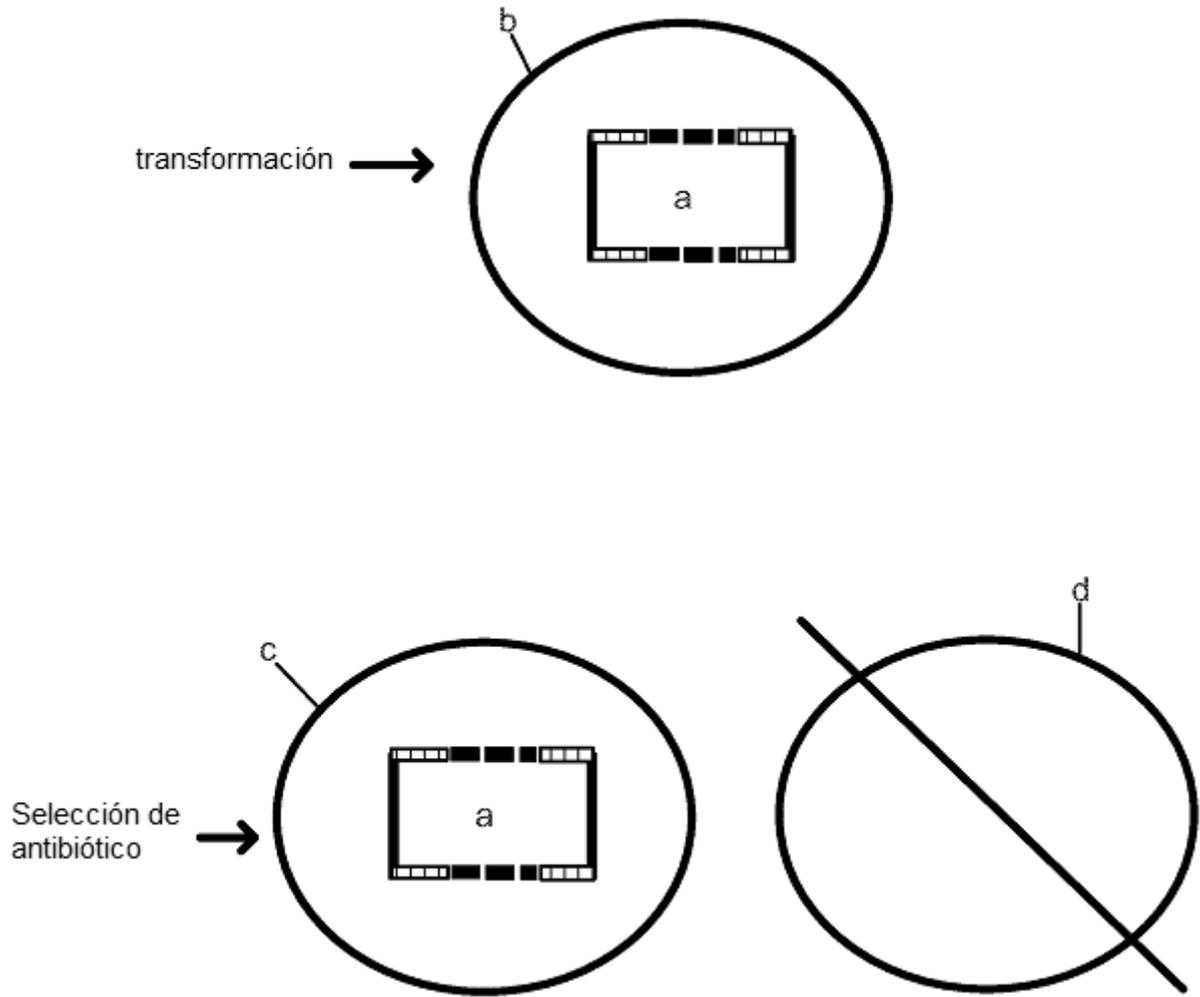


FIG. 4

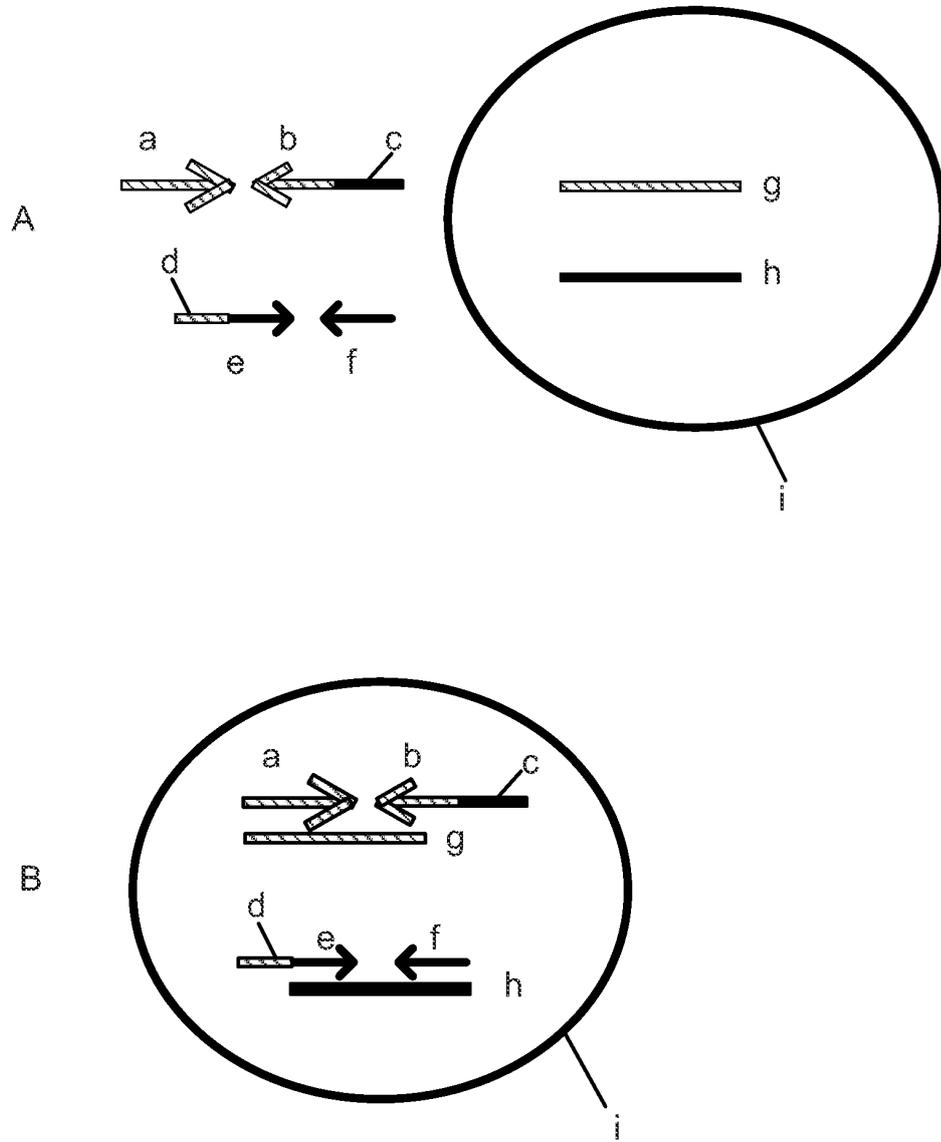


FIG. 5

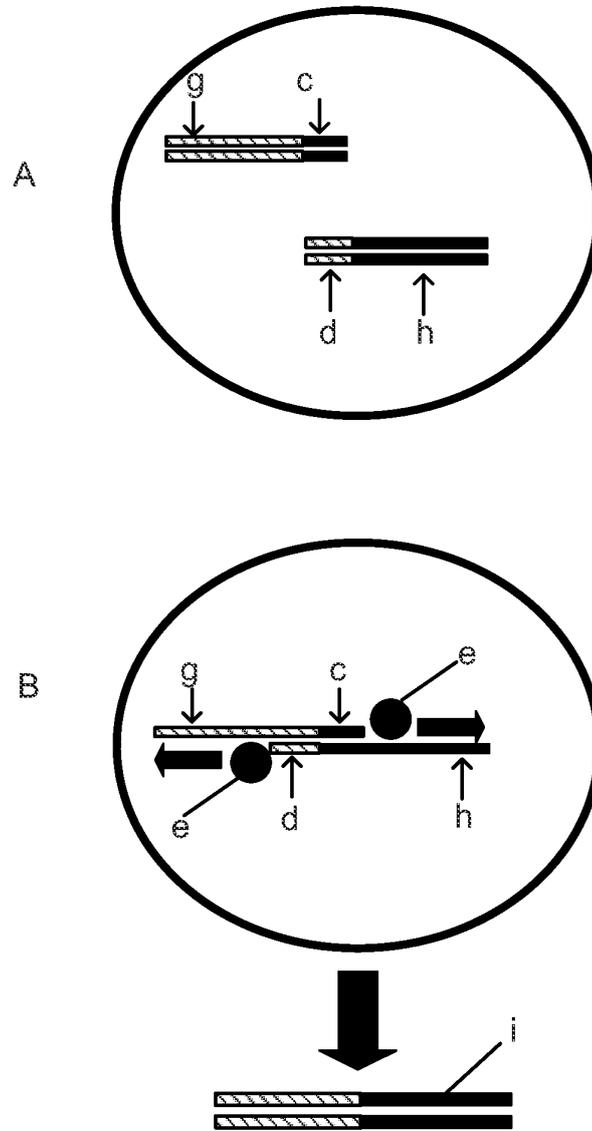


FIG. 6

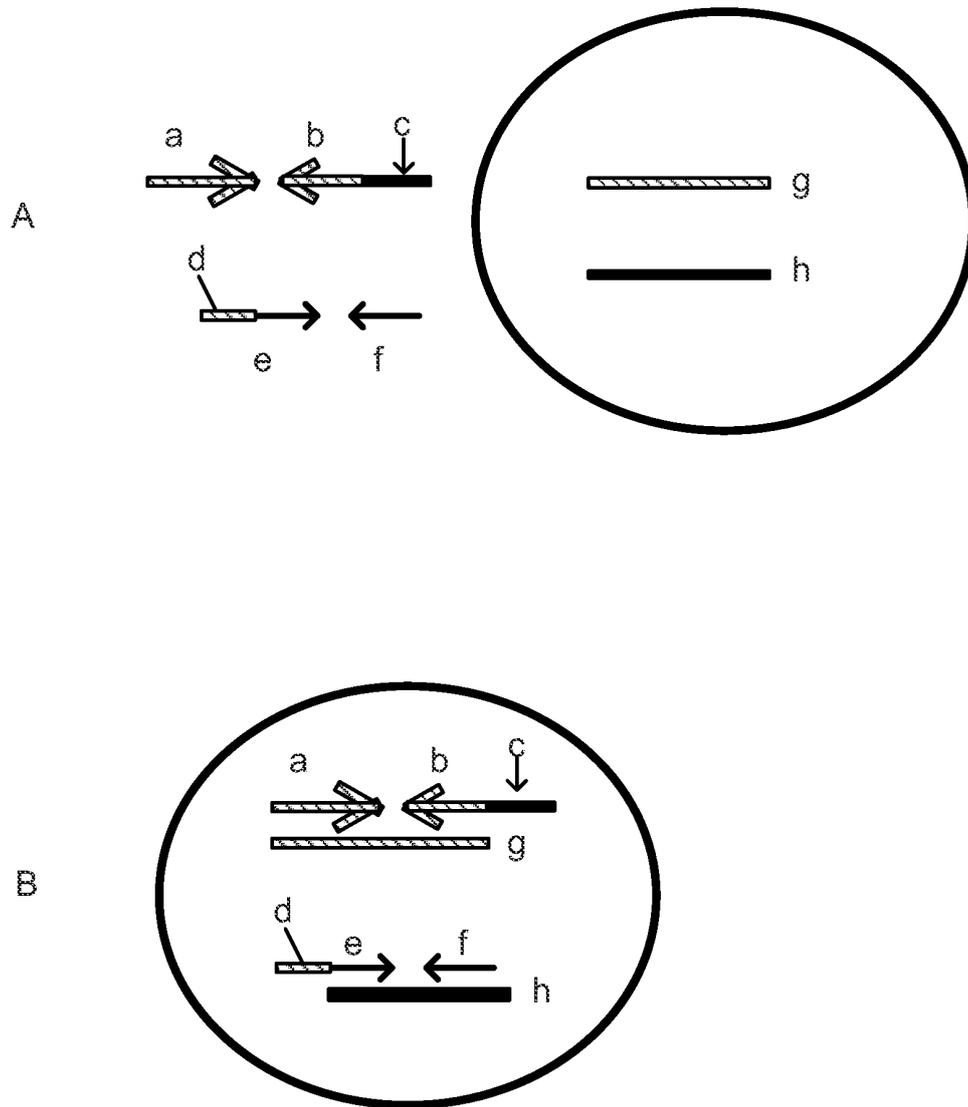


FIG. 7

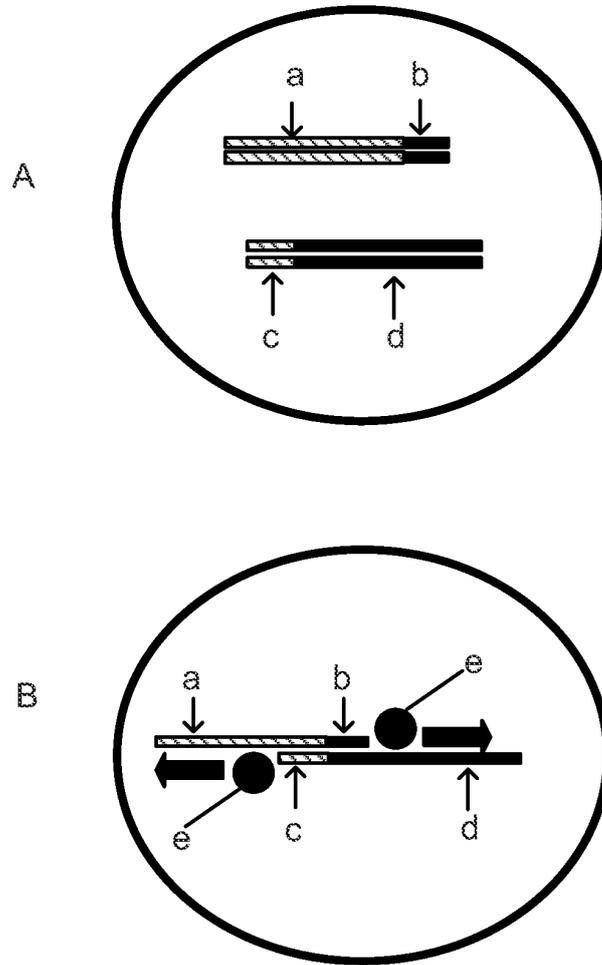


FIG. 8

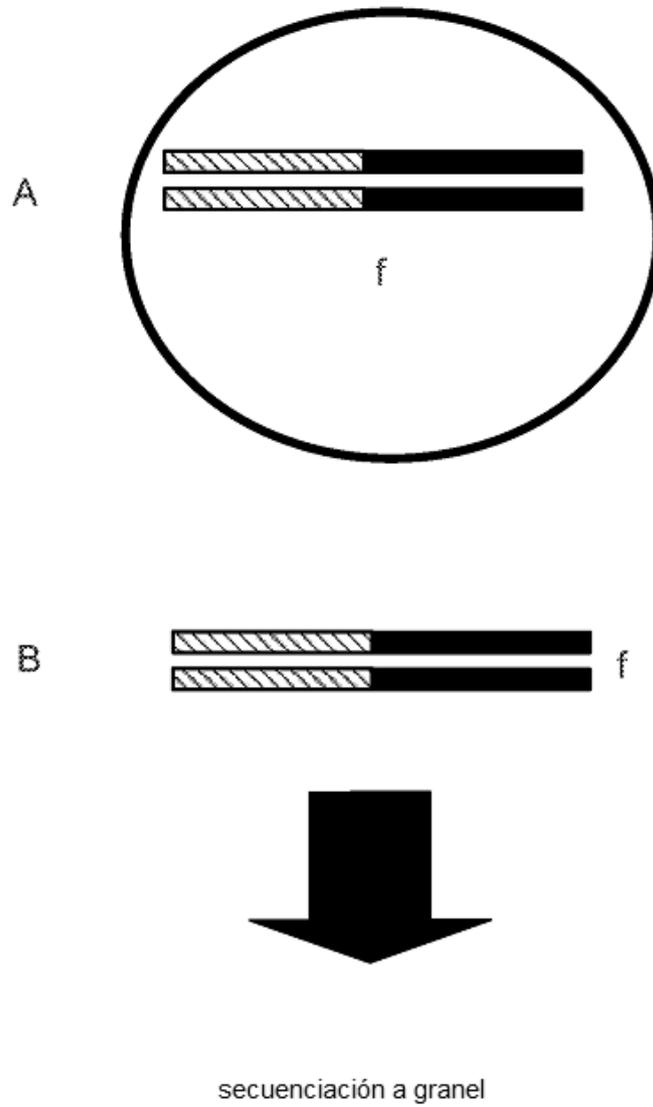


FIG. 9

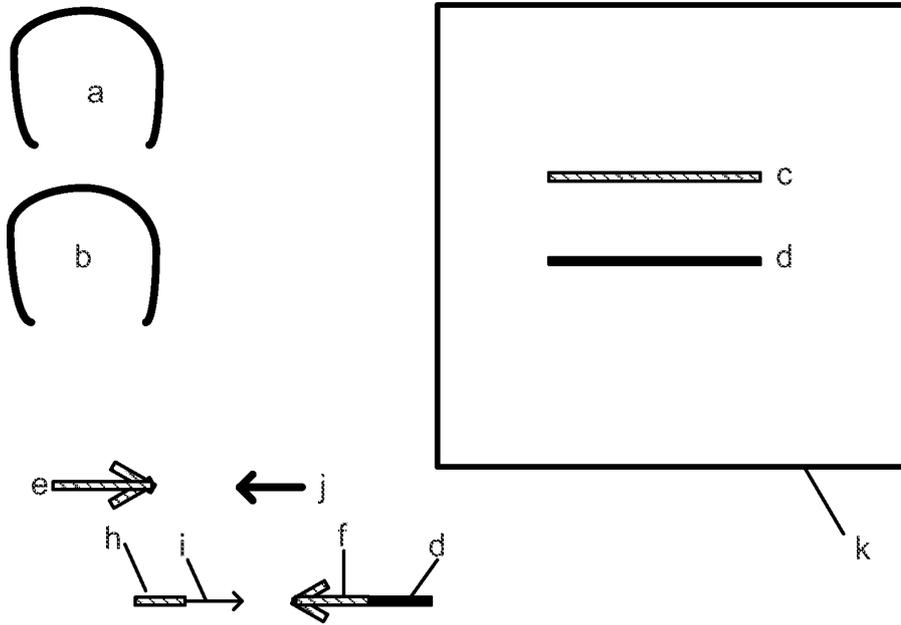


FIG. 10

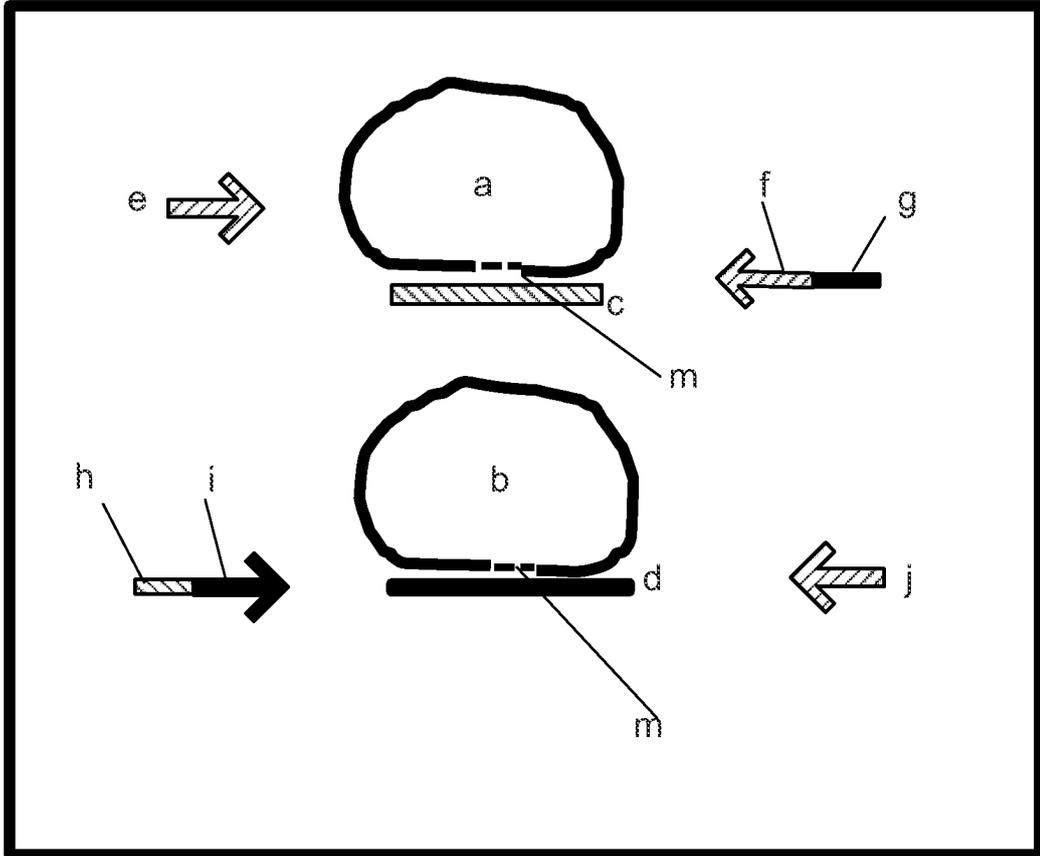


FIG. 11

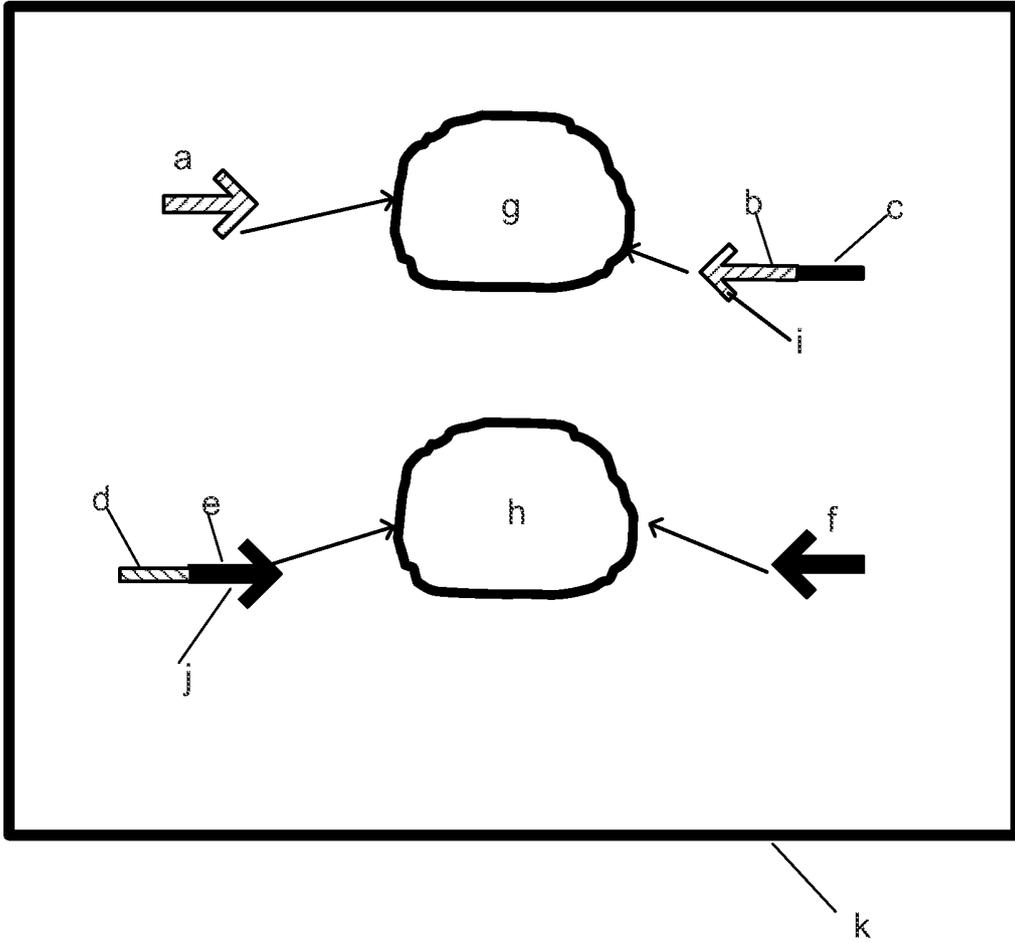


FIG. 12

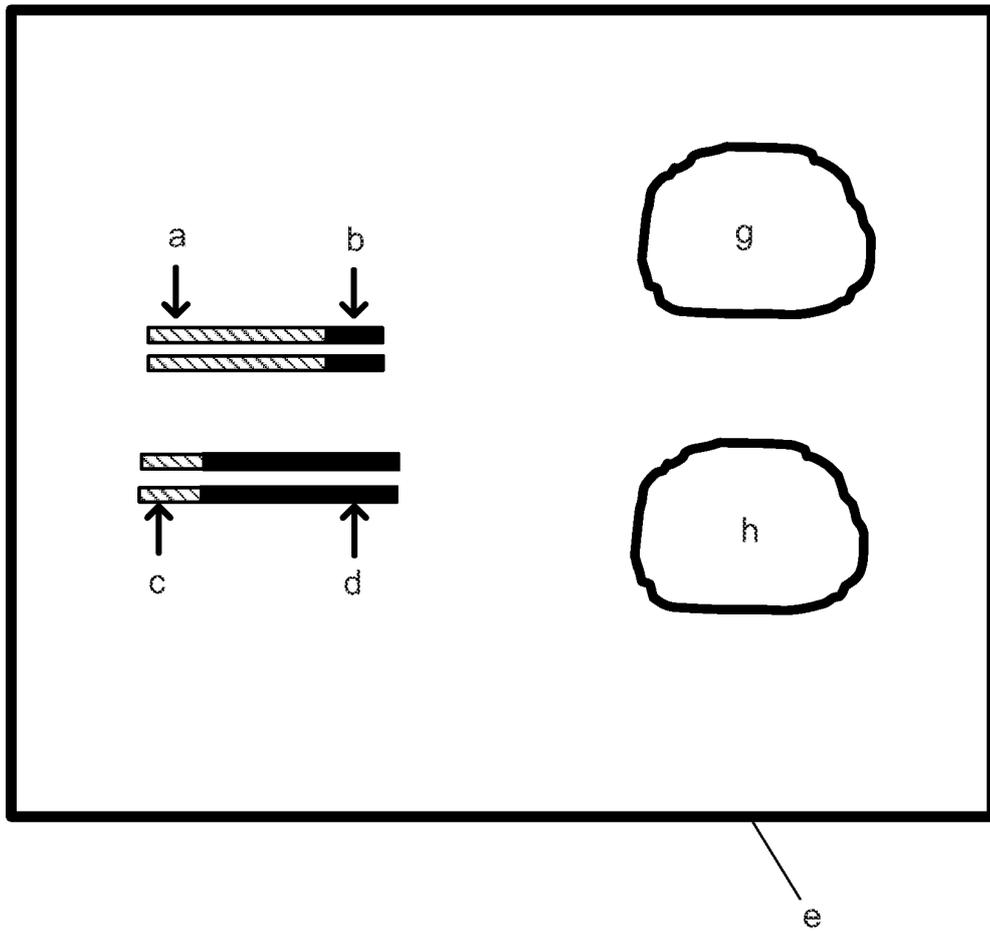


FIG. 13

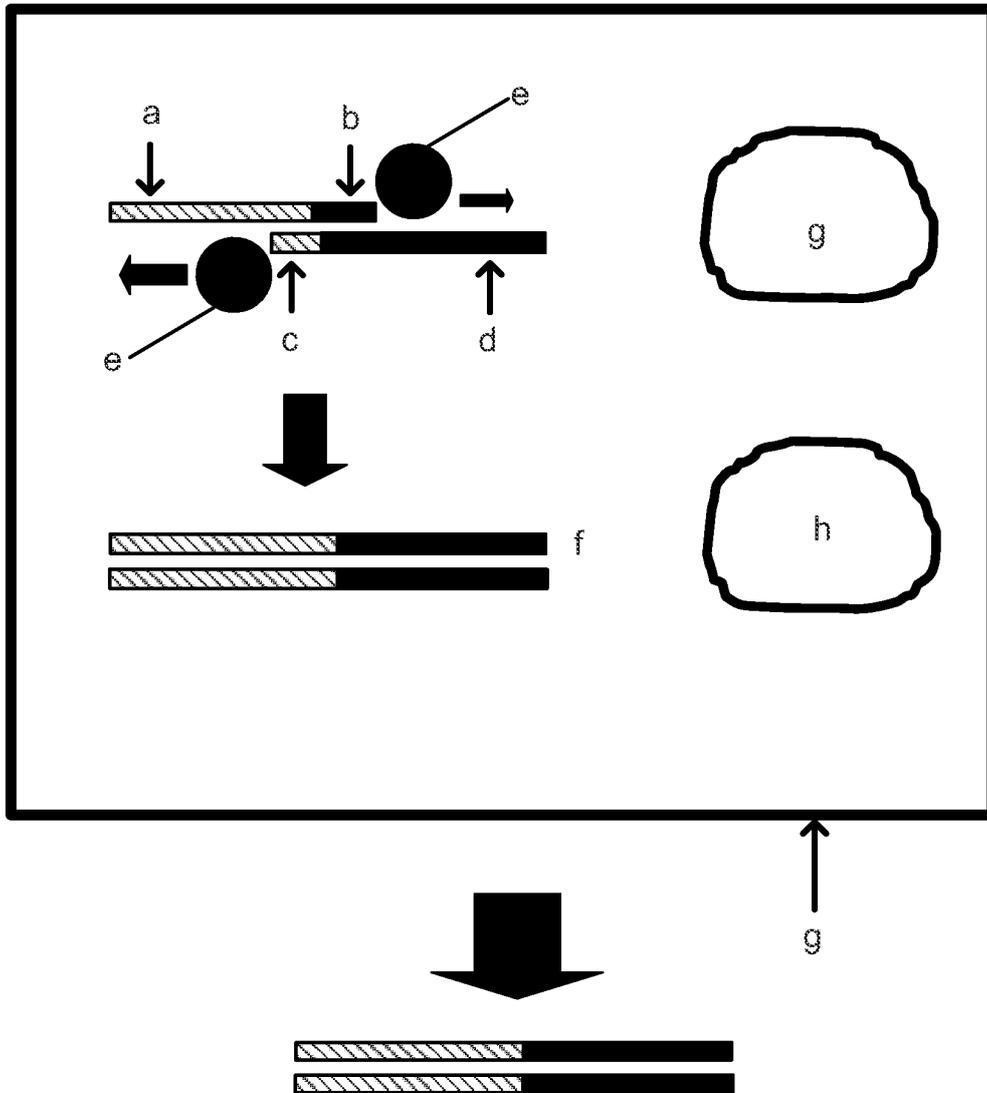


FIG. 14

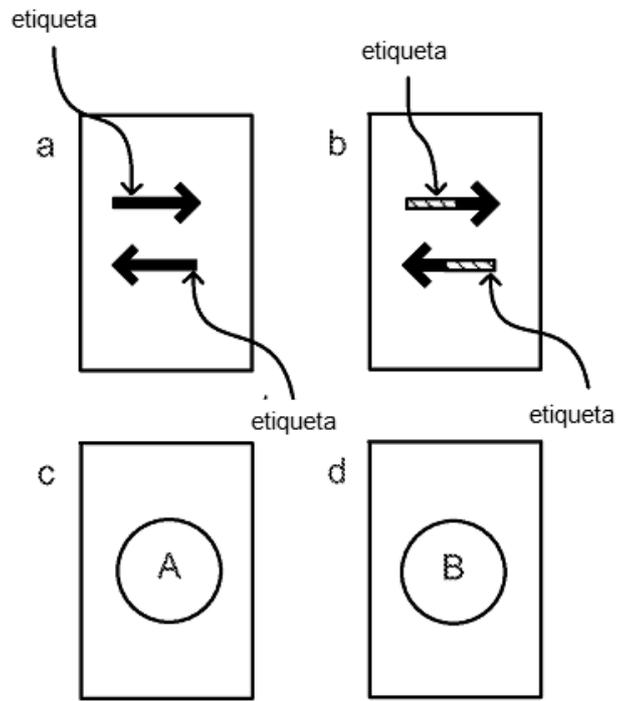


FIG. 15

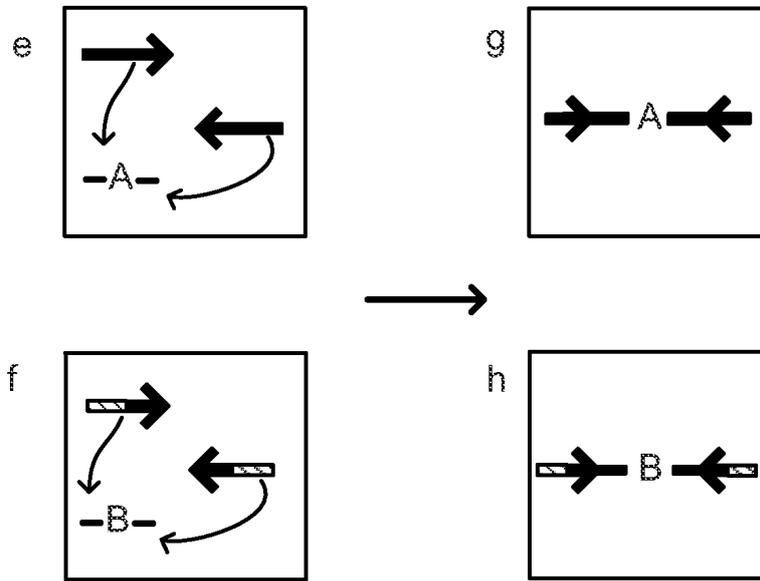


FIG. 16

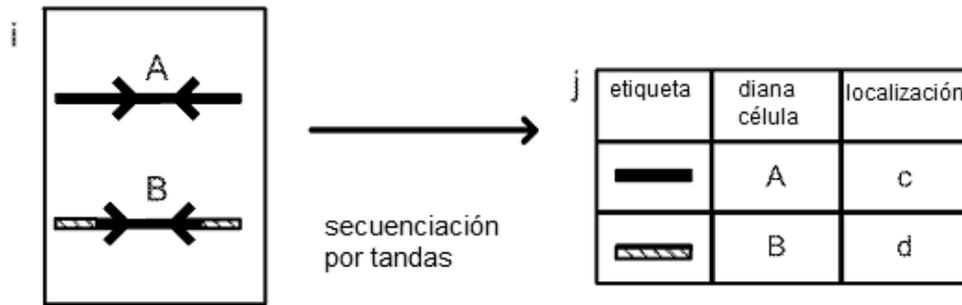


FIG. 17

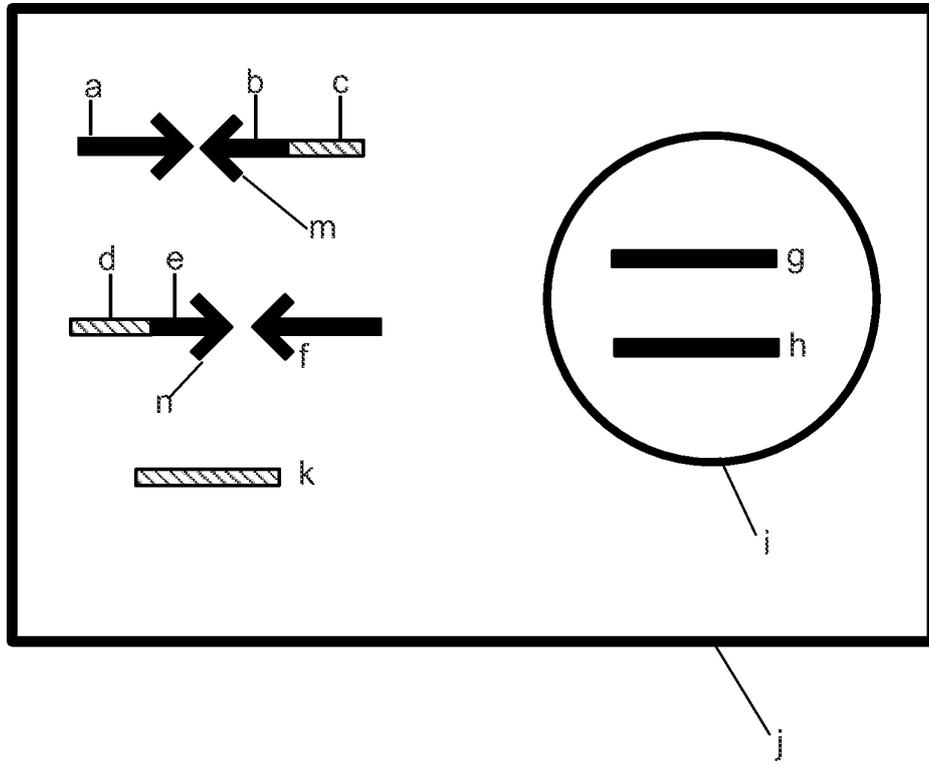


FIG. 18

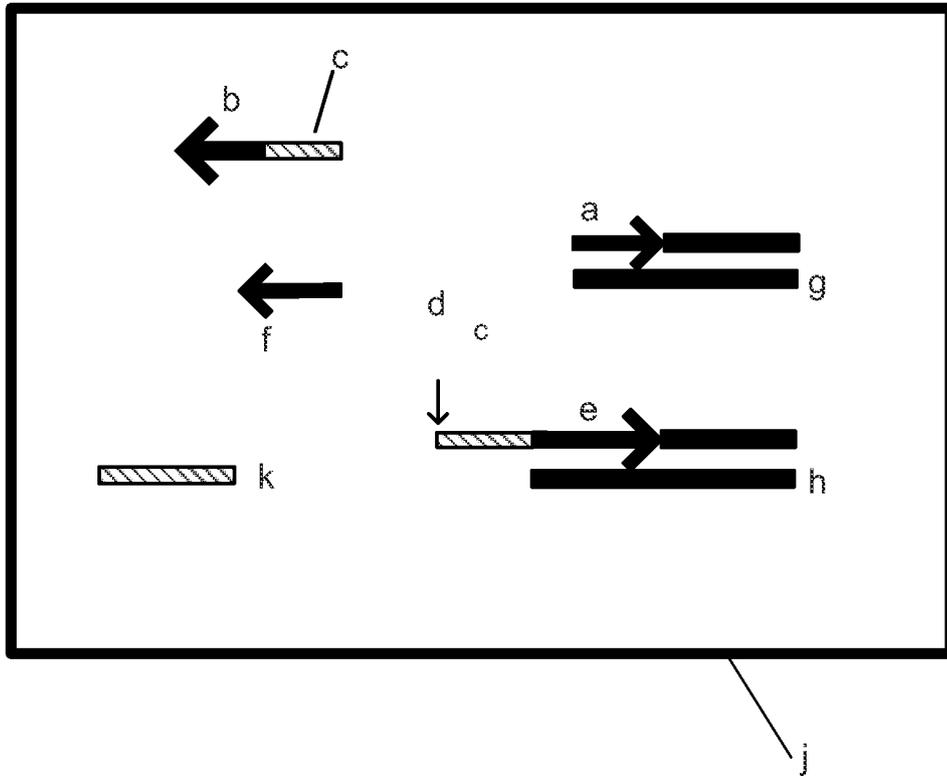


FIG. 19

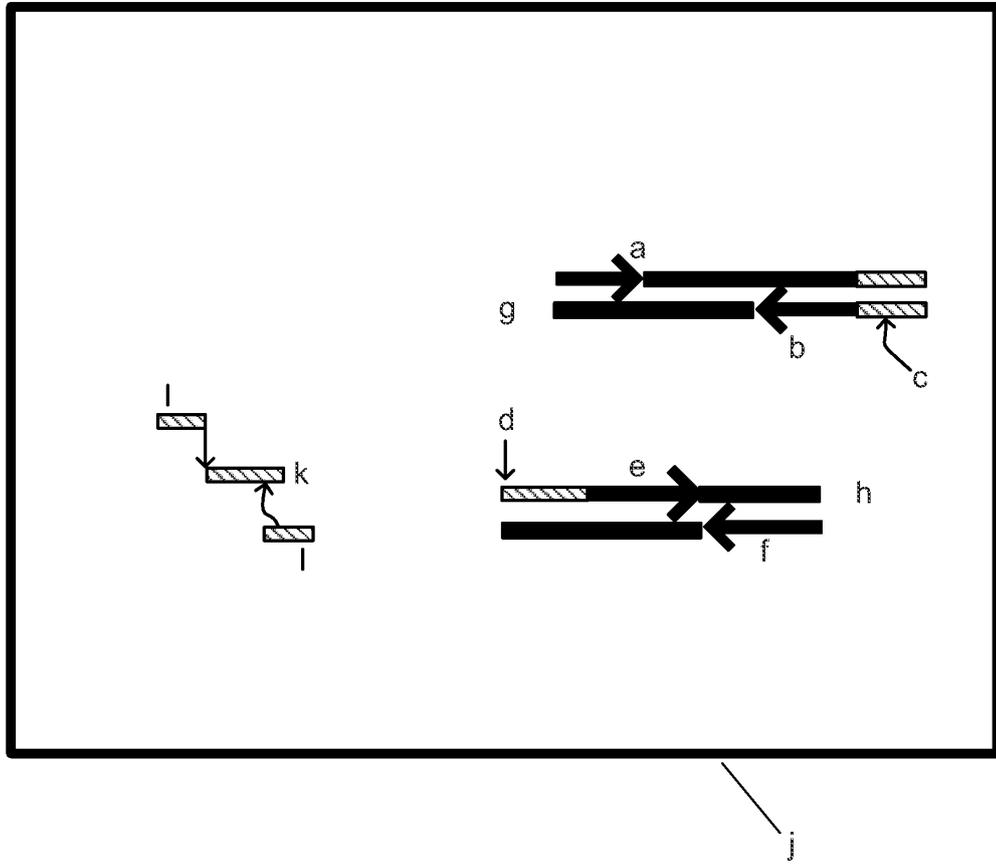
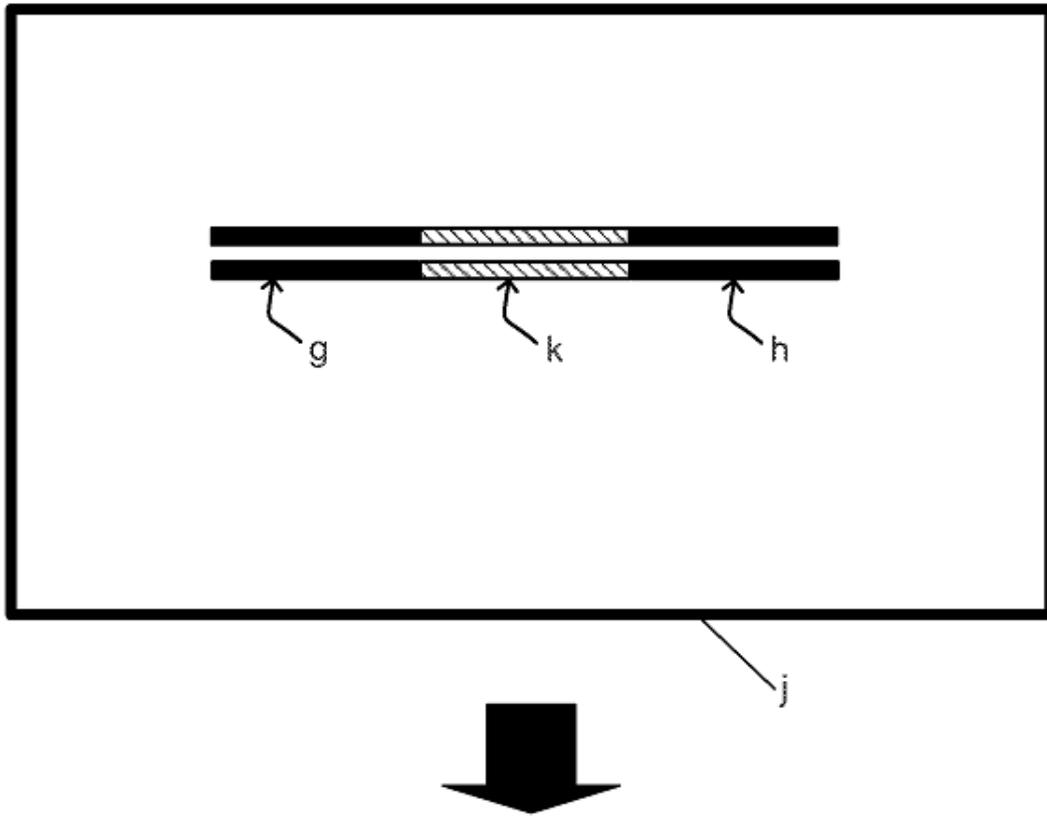


FIG. 20



emulsión inversa y amplicón principal de la  
secuencia a granel

FIG. 21

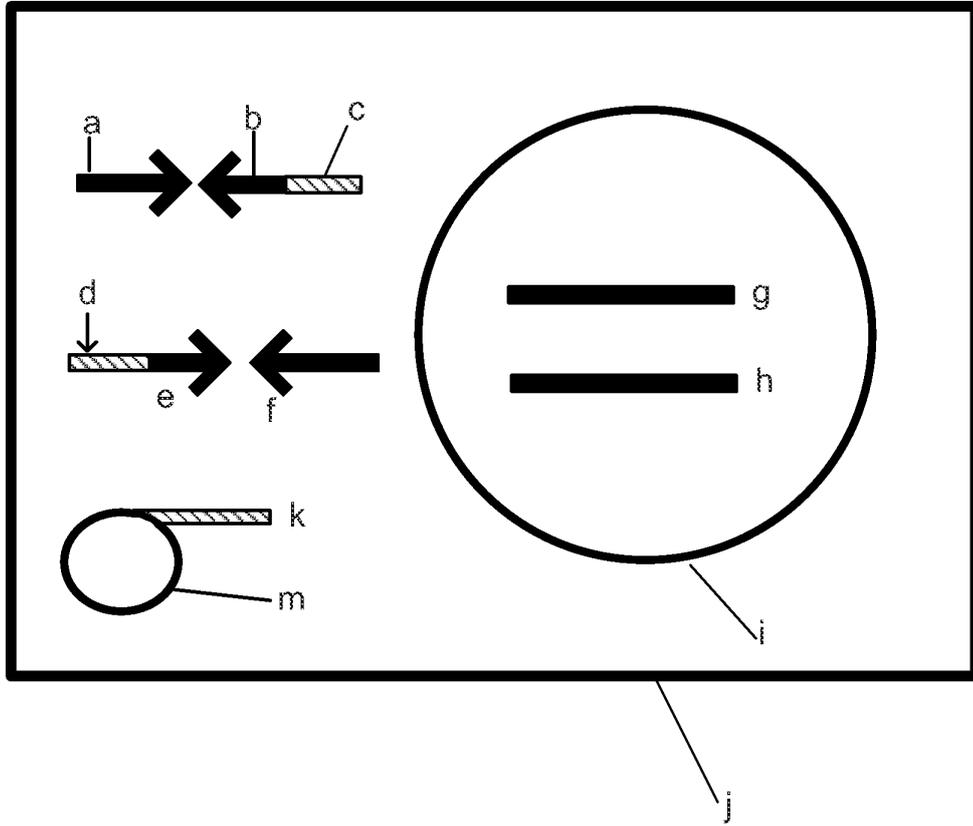


FIG. 22

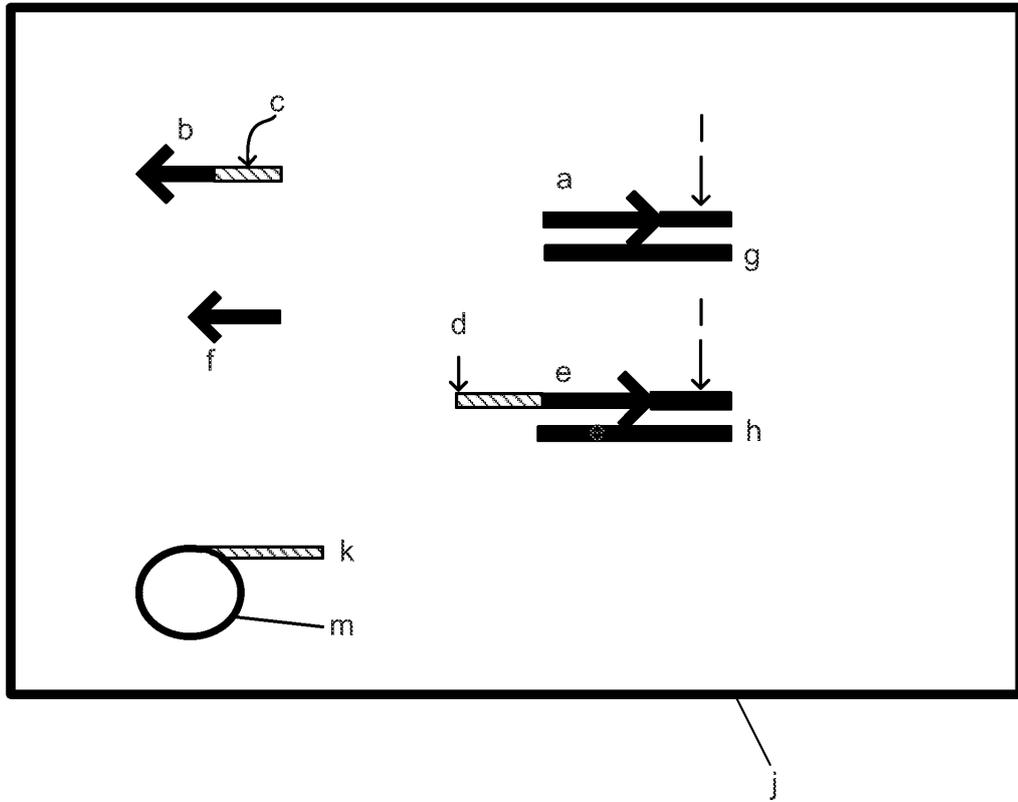


FIG. 23

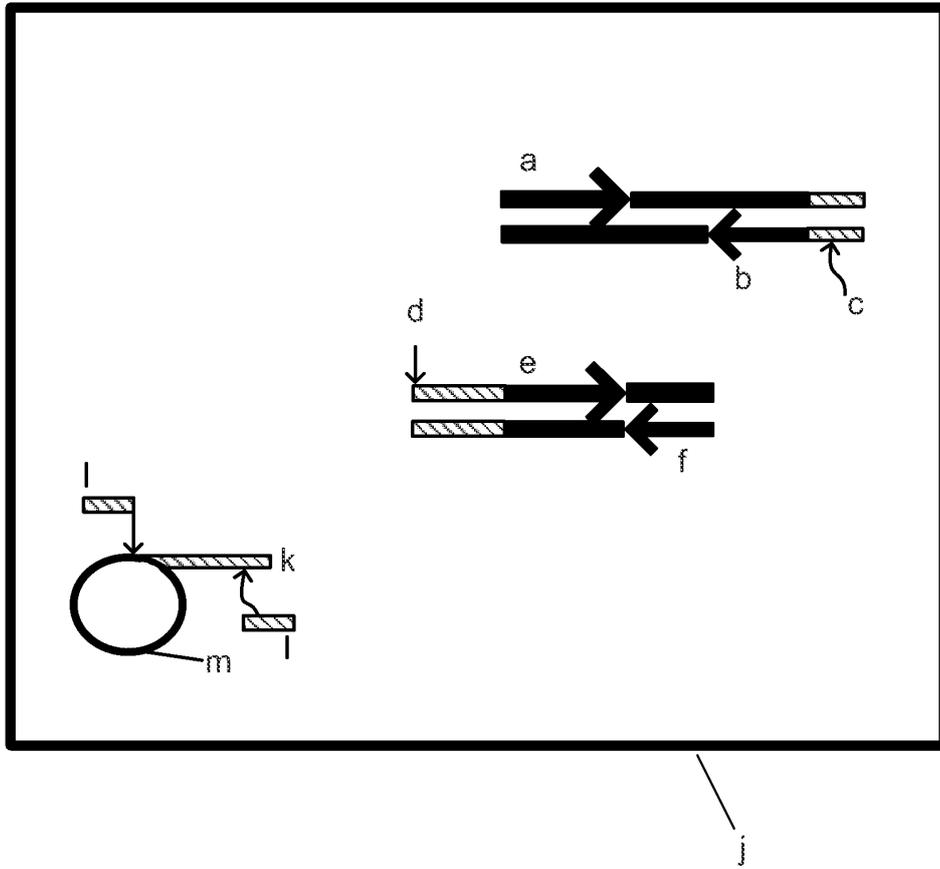
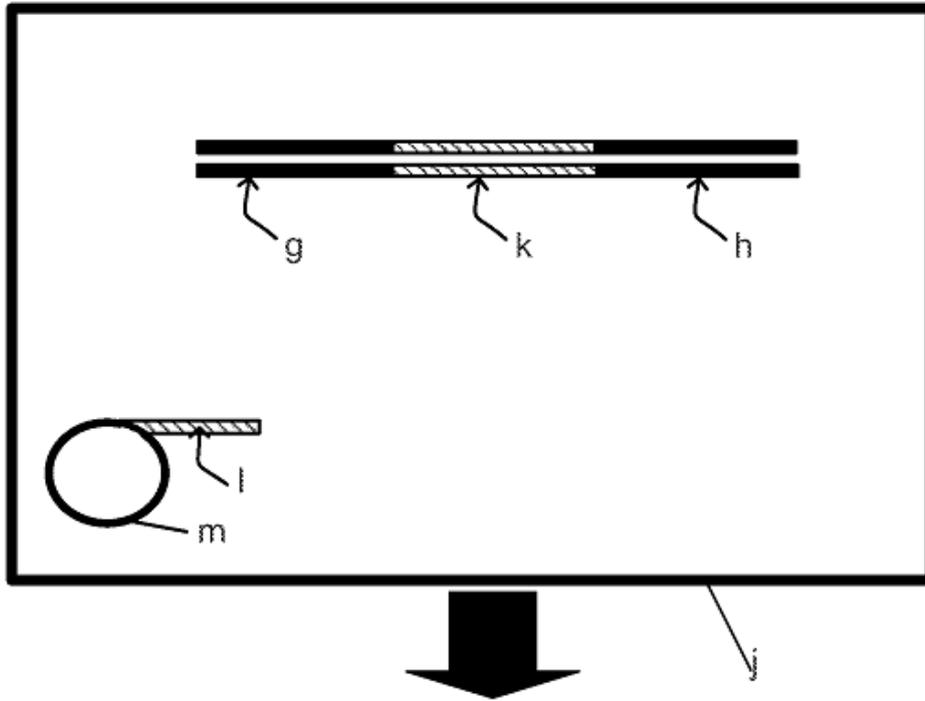


FIG. 24



emulsión inversa y amplicones principales de la secuencia a granel

FIG. 25

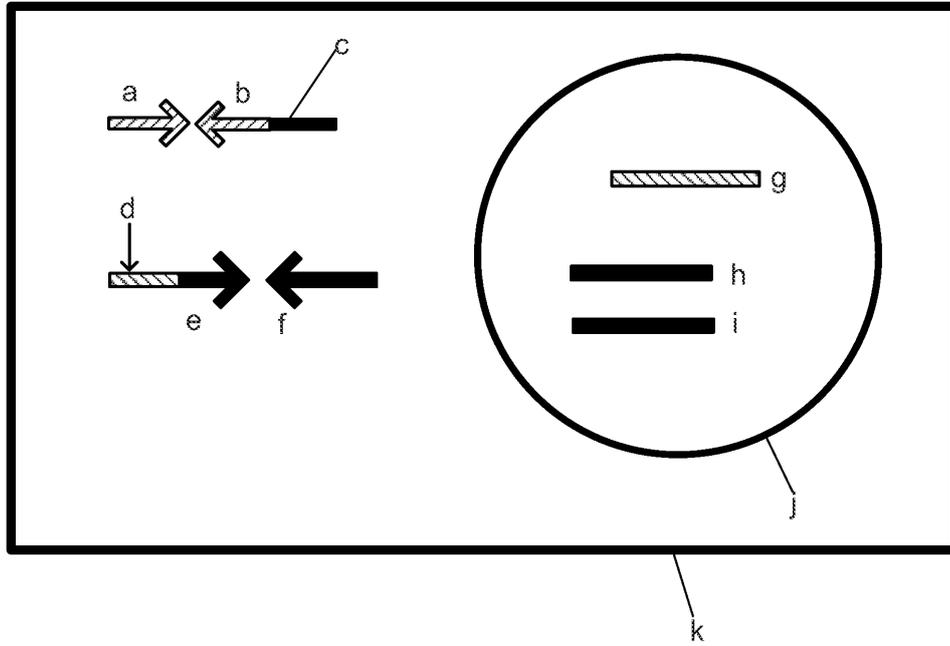


FIG. 26

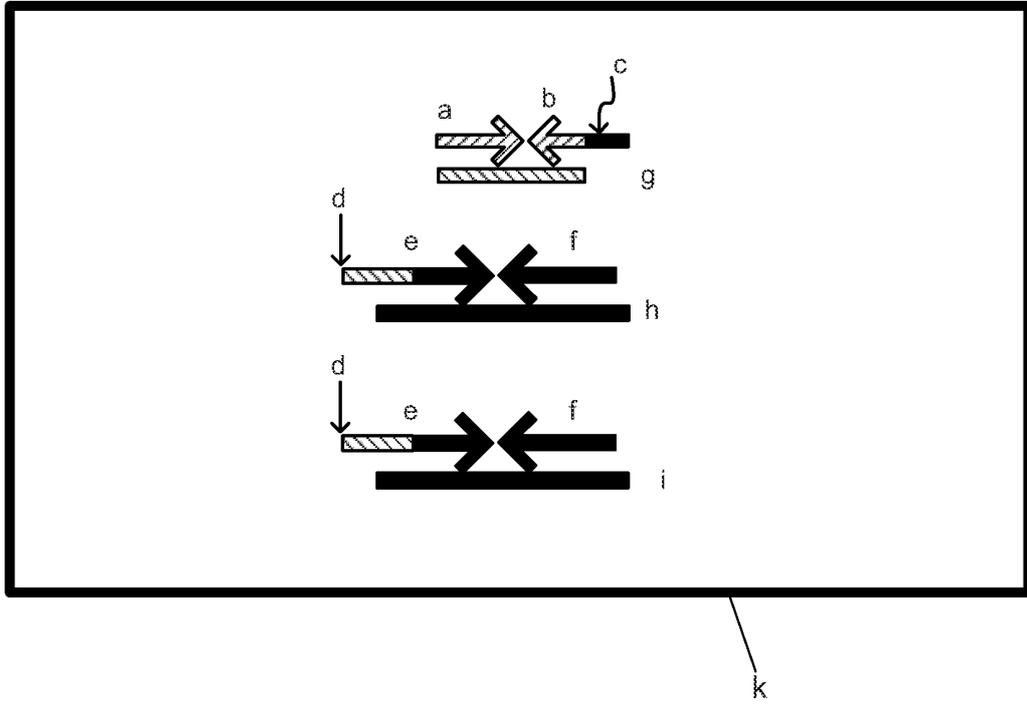


FIG. 27

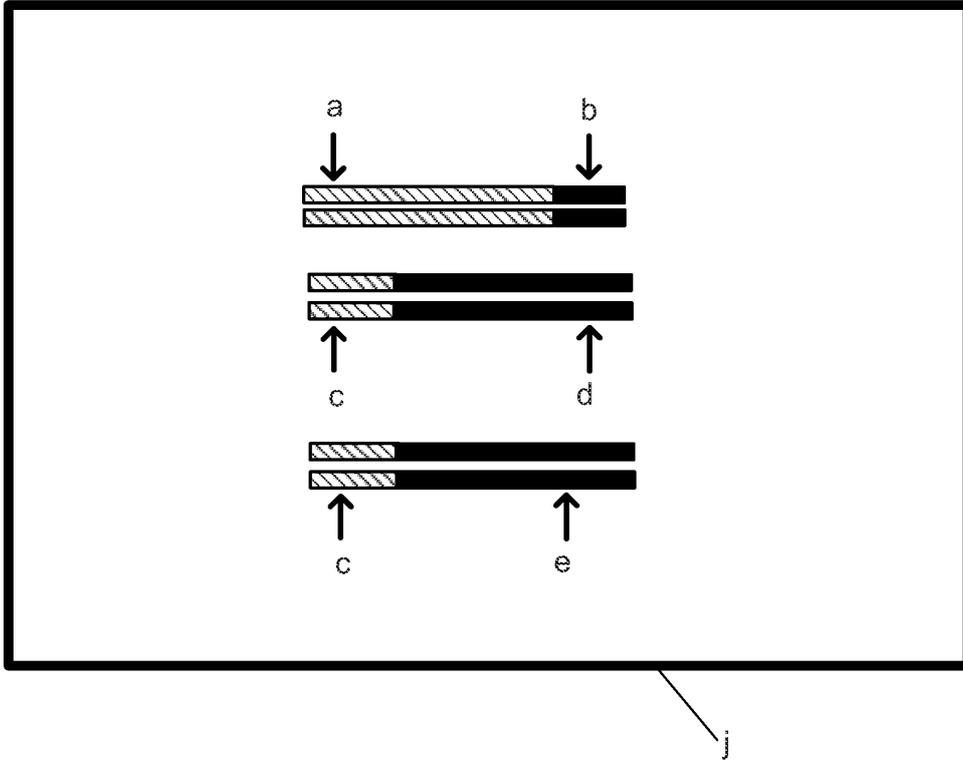


FIG. 28

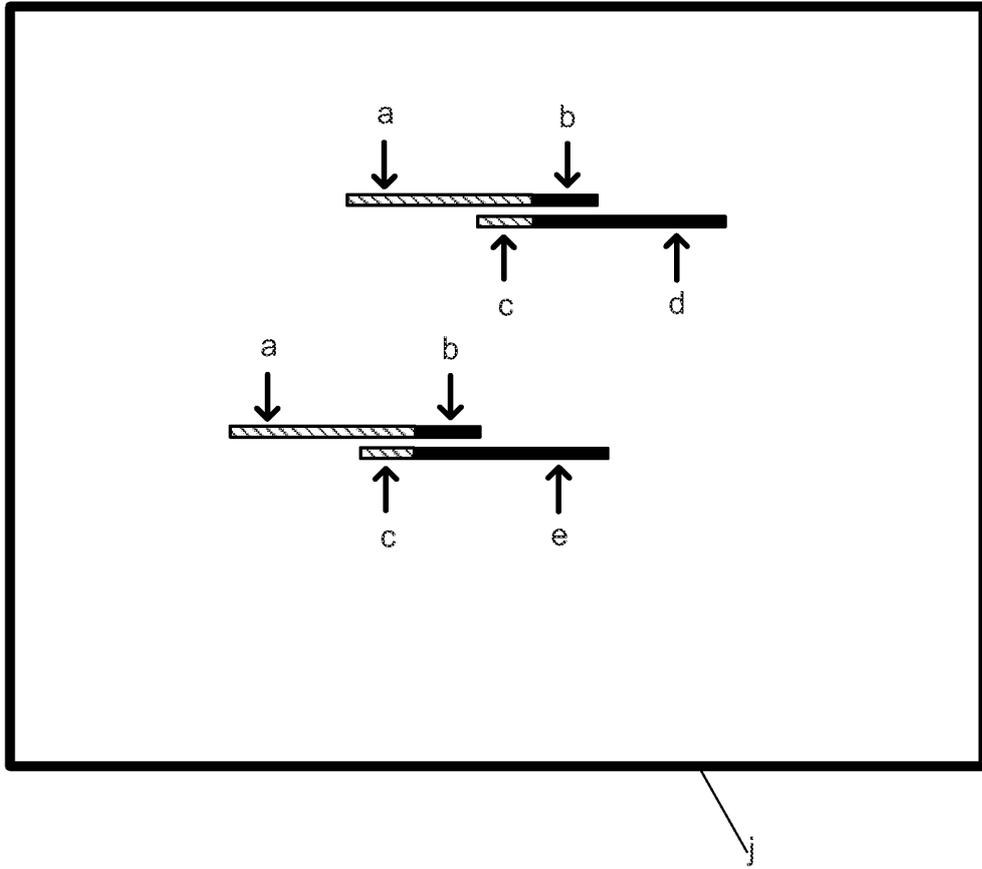
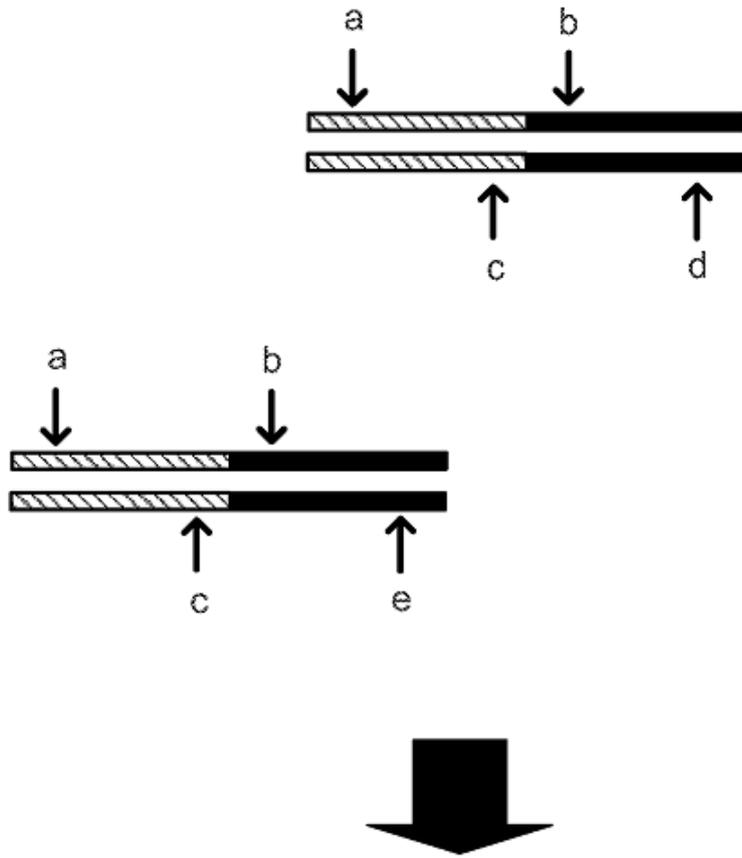


FIG. 29



emulsión inversa y secuencia a granel

FIG. 30

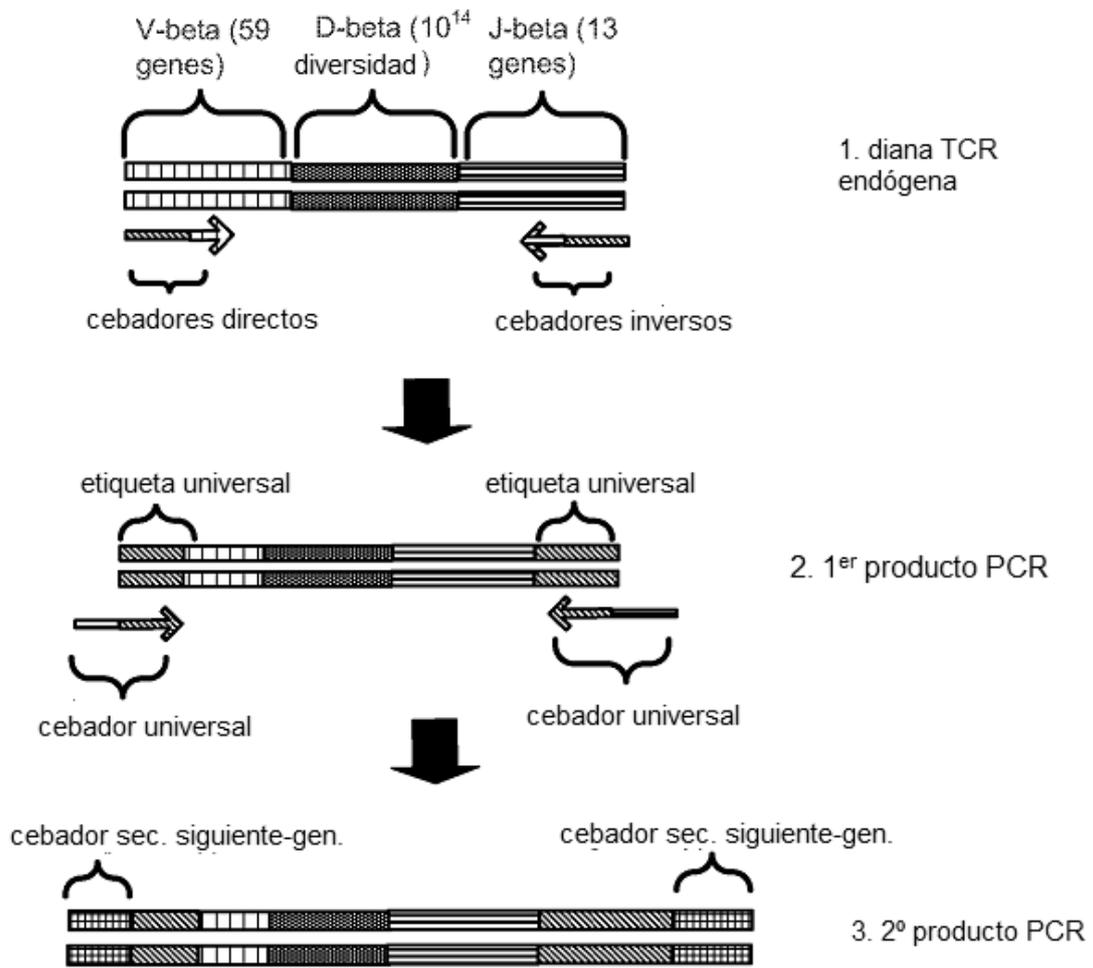


FIG. 31