(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 979 395**

(51) Int. Cl.:

G16B 30/00 (2009.01)
G16B 40/00 (2009.01)
G16B 40/10 (2009.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(86) Fecha de presentación y número de la solicitud internacional: **25.05.2017 PCT/US2017/034576**

(87) Fecha y número de publicación internacional: **30.11.2017 WO17205691**

(96) Fecha de presentación y número de la solicitud europea: **25.05.2017 E 17728398 (3)**

(97) Fecha y número de publicación de la concesión europea: **10.04.2024 EP 3465502**

(54) Título: **Métodos de ajuste del recuento de etiquetas moleculares**

(30) Prioridad:

26.05.2016 US 201662342137 P
31.08.2016 US 201662381945 P
29.09.2016 US 201662401720 P

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
25.09.2024

(73) Titular/es:

BECTON, DICKINSON AND COMPANY (100.0%)
1 Becton Drive
Franklin Lakes, NJ 07417, US

(72) Inventor/es:

FAN, JUE;
TSAI, JENNIFER;
SHUM, ELEEN;
DENG, LISHA y
FU, GLENN, K.

(74) Agente/Representante:

IZQUIERDO BLANCO, María Alicia

DESCRIPCIÓN

Métodos de ajuste del recuento de etiquetas moleculares

ANTECEDENTES

Campo

[0001] La presente divulgación se refiere en general al campo de los códigos de barras moleculares y, más particularmente, a la corrección de errores de secuenciación y PCR utilizando etiquetas moleculares.

Descripción de la técnica relacionada

[0002] Los métodos y técnicas tales como códigos de barras estocásticos son útiles para el análisis celular, en particular para descifrar perfiles de expresión génica para determinar los estados de las células usando, por ejemplo, transcripción inversa, amplificación por reacción en hebra de la polimerasa (PCR) y secuenciación de próxima generación (NGS). Sin embargo, estos métodos y técnicas pueden introducir errores como errores de sustitución (incluidas una o más bases) y errores de no sustitución, si no se corrigen, pueden dar lugar a recuentos moleculares sobreestimados. Por lo tanto, existe una necesidad de métodos y técnicas capaces de corregir diversos errores para lograr recuentos moleculares precisos estimados mediante códigos de barras estocásticos.

[0003] Quan Peng ET AL: "Reducing amplification artifacts in high multiplex amplicon sequencing by using molecular barcodes | BMC Genomics | Full Text", 11 de marzo de 2016 (2016-03-11) revela que se adjuntan códigos de barras de secuencia únicos a cada molécula original de ADN o ARN, esa cuantificación objetivo se realiza contando el número de códigos de barras moleculares únicos en las lecturas, que los códigos de barras están agrupados que están dentro de la distancia de edición y que los grupos de códigos de barras se utilizan para contar moléculas; no describe, entre otras cosas, ajustar la distribución de las etiquetas moleculares de la diana y sus apariciones a dos distribuciones binomiales negativas.

[0004] El documento WO 2015/002908 A1 desvela el perfilado basado en secuencias de poblaciones de ácidos nucleicos mediante amplificación múltiple y unión de una o más etiquetas de secuencia a ácidos nucleicos diana y/o copias de los mismos, seguido de una secuenciación de alto rendimiento del producto de amplificación.

[0005] El documento US 2015/119256 A1 describe el recuento digital de moléculas individuales de alta sensibilidad mediante el etiquetado estocástico de una colección de moléculas idénticas mediante la unión de un conjunto diverso de etiquetas.

RESUMEN

[0006] El alcance de la invención está definido por las reivindicaciones adjuntas. En el presente documento se describen métodos para determinar el número de objetivos. En algunas formas de realización, el método comprende: (a) codificar de barras estocásticamente una pluralidad de objetivos usando una pluralidad de códigos de barras estocásticos para crear una pluralidad de objetivos con códigos de barras estocásticos, en donde cada uno de la pluralidad de códigos de barras estocásticos comprende una etiqueta molecular; (b) obtener datos de secuenciación de los objetivos con códigos de barras estocásticos; y (c) para una o más de la pluralidad de dianas: (i) contar el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación; (ii) determinar un estado de calidad de la diana en los datos de secuenciación obtenidos en (b); (iii) determinar uno o más errores de datos de secuenciación en los datos de secuenciación obtenidos en (b), en donde determinar uno o más errores de datos de secuenciación en los datos de secuenciación comprende determinar uno o más de: el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación, el estado de calidad de la diana en los datos de secuenciación y el número de etiquetas moleculares con secuencias distintas en la pluralidad de códigos de barras estocásticos; y (iv) estimar el número de la diana, en donde el número de la diana estimado se correlaciona con el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación contados en (i) ajustados de acuerdo con uno o más errores de datos de secuenciación determinados en (iii). Los pasos (i), (ii), (iii) y (iv) se pueden realizar para cada uno de la pluralidad de objetivos. El método puede ser multiplexado.

[0007] En algunas formas de realización, el método comprende además: colapsar los datos de secuenciación obtenidos en (b) antes de determinar uno o más errores de datos de secuenciación. Colapsar los datos de secuenciación obtenidos en (b) comprende: atribuir copias de la diana con etiquetas moleculares similares y con apariciones menores que un umbral de ocurrencia de colapso predeterminado como si tuvieran la misma etiqueta molecular para la pluralidad de dianas, en donde dos copias de una diana tienen etiquetas moleculares similares si las etiquetas moleculares de las dos copias de la diana difieren en al menos una base en la secuencia.

[0008] En algunas formas de realización, el umbral de aparición de colapso predeterminado puede ser 7 si los códigos de barras estocásticos comprenden aproximadamente 6561 etiquetas moleculares con secuencias distintas. El umbral de aparición de colapso predeterminado puede ser 17 si los códigos de barras estocásticos comprenden aproximadamente

65536 etiquetas moleculares con secuencias distintas. Dos copias de la diana tienen etiquetas moleculares similares si las etiquetas moleculares de las dos copias de la diana difieren en al menos una base en la secuencia. En algunas formas de realización, el marcador molecular comprende de 5 a 20 nucleótidos. Las etiquetas moleculares de diferentes códigos de barras estocásticos pueden ser diferentes entre sí. La pluralidad de códigos de barras estocásticos comprende aproximadamente 6561 etiquetas moleculares con secuencias distintas. La pluralidad de códigos de barras estocásticos comprende aproximadamente 65536 etiquetas moleculares con secuencias distintas.

[0009] En algunas formas de realización, los datos de secuenciación comprenden secuencias de la pluralidad de dianas con longitudes de lectura de 50 o más nucleótidos. Los datos de secuenciación comprenden secuencias de la pluralidad de objetivos con longitudes de lectura de 75 o más nucleótidos. Los datos de secuenciación comprenden secuencias de la pluralidad de objetivos con longitudes de lectura de 100 o más nucleótidos. Los datos de secuenciación obtenidos en (b) se pueden generar realizando una amplificación por reacción en hebra de la polimerasa (PCR) en la pluralidad de objetivos con códigos de barras estocásticos.

[0010] En algunas formas de realización, uno o más errores de datos de secuenciación pueden ser un error introducido por PCR, un error introducido por secuenciación, un error causado por contaminación de código de barras, un error de preparación de biblioteca o cualquier combinación de los mismos. El error introducido por la PCR puede ser el resultado de un error de amplificación por PCR, un sesgo de amplificación por PCR, una amplificación por PCR insuficiente o cualquier combinación de los mismos. El error introducido por la secuenciación puede ser el resultado de una llamada de base inexacta, una secuenciación insuficiente o cualquier combinación de los mismos.

[0011] En algunas formas de realización, el estado de calidad de la diana en los datos de secuenciación puede ser secuenciación completa, secuenciación incompleta o secuenciación saturada. El estado de calidad de la diana en los datos de secuenciación se puede determinar mediante el número de etiquetas moleculares con secuencias distintas en la pluralidad de códigos de barras estocásticos y la cantidad de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación contados. El estado de calidad del objetivo en los datos de secuenciación se puede clasificar como secuenciación incompleta si el estado de calidad del objetivo en los datos de secuenciación obtenidos en (b) no es una secuenciación completa ni una secuenciación saturada.

[0012] En algunas formas de realización, el estado de calidad de secuenciación completa se puede determinar mediante un índice de dispersión relativo a la distribución de Poisson mayor o igual a un umbral de dispersión de secuenciación completa predeterminado, en donde el umbral de dispersión de secuenciación completa predeterminado puede ser 0,9, 1 o 4. El estado de calidad de la secuenciación completa puede determinarse además mediante una etiqueta molecular con una aparición mayor o igual a un umbral de aparición de secuenciación completa predeterminado en los datos de secuenciación obtenidos en (b), en donde el umbral de aparición de secuenciación completa predeterminado puede ser 10 o 18.

[0013] En algunas formas de realización, el estado de calidad de la secuenciación saturada puede determinarse si la diana tiene un número de etiquetas moleculares con secuencias distintas mayor que un umbral de saturación predeterminado. El estado de calidad de la secuenciación saturada puede determinarse además mediante otra diana de la pluralidad de dianas que tiene una serie de etiquetas moleculares con secuencias distintas que sean mayores que el umbral de saturación predeterminado. El umbral de saturación predeterminado puede ser 6557 si los códigos de barras estocásticos comprenden aproximadamente 6561 etiquetas moleculares con secuencias distintas. El umbral de saturación predeterminado puede ser 65532 si los códigos de barras estocásticos comprenden aproximadamente 65536 etiquetas moleculares con secuencias distintas.

[0014] En algunas formas de realización, el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación contados en (i) se ajusta en (iv), si la diana tiene el estado de calidad de secuenciación completo, determinando todos las etiquetas moleculares secundarias para una o más etiquetas moleculares originales; realizar un primer análisis estadístico para al menos una etiqueta molecular secundaria y la etiqueta molecular principal; y atribuir la aparición de la etiqueta molecular secundaria a la etiqueta molecular principal si se acepta la hipótesis nula del primer análisis estadístico.

[0015] En algunas formas de realización, uno o más etiquetas moleculares originales comprenden etiquetas moleculares con apariciones mayores o iguales a un umbral principal de secuenciación completa predeterminado, en donde el umbral principal de secuenciación completa predeterminado es igual al umbral de aparición de secuenciación completa predeterminada. Las etiquetas moleculares secundarias comprenden etiquetas moleculares que difieren de la etiqueta molecular principal en una base y tienen apariciones menores que o iguales a un umbral secundario de secuenciación completa predeterminado, en donde el umbral secundario de secuenciación completa predeterminado puede ser 3 o 5. La hipótesis nula del primer análisis estadístico puede aceptarse si la probabilidad de que la hipótesis nula sea verdadera está por debajo de las tasas de descubrimiento falso, donde la tasa de descubrimiento falso es del 5 % o del 10 %. El primer análisis estadístico puede ser una prueba binomial múltiple.

[0016] En algunas formas de realización, el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación contados en (i) se ajusta en (iv), si la diana tiene el estado de calidad de secuenciación completo, umbralizando las etiquetas moleculares de la diana para determinar etiquetas moleculares

verdaderas y etiquetas moleculares falsas asociadas con la diana en los datos de secuenciación obtenidos en (b). Establecer un umbral para las etiquetas moleculares de la diana comprende realizar un segundo análisis estadístico sobre las etiquetas moleculares de la diana.

[0017] En algunas formas de realización, realizar el segundo análisis estadístico comprende: ajustar la distribución de las etiquetas moleculares de la diana y sus apariciones a dos distribuciones de Poisson; determinar el número de etiquetas moleculares verdaderas n usando las dos distribuciones de Poisson; y eliminar las etiquetas moleculares falsas de los datos de secuenciación obtenidos en (b), en el que las etiquetas moleculares falsas comprenden etiquetas moleculares con apariciones inferiores a la aparición del enésimo marcador molecular más abundante, y en el que las etiquetas moleculares verdaderas comprenden etiquetas moleculares con apariciones mayores mayor o igual a la aparición de la enésima etiqueta molecular más abundante. Las dos distribuciones de Poisson comprenden una primera distribución de Poisson correspondiente a las etiquetas moleculares verdaderas y una segunda distribución de Poisson para las etiquetas moleculares falsas.

[0018] En algunas formas de realización, el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación contados en (i) se puede ajustar en (iv) si el estado de calidad de la diana en los datos de secuenciación obtenidos en (b) es el estado de calidad de la secuenciación incompleta, que determina si el objetivo tiene ruido en los datos de secuenciación obtenidos en (b); y eliminar el objetivo ruidoso de los datos de secuenciación obtenidos en (b). La diana puede ser ruidosa si la aparición de las etiquetas moleculares de las dianas ruidosas es menor o igual a un umbral de diana ruidosa de secuenciación incompleta, en donde el umbral del gen ruidoso de secuenciación incompleta es 5. El umbral de diana ruidosa de secuenciación incompleta puede igualar la mediana o ocurrencia media de las etiquetas moleculares de la pluralidad de dianas con estados de calidad de secuenciación completa.

[0019] En algunas formas de realización, el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación contados en (i) se puede ajustar en (iv) si el estado de calidad de la diana en los datos de secuenciación obtenidos en (b) es el estado de calidad de la secuenciación incompleta, que establece un umbral para las etiquetas moleculares del objetivo para determinar las etiquetas moleculares verdaderas y las etiquetas moleculares falsas en los datos de secuenciación obtenidos en (b).

[0020] En algunas formas de realización, establecer un umbral para las etiquetas moleculares de la diana comprende realizar un tercer análisis estadístico en las etiquetas moleculares. Realizar el tercer análisis estadístico de las etiquetas moleculares comprende: determinar el número de etiquetas moleculares verdaderas n usando un modelo de Poisson truncado en cero; y eliminar las etiquetas moleculares falsas de los datos de secuenciación obtenidos en (b), en el que las etiquetas moleculares falsas comprenden etiquetas moleculares con apariciones inferiores a la aparición del enésimo marcador molecular más abundante, y en el que las etiquetas moleculares verdaderas comprenden etiquetas moleculares con apariciones mayores mayor o igual a la aparición de la enésima etiqueta molecular más abundante.

[0021] En algunas formas de realización, al menos el 50 % o el 80 % de las etiquetas moleculares en los datos de secuenciación obtenidos en (b) se pueden retener después de que los datos de secuenciación contados en (i) se ajusten de acuerdo con uno o más errores de datos de secuenciación determinados en (iii).

[0022] En algunas formas de realización, codificar de forma estocástica la pluralidad de objetivos comprende hibridar la pluralidad de códigos de barras estocásticos con la pluralidad de objetivos para crear los objetivos con código de barras estocástico. Codificar de forma estocástica la pluralidad de objetivos comprende generar una biblioteca indexada de los objetivos con código de barras estocástico. La generación de una biblioteca indexada de objetivos con códigos de barras estocásticos se puede realizar con un soporte sólido que comprende la pluralidad de códigos de barras estocásticos. El soporte sólido comprende una pluralidad de partículas sintéticas asociadas con la pluralidad de códigos de barras estocásticos. El soporte sólido comprende la pluralidad de códigos de barras estocásticos en dos o tres dimensiones. El soporte sólido comprende un polímero, una matriz, un hidrogel, un dispositivo de conjunto de agujas, un anticuerpo o cualquier combinación de los mismos.

[0023] En algunas formas de realización, cada uno de la pluralidad de códigos de barras estocásticos comprende una o más de una etiqueta de muestra, una etiqueta universal y una etiqueta de célula, en donde la etiqueta de muestra puede ser la misma para la pluralidad de códigos de barras estocásticos en el soporte sólido en donde las etiquetas universales son las mismas para la pluralidad de códigos de barras estocásticos en el soporte sólido y las etiquetas de célula pueden ser las mismas para la pluralidad de códigos de barras estocásticos en el soporte sólido. La etiqueta de muestra comprende de 5 a 20 nucleótidos. La etiqueta universal comprende de 5 a 20 nucleótidos. El marcador celular comprende de 5 a 20 nucleótidos.

[0024] En algunas formas de realización, las partículas sintéticas pueden ser perlas. Las perlas pueden ser perlas de gel de sílice, perlas de vidrio de poro controlado, perlas magnéticas, perlas Dynabeads, perlas de Sephadex/Sepharose, perlas de celulosa, perlas de poliestireno o cualquier combinación de las mismas.

[0025] En algunas formas de realización, la pluralidad de objetivos puede estar comprendida en una muestra. La muestra comprende una o más células. La muestra puede ser una sola célula. La una o más células comprenden uno o

más tipos de células. Al menos uno de uno o más tipos de células es una célula cerebral, una célula cardíaca, una célula cancerosa, una célula tumoral circulante, una célula orgánica, una célula epitelial, una célula metastásica, una célula benigna, una célula primaria, una célula circulatoria o cualquier combinación de las mismas.

[0026] En algunas formas de realización, la pluralidad de dianas comprende ácidos ribonucleicos (ARN), ARN mensajeros (ARNm), microARN, pequeños ARN interferentes (ARNip), productos de degradación de ARN, ARN que comprende cada uno una cola poli(A) o cualquier combinación de los mismos.

[0027] En algunas formas de realización, el método puede comprender además lisar una o más células. Lisar una o más células comprende calentar la muestra, poner en contacto la muestra con un detergente, cambiar el pH de la muestra o cualquier combinación de los mismos.

[0028] En este documento se describen métodos para determinar el número de objetivos. En algunas formas de realización, el método comprende: (a) codificar de barras estocásticamente una pluralidad de objetivos usando una pluralidad de códigos de barras estocásticos para crear una pluralidad de objetivos con códigos de barras estocásticos, en donde cada uno de la pluralidad de códigos de barras estocásticos comprende una etiqueta molecular; (b) obtener datos de secuenciación de los objetivos con códigos de barras estocásticos; y (c) para una o más de la pluralidad de dianas: (i) contar el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación; (ii) identificar grupos de etiquetas moleculares de la diana usando adyacencia direccional; (iii) colapsar los datos de secuenciación obtenidos en (b) usando los grupos de etiquetas moleculares de la diana identificada en (ii); y (iv) estimar el número de la diana, en donde el número de la diana estimado se correlaciona con el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación contados en (i) después de colapsar los datos de secuenciación en (ii). La pluralidad de dianas comprende dianas de todo el transcriptoma de una célula.

[0029] En algunas formas de realización, las etiquetas moleculares de la diana dentro de un grupo están dentro de un umbral de adyacencia direccional predeterminado entre sí. El umbral de adyacencia direccional es una distancia de Hamming de uno. Las etiquetas moleculares de la diana dentro del grupo comprenden uno o más etiquetas moleculares principales y etiquetas moleculares secundarios de uno o más etiquetas moleculares principales, en donde la aparición del marcador molecular principal es mayor o igual a un umbral de aparición de adyacencia direccional predeterminado. El umbral de ocurrencia de adyacencia direccional predeterminado puede tener el doble de aparición que una etiqueta molecular secundaria menos uno.

[0030] En algunas formas de realización, colapsar los datos de secuenciación obtenidos en (b) usando los grupos de etiquetas moleculares de la diana identificada en (ii) comprende: atribuir la aparición de la etiqueta molecular secundaria a la etiqueta molecular principal.

[0031] En algunas formas de realización, el método puede comprender además: determinar una profundidad de secuenciación del objetivo. Estimar el número del objetivo, si la profundidad de secuenciación del objetivo está por encima de un umbral de profundidad de secuenciación predeterminado, comprende ajustar los datos de secuenciación contados en (i). El umbral de profundidad de secuenciación predeterminado puede estar entre 15 y 20. Ajustar los datos de secuenciación contados en (i) comprende: establecer un umbral de etiquetas moleculares de la diana para determinar etiquetas moleculares verdaderas y etiquetas moleculares falsas asociadas con la diana en los datos de secuenciación obtenidos en (b). Establecer un umbral para las etiquetas moleculares de la diana comprende realizar un análisis estadístico de las etiquetas moleculares de la diana. Realizar el análisis estadístico comprende: ajustar la distribución de las etiquetas moleculares del objetivo y sus apariciones a dos distribuciones de Poisson; determinar el número de etiquetas moleculares verdaderas n usando las dos distribuciones de Poisson; y eliminar las etiquetas moleculares falsas de los datos de secuenciación obtenidos en (b), en el que las etiquetas moleculares falsas comprenden etiquetas moleculares con apariciones inferiores a la aparición del enésimo marcador molecular más abundante, y en el que las etiquetas moleculares verdaderas comprenden etiquetas moleculares con apariciones mayores mayor o igual a la aparición de la enésima etiqueta molecular más abundante.

[0032] En este documento se describen sistemas informáticos para determinar el número de objetivos. En algunas formas de realización, el sistema informático comprende: una memoria legible por computadora que almacena instrucciones ejecutables; y uno o más procesadores de computadora en comunicación con la memoria legible por computadora, en donde uno o más procesadores de computadora están programados mediante las instrucciones ejecutables para realizar (a) codificar de forma estocástica una pluralidad de objetivos usando una pluralidad de códigos de barras estocásticos para crear una pluralidad de códigos de barras estocásticos objetivos con códigos de barras, en los que cada uno de la pluralidad de códigos de barras estocásticos comprende una etiqueta molecular; (b) obtener datos de secuenciación de los objetivos con códigos de barras estocásticos; y (c) para una o más de la pluralidad de dianas: (i) contar el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación; (ii) determinar un estado de calidad de la diana en los datos de secuenciación obtenidos en (b); (iii) determinar uno o más errores de datos de secuenciación en los datos de secuenciación obtenidos en (b), en donde determinar uno o más errores de datos de secuenciación en los datos de secuenciación comprende determinar uno o más de: el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación, el estado de calidad de la diana en los datos de secuenciación y el número de etiquetas moleculares con secuencias distintas en la pluralidad de

códigos de barras estocásticos; y (iv) estimar el número de la diana, en donde el número de la diana estimado se correlaciona con el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación contados en (i) ajustados de acuerdo con uno o más errores de datos de secuenciación determinados en (iii). Pasos (i), (ii), (iii) y (iv) para cada uno de la pluralidad de objetivos. Los pasos (a), (b), (c), (i), (ii), (iii) y (iv) pueden multiplexarse.

[0033] En algunas formas de realización, las instrucciones ejecutables pueden programar además uno o más procesadores de computadora para realizar: colapsar los datos de secuenciación obtenidos en (b) antes de determinar uno o más errores de datos de secuenciación. Colapsar los datos de secuenciación obtenidos en (b) comprende: atribuir copias de la diana con etiquetas moleculares similares y con apariciones menores que un umbral de ocurrencia de colapso predeterminado como si tuvieran la misma etiqueta molecular para la pluralidad de dianas, en donde dos copias de una diana tienen etiquetas moleculares similares si las etiquetas moleculares de las dos copias de la diana difieren en al menos una base en la secuencia.

[0034] En algunas formas de realización, el umbral de aparición de colapso predeterminado puede ser 7 si los códigos de barras estocásticos comprenden aproximadamente 6561 etiquetas moleculares con secuencias distintas. El umbral de aparición de colapso predeterminado puede ser 17 si los códigos de barras estocásticos comprenden aproximadamente 65536 etiquetas moleculares con secuencias distintas. Dos copias de la diana tienen etiquetas moleculares similares si las etiquetas moleculares de las dos copias de la diana difieren en al menos una base en la secuencia. En algunas formas de realización, el marcador molecular comprende de 5 a 20 nucleótidos. Las etiquetas moleculares de diferentes códigos de barras estocásticos pueden ser diferentes entre sí. La pluralidad de códigos de barras estocásticos comprende aproximadamente 6561 etiquetas moleculares con secuencias distintas. La pluralidad de códigos de barras estocásticos comprende aproximadamente 65536 etiquetas moleculares con secuencias distintas.

[0035] En algunas formas de realización, los datos de secuenciación comprenden secuencias de la pluralidad de dianas con longitudes de lectura de 50 o más nucleótidos. Los datos de secuenciación comprenden secuencias de la pluralidad de objetivos con longitudes de lectura de 75 o más nucleótidos. Los datos de secuenciación comprenden secuencias de la pluralidad de objetivos con longitudes de lectura de 100 o más nucleótidos. Los datos de secuenciación obtenidos en (b) se pueden generar realizando una amplificación por reacción en hebra de la polimerasa (PCR) en la pluralidad de objetivos con códigos de barras estocásticos.

[0036] En algunas formas de realización, uno o más errores de datos de secuenciación pueden ser un error introducido por PCR, un error introducido por secuenciación, un error causado por contaminación de código de barras, un error de preparación de biblioteca o cualquier combinación de los mismos. El error introducido por la PCR puede ser el resultado de un error de amplificación por PCR, un sesgo de amplificación por PCR, una amplificación por PCR insuficiente o cualquier combinación de los mismos. El error introducido por la secuenciación puede ser el resultado de una llamada de base inexacta, una secuenciación insuficiente o cualquier combinación de los mismos.

[0037] En algunas formas de realización, las instrucciones ejecutables pueden programar además uno o más procesadores de computadora para determinar el estado de calidad del objetivo en los datos de secuenciación para que sea secuenciación completa, secuenciación incompleta o secuenciación saturada. El estado de calidad del objetivo en los datos de secuenciación se puede determinar por el número de etiquetas moleculares con secuencias distintas en la pluralidad de códigos de barras estocásticos y el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación contados. El estado de calidad del objetivo en los datos de secuenciación se puede clasificar como secuenciación incompleta si el estado de calidad del objetivo en los datos de secuenciación obtenidos en (b) no es una secuenciación completa ni una secuenciación saturada.

[0038] En algunas formas de realización, las instrucciones ejecutables pueden programar además uno o más procesadores de computadora para determinar el estado de calidad de secuenciación completa mediante un índice de dispersión relativo a la distribución de Poisson mayor o igual a un umbral de dispersión de secuenciación completa predeterminado, en donde el predeterminado el umbral de dispersión de secuenciación completa puede ser 0,9, 1 o 4. El estado de calidad de la secuenciación completa puede determinarse además mediante una etiqueta molecular con una aparición mayor o igual a un umbral de aparición de secuenciación completa predeterminado en los datos de secuenciación obtenidos en (b), en el que el umbral de aparición de secuenciación completa predeterminado puede ser 10 o 18.

[0039] En algunas formas de realización, las instrucciones ejecutables pueden programar además uno o más procesadores de computadora para determinar el estado de calidad de secuenciación saturada por el objetivo que tiene una cantidad de etiquetas moleculares con secuencias distintas que han sido mayores que un umbral de saturación predeterminado. El estado de calidad de la secuenciación saturada puede determinarse además mediante otra diana de la pluralidad de dianas que tiene una serie de etiquetas moleculares con secuencias distintas que sean mayores que el umbral de saturación predeterminado. El umbral de saturación predeterminado puede ser 6557 si los códigos de barras estocásticos comprenden aproximadamente 6561 etiquetas moleculares con secuencias distintas. El umbral de saturación predeterminado puede ser 65532 si los códigos de barras estocásticos comprenden aproximadamente 65536 etiquetas moleculares con secuencias distintas.

[0040] En algunas formas de realización, las instrucciones ejecutables pueden programar además uno o más procesadores de computadora para ajustar en (iv) el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación contados en (i) por, si el objetivo tiene el estado de calidad de secuenciación completo, determinando todas las etiquetas moleculares secundarias para una o más etiquetas moleculares principales; realizar un primer análisis estadístico para al menos una etiqueta molecular secundaria y la etiqueta molecular principal; y atribuir la aparición de la etiqueta molecular secundaria a la etiqueta molecular principal si se acepta la hipótesis nula del primer análisis estadístico.

[0041] En algunas formas de realización, uno o más etiquetas moleculares originales comprenden etiquetas moleculares con apariciones mayores o iguales a un umbral principal de secuenciación completa predeterminado, en donde el umbral principal de secuenciación completa predeterminado es igual al umbral de aparición de secuenciación completa predeterminada. Las etiquetas moleculares secundarias comprenden etiquetas moleculares que difieren de la etiqueta molecular principal en una base y tienen apariciones menores o iguales a un umbral secundario de secuenciación completa predeterminado, en donde el umbral secundario de secuenciación completa predeterminado puede ser 3 o 5. La hipótesis nula del primer análisis estadístico puede aceptarse si la probabilidad de que la hipótesis nula sea verdadera está por debajo de las tasas de descubrimiento falso, donde la tasa de descubrimiento falso es del 5 % o del 10 %. El primer análisis estadístico puede ser una prueba binomial múltiple.

[0042] En algunas formas de realización, las instrucciones ejecutables pueden programar además uno o más procesadores de computadora para ajustar en (iv) el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación contados en (i) por, si el La diana tiene el estado de calidad de secuenciación completo, umbralizando las etiquetas moleculares de la diana para determinar las etiquetas moleculares verdaderas y las etiquetas moleculares falsas asociadas con la diana en los datos de secuenciación obtenidos en (b). Establecer un umbral para las etiquetas moleculares de la diana comprende realizar un segundo análisis estadístico sobre las etiquetas moleculares de la diana.

[0043] En algunas formas de realización, las instrucciones ejecutables pueden programar además uno o más procesadores de computadora para realizar el segundo análisis estadístico: ajustando la distribución de las etiquetas moleculares del objetivo y sus apariciones a dos distribuciones de Poisson; determinar el número de etiquetas moleculares verdaderas n usando las dos distribuciones de Poisson; y eliminar las etiquetas moleculares falsas de los datos de secuenciación obtenidos en (b), en el que las etiquetas moleculares falsas comprenden etiquetas moleculares con apariciones inferiores a la aparición del enésimo marcador molecular más abundante, y en el que las etiquetas moleculares verdaderas comprenden etiquetas moleculares con apariciones mayores mayor o igual a la aparición de la enésima etiqueta molecular más abundante. Las dos distribuciones de Poisson comprenden una primera distribución de Poisson correspondiente a las etiquetas moleculares verdaderas y una segunda distribución de Poisson para las etiquetas moleculares falsas.

[0044] En algunas formas de realización, las instrucciones ejecutables pueden programar además uno o más procesadores de computadora para ajustar en (iv) el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación contados en (i) por, si el estado de calidad del objetivo en los datos de secuenciación obtenidos en (b) es el estado de calidad de la secuenciación incompleta, determinando si el objetivo tiene ruido en los datos de secuenciación obtenidos en (b); y eliminar el objetivo ruidoso de los datos de secuenciación obtenidos en (b). La diana puede ser ruidosa si la aparición de las etiquetas moleculares de las dianas ruidosas es menor o igual a un umbral de diana ruidosa de secuenciación incompleta, en donde el umbral del gen ruidoso de secuenciación incompleta es 5. El umbral de diana ruidosa de secuenciación incompleta puede igualar la mediana o ocurrencia media de las etiquetas moleculares de la pluralidad de dianas con estados de calidad de secuenciación completa.

[0045] En algunas formas de realización, las instrucciones ejecutables pueden programar además uno o más procesadores de computadora para ajustar en (iv) el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación contados en (i) por, si el estado de calidad del objetivo en los datos de secuenciación obtenidos en (b) es el estado de calidad de secuenciación incompleta, umbralizando las etiquetas moleculares del objetivo para determinar las etiquetas moleculares verdaderas y las etiquetas moleculares falsas en los datos de secuenciación obtenidos en (b).

[0046] En algunas formas de realización, las instrucciones ejecutables pueden programar además uno o más procesadores informáticos para establecer un umbral para las etiquetas moleculares de la diana realizando un tercer análisis estadístico en las etiquetas moleculares. Realizar el tercer análisis estadístico de las etiquetas moleculares comprende: determinar el número de etiquetas moleculares verdaderas n usando un modelo de Poisson truncado en cero; y eliminar las etiquetas moleculares falsas de los datos de secuenciación obtenidos en (b), en el que las etiquetas moleculares falsas comprenden etiquetas moleculares con apariciones inferiores a la aparición del enésimo marcador molecular más abundante, y en el que las etiquetas moleculares verdaderas comprenden etiquetas moleculares con apariciones mayores mayor o igual a la aparición de la enésima etiqueta molecular más abundante.

[0047] En algunas formas de realización, al menos el 50 % o el 80 % de las etiquetas moleculares en los datos de secuenciación obtenidos en (b) se pueden retener después de que los datos de secuenciación contados en (i) se ajusten de acuerdo con uno o más errores de datos de secuenciación determinados en (iii).

[0048] En el presente documento se describen sistemas informáticos para determinar el número de objetivos. En algunas formas de realización, el sistema informático comprende: una memoria legible por computadora que almacena instrucciones ejecutables; y uno o más procesadores de computadora en comunicación con la memoria legible por computadora, en donde uno o más procesadores de computadora están programados mediante las instrucciones ejecutables para realizar: (a) codificar de barras estocásticamente una pluralidad de objetivos usando una pluralidad de códigos de barras estocásticos para crear una pluralidad de objetivos con códigos de barras estocásticos, en los que cada uno de la pluralidad de códigos de barras estocásticos comprende una etiqueta molecular; (b) obtener datos de secuenciación de los objetivos con códigos de barras estocásticos; y (c) para una o más de la pluralidad de dianas: (i) contar el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación; (ii) identificar grupos de etiquetas moleculares de la diana usando adyacencia direccional; (iii) colapsar los datos de secuenciación obtenidos en (b) usando los grupos de etiquetas moleculares de la diana identificada en (ii); y (iv) estimar el número de la diana, en donde el número de la diana estimado se correlaciona con el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación contados en (i) después de colapsar los datos de secuenciación en (ii). La pluralidad de dianas comprende dianas de todo el transcriptoma de una célula.

[0049] En algunas formas de realización, las etiquetas moleculares de la diana dentro de un grupo están dentro de un umbral de adyacencia direccional predeterminado entre sí. El umbral de adyacencia direccional es una distancia de Hamming de uno. Las etiquetas moleculares de la diana dentro del grupo comprenden uno o más etiquetas moleculares principales y etiquetas moleculares secundarios de uno o más etiquetas moleculares principales, en donde la aparición del marcador molecular principal es mayor o igual a un umbral de aparición de adyacencia direccional predeterminado. El umbral de aparición de adyacencia direccional predeterminado puede ser el doble de la aparición de una etiqueta molecular secundaria menos uno.

[0050] En algunas formas de realización, colapsar los datos de secuenciación obtenidos en (b) usando los grupos de etiquetas moleculares de la diana identificada en (ii) comprende: atribuir la aparición de la etiqueta molecular secundaria a la etiqueta molecular principal.

[0051] En algunas formas de realización, las instrucciones ejecutables pueden programar además uno o más procesadores de computadora para: determinar una profundidad de secuenciación del objetivo. Estimar el número del objetivo, si la profundidad de secuenciación del objetivo está por encima de un umbral de profundidad de secuenciación predeterminado, comprende ajustar los datos de secuenciación contados en (i). El umbral de profundidad de secuenciación predeterminado puede estar entre 15 y 20. Ajustar los datos de secuenciación contados en (i) comprende: establecer un umbral de etiquetas moleculares de la diana para determinar etiquetas moleculares verdaderas y etiquetas moleculares falsas asociadas con la diana en los datos de secuenciación obtenidos en (b). Establecer un umbral para las etiquetas moleculares de la diana comprende realizar un análisis estadístico de las etiquetas moleculares de la diana. Realizar el análisis estadístico comprende: ajustar la distribución de las etiquetas moleculares del objetivo y sus apariciones a dos distribuciones de Poisson; determinar el número de etiquetas moleculares verdaderas n usando las dos distribuciones de Poisson; y eliminar las etiquetas moleculares falsas de los datos de secuenciación obtenidos en (b), en el que las etiquetas moleculares falsas comprenden etiquetas moleculares con apariciones inferiores a la aparición del enésimo marcador molecular más abundante, y en el que las etiquetas moleculares verdaderas comprenden etiquetas moleculares con apariciones mayores mayor o igual a la aparición de la enésima etiqueta molecular más abundante.

[0052] En el presente documento se divulgan uno o más medios legibles por computadora no transitorios que comprenden códigos ejecutables que, cuando se ejecutan, hacen que uno o más dispositivos informáticos determinen el número de objetivos. En algunas formas de realización, los códigos ejecutables, cuando se ejecutan, hacen que uno o más dispositivos informáticos realicen un proceso que comprende: (a) codificar con barras estocásticamente una pluralidad de objetivos usando una pluralidad de códigos de barras estocásticos para crear una pluralidad de objetivos con códigos de barras estocásticos, en donde cada de la pluralidad de códigos de barras estocásticos comprende una etiqueta molecular; (b) obtener datos de secuenciación de los objetivos con códigos de barras estocásticos; y (c) para una o más de la pluralidad de dianas: (i) contar el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación; (ii) determinar un estado de calidad de la diana en los datos de secuenciación obtenidos en (b); (iii) determinar uno o más errores de datos de secuenciación en los datos de secuenciación obtenidos en (b), en donde determinar uno o más errores de datos de secuenciación en los datos de secuenciación comprende determinar uno o más de: el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación, el estado de calidad de la diana en los datos de secuenciación y el número de etiquetas moleculares con secuencias distintas en la pluralidad de códigos de barras estocásticos; y (iv) estimar el número de la diana, en donde el número de la diana estimado se correlaciona con el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación contados en (i) ajustados de acuerdo con uno o más errores de datos de secuenciación determinados en (iii). Los pasos (i), (ii), (iii) y (iv) se pueden realizar para cada uno de la pluralidad de objetivos. El método puede ser multiplexado.

[0053] En algunas formas de realización, el proceso comprende además: colapsar los datos de secuenciación obtenidos en (b) antes de determinar uno o más errores de datos de secuenciación. Colapsar los datos de secuenciación obtenidos en (b) comprende: atribuir copias de la diana con etiquetas moleculares similares y con apariciones menores que un umbral de ocurrencia de colapso predeterminado como si tuvieran la misma etiqueta molecular para la pluralidad de

dianas, en donde dos copias de una diana tienen etiquetas moleculares similares si las etiquetas moleculares de las dos copias de la diana difieren en al menos una base en la secuencia.

[0054] En algunas formas de realización, el umbral de aparición de colapso predeterminado puede ser 7 si los códigos de barras estocásticos comprenden aproximadamente 6561 etiquetas moleculares con secuencias distintas. El umbral de aparición de colapso predeterminado puede ser 17 si los códigos de barras estocásticos comprenden aproximadamente 65536 etiquetas moleculares con secuencias distintas. Dos copias de la diana tienen etiquetas moleculares similares si las etiquetas moleculares de las dos copias de la diana difieren en al menos una base en la secuencia. En algunas formas de realización, el marcador molecular comprende de 5 a 20 nucleótidos. Las etiquetas moleculares de diferentes códigos de barras estocásticos pueden ser diferentes entre sí. La pluralidad de códigos de barras estocásticos comprende aproximadamente 6561 etiquetas moleculares con secuencias distintas. La pluralidad de códigos de barras estocásticos comprende aproximadamente 65536 etiquetas moleculares con secuencias distintas.

[0055] En algunas formas de realización, los datos de secuenciación comprenden secuencias de la pluralidad de dianas con longitudes de lectura de 50 o más nucleótidos. Los datos de secuenciación comprenden secuencias de la pluralidad de objetivos con longitudes de lectura de 75 o más nucleótidos. Los datos de secuenciación comprenden secuencias de la pluralidad de objetivos con longitudes de lectura de 100 o más nucleótidos. Los datos de secuenciación obtenidos en (b) se pueden generar realizando una amplificación por reacción en hebra de la polimerasa (PCR) en la pluralidad de objetivos con códigos de barras estocásticos.

[0056] En algunas formas de realización, uno o más errores de datos de secuenciación pueden ser un error introducido por PCR, un error introducido por secuenciación, un error causado por contaminación de código de barras, un error de preparación de biblioteca o cualquier combinación de los mismos. El error introducido por la PCR puede ser el resultado de un error de amplificación por PCR, un sesgo de amplificación por PCR, una amplificación por PCR insuficiente o cualquier combinación de los mismos. El error introducido por la secuenciación puede ser el resultado de una llamada de base inexacta, una secuenciación insuficiente o cualquier combinación de los mismos.

[0057] En algunas formas de realización, el estado de calidad de la diana en los datos de secuenciación puede ser secuenciación completa, secuenciación incompleta o secuenciación saturada. El estado de calidad de la diana en los datos de secuenciación se puede determinar mediante el número de etiquetas moleculares con secuencias distintas en la pluralidad de códigos de barras estocásticos y la cantidad de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación contados. El estado de calidad del objetivo en los datos de secuenciación se puede clasificar como secuenciación incompleta si el estado de calidad del objetivo en los datos de secuenciación obtenidos en (b) no es una secuenciación completa ni una secuenciación saturada.

[0058] En algunas formas de realización, el estado de calidad de la secuenciación completa se puede determinar mediante un índice de dispersión relativo a la distribución de Poisson mayor o igual a un umbral de dispersión de secuenciación completa predeterminado, en donde el umbral de dispersión de secuenciación completa predeterminado puede ser 0,9, 1 o 4. El estado de calidad de la secuenciación completa puede determinarse además mediante una etiqueta molecular con una aparición mayor o igual a un umbral de aparición de secuenciación completa predeterminado en los datos de secuenciación obtenidos en (b), en donde el umbral de aparición de secuenciación completa predeterminado puede ser 10 o 18.

[0059] En algunas formas de realización, el estado de calidad de la secuenciación saturada se puede determinar si la diana tiene un número de etiquetas moleculares con secuencias distintas que son mayores que un umbral de saturación predeterminado. El estado de calidad de la secuenciación saturada puede determinarse además mediante otra diana de la pluralidad de dianas que tiene una serie de etiquetas moleculares con secuencias distintas que sean mayores que el umbral de saturación predeterminado. El umbral de saturación predeterminado puede ser 6557 si los códigos de barras estocásticos comprenden aproximadamente 6561 etiquetas moleculares con secuencias distintas. El umbral de saturación predeterminado puede ser 65532 si los códigos de barras estocásticos comprenden aproximadamente 65536 etiquetas moleculares con secuencias distintas.

[0060] En algunas formas de realización, el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación contados en (i) se ajusta en (iv), si la diana tiene el estado de calidad de secuenciación completo, determinando todas las etiquetas moleculares secundarias para una o más etiquetas moleculares originales; realizar un primer análisis estadístico para al menos una etiqueta molecular secundaria y la etiqueta molecular principal; y atribuir la aparición de la etiqueta molecular secundaria a la etiqueta molecular principal si se acepta la hipótesis nula del primer análisis estadístico.

[0061] En algunas formas de realización, uno o más etiquetas moleculares originales comprenden etiquetas moleculares con apariciones mayores o iguales a un umbral principal de secuenciación completa predeterminado, en donde el umbral principal de secuenciación completa predeterminado es igual al umbral de aparición de secuenciación completa predeterminada. Las etiquetas moleculares secundarias comprenden etiquetas moleculares que difieren de la etiqueta molecular principal en una base y tienen apariciones menores o iguales a un umbral secundario de secuenciación completa predeterminado, en donde el umbral secundario de secuenciación completa predeterminado puede ser 3 o 5. La hipótesis nula del primer análisis estadístico puede aceptarse si la probabilidad de que la hipótesis nula sea verdadera

está por debajo de las tasas de descubrimiento falso, donde la tasa de descubrimiento falso es del 5 % o del 10 %. El primer análisis estadístico puede ser una prueba binomial múltiple.

[0062] En algunas formas de realización, el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación contados en (i) se ajustan en (iv), si la diana tiene el estado de calidad de secuenciación completo, umbralizando las etiquetas moleculares de la diana para determinar las etiquetas moleculares verdaderas y las etiquetas moleculares falsas asociadas con la diana en los datos de secuenciación obtenidos en (b). Establecer un umbral para las etiquetas moleculares de la diana comprende realizar un segundo análisis estadístico sobre las etiquetas moleculares de la diana.

[0063] En algunas formas de realización, realizar el segundo análisis estadístico comprende: ajustar la distribución de las etiquetas moleculares de la diana y sus apariciones a dos distribuciones de Poisson; determinar el número de etiquetas moleculares verdaderas n usando las dos distribuciones de Poisson; y eliminar las etiquetas moleculares falsas de los datos de secuenciación obtenidos en (b), en el que las etiquetas moleculares falsas comprenden etiquetas moleculares con apariciones inferiores a la aparición del enésimo marcador molecular más abundante, y en el que las etiquetas moleculares verdaderas comprenden etiquetas moleculares con apariciones mayores mayor o igual a la aparición de la enésima etiqueta molecular más abundante. Las dos distribuciones de Poisson comprenden una primera distribución de Poisson correspondiente a las etiquetas moleculares verdaderas y una segunda distribución de Poisson para las etiquetas moleculares falsas.

[0064] En algunas formas de realización, el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación contados en (i) se puede ajustar en (iv) si el estado de calidad de la diana en los datos de secuenciación obtenidos en (b) es el estado de calidad de la secuenciación incompleta, que determina si el objetivo tiene ruido en los datos de secuenciación obtenidos en (b); y eliminar el objetivo ruidoso de los datos de secuenciación obtenidos en (b). La diana puede ser ruidosa si la aparición de las etiquetas moleculares de las dianas ruidosas es menor o igual a un umbral de diana ruidosa de secuenciación incompleta, en donde el umbral del gen ruidoso de secuenciación incompleta es 5. El umbral de diana ruidosa de secuenciación incompleta puede igualar la mediana o ocurrencia media de las etiquetas moleculares de la pluralidad de dianas con estados de calidad de secuenciación completa.

[0065] En algunas formas de realización, el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación contados en (i) se puede ajustar en (iv) si el estado de calidad de la diana en los datos de secuenciación obtenidos en (b) es el estado de calidad de la secuenciación incompleta, que establece un umbral para las etiquetas moleculares del objetivo para determinar las etiquetas moleculares verdaderas y las etiquetas moleculares falsas en los datos de secuenciación obtenidos en (b).

[0066] En algunas formas de realización, establecer un umbral para las etiquetas moleculares de la diana comprende realizar un tercer análisis estadístico en las etiquetas moleculares. Realizar el tercer análisis estadístico de las etiquetas moleculares comprende: determinar el número de etiquetas moleculares verdaderas n usando un modelo de Poisson truncado en cero; y eliminar las etiquetas moleculares falsas de los datos de secuenciación obtenidos en (b), en el que las etiquetas moleculares falsas comprenden etiquetas moleculares con apariciones inferiores a la aparición del enésimo marcador molecular más abundante, y en el que las etiquetas moleculares verdaderas comprenden etiquetas moleculares con apariciones mayores mayor o igual a la aparición de la enésima etiqueta molecular más abundante.

[0067] En algunas formas de realización, al menos el 50 % o el 80 % de las etiquetas moleculares en los datos de secuenciación obtenidos en (b) se pueden retener después de que los datos de secuenciación contados en (i) se ajusten de acuerdo con uno o más errores de datos de secuenciación determinados en (iii).

[0068] En el presente documento se divulgan uno o más medios legibles por computadora no transitorios que comprenden códigos ejecutables que, cuando se ejecutan, hacen que uno o más dispositivos informáticos determinen el número de objetivos. En algunas formas de realización, los códigos ejecutables, cuando se ejecutan, hacen que uno o más dispositivos informáticos realicen un proceso que comprende: (a) codificar con barras estocásticamente una pluralidad de objetivos usando una pluralidad de códigos de barras estocásticos para crear una pluralidad de objetivos con códigos de barras estocásticos, en donde cada de la pluralidad de códigos de barras estocásticos comprende una etiqueta molecular; (b) obtener datos de secuenciación de los objetivos con códigos de barras estocásticos; y (c) para una o más de la pluralidad de dianas: (i) contar el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación; (ii) identificar grupos de etiquetas moleculares de la diana usando adyacencia direccional; (iii) colapsar los datos de secuenciación obtenidos en (b) usando los grupos de etiquetas moleculares de la diana identificada en (ii); y (iv) estimar el número de la diana, en donde el número de la diana estimado se correlaciona con el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación contados en (i) después de colapsar los datos de secuenciación en (ii). La pluralidad de dianas comprende dianas de todo el transcriptoma de una célula.

[0069] En algunas formas de realización, las etiquetas moleculares de la diana dentro de un grupo están dentro de un umbral de adyacencia direccional predeterminado entre sí. El umbral de adyacencia direccional es una distancia de Hamming de uno. Las etiquetas moleculares de la diana dentro del grupo comprenden uno o más etiquetas moleculares

principales y etiquetas moleculares secundarios de uno o más etiquetas moleculares principales, en donde la aparición del marcador molecular principal es mayor o igual a un umbral de aparición de adyacencia direccional predeterminado. El umbral de aparición de adyacencia direccional predeterminado puede ser el doble de la aparición de una etiqueta molecular secundaria menos uno.

[0070] En algunas formas de realización, colapsar los datos de secuenciación obtenidos en (b) usando los grupos de etiquetas moleculares de la diana identificada en (ii) comprende: atribuir la aparición de la etiqueta molecular secundaria a la etiqueta molecular principal.

[0071] En algunas formas de realización, el método puede comprender además: determinar una profundidad de secuenciación del objetivo. Estimar el número del objetivo, si la profundidad de secuenciación del objetivo está por encima de un umbral de profundidad de secuenciación predeterminado, comprende ajustar los datos de secuenciación contados en (i). El umbral de profundidad de secuenciación predeterminado puede estar entre 15 y 20. Ajustar los datos de secuenciación contados en (i) comprende: establecer un umbral de etiquetas moleculares de la diana para determinar etiquetas moleculares verdaderas y etiquetas moleculares falsas asociadas con la diana en los datos de secuenciación obtenidos en (b). Establecer un umbral para las etiquetas moleculares de la diana comprende realizar un análisis estadístico de las etiquetas moleculares de la diana. Realizar el análisis estadístico comprende: ajustar la distribución de las etiquetas moleculares del objetivo y sus apariciones a dos distribuciones de Poisson; determinar el número de etiquetas moleculares verdaderas n usando las dos distribuciones de Poisson; y eliminar las etiquetas moleculares falsas de los datos de secuenciación obtenidos en (b), en el que las etiquetas moleculares falsas comprenden etiquetas moleculares con apariciones inferiores a la aparición del enésimo marcador molecular más abundante, y en el que las etiquetas moleculares verdaderas comprenden etiquetas moleculares con apariciones mayores mayor o igual a la aparición de la enésima etiqueta molecular más abundante.

[0072] En el presente documento se describen métodos para corregir errores de PCR o secuenciación. En algunas formas de realización, el método puede comprender: (a) obtener datos de secuenciación de los objetivos con códigos de barras estocásticos; y (b) para una o más de la pluralidad de dianas: (i) contar el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación; (ii) identificar grupos de etiquetas moleculares de la diana usando adyacencia direccional; (iii) colapsar los datos de secuenciación obtenidos en (a) usando los grupos de etiquetas moleculares de la diana identificada en (ii); y (iv) estimar el número de la diana, en donde el número de la diana estimado se correlaciona con el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación contados en (i) después de colapsar los datos de secuenciación en (ii). La pluralidad de dianas comprende dianas de todo el transcriptoma de una célula. En algunas formas de realización, el método se puede utilizar para determinar el número de objetivos. El método puede comprender además (c) codificar de barras estocásticamente la pluralidad de objetivos usando la pluralidad de códigos de barras estocásticos para crear la pluralidad de objetivos con códigos de barras estocásticos; y (d) secuenciar los objetivos con códigos de barras estocásticos para generar los datos de secuenciación de los objetivos con códigos de barras estocásticos recibidos.

[0073] En algunas formas de realización, las etiquetas moleculares de la diana dentro de un grupo están dentro de un umbral de adyacencia direccional predeterminado entre sí. El umbral de adyacencia direccional es una distancia de Hamming de uno. Las etiquetas moleculares de la diana dentro del grupo comprenden uno o más etiquetas moleculares principales y etiquetas moleculares secundarios de uno o más etiquetas moleculares principales, en donde la aparición del marcador molecular principal es mayor o igual a un umbral de aparición de adyacencia direccional predeterminado. El umbral de aparición de adyacencia direccional predeterminado puede ser el doble de la aparición de una etiqueta molecular secundaria menos uno.

[0074] En algunas formas de realización, colapsar los datos de secuenciación obtenidos en (b) usando los grupos de etiquetas moleculares de la diana identificada en (ii) comprende: atribuir la aparición de la etiqueta molecular secundaria a la etiqueta molecular original.

[0075] En algunas formas de realización, el método comprende además: determinar una profundidad de secuenciación del objetivo. Estimar el número del objetivo, si la profundidad de secuenciación del objetivo está por encima de un umbral de profundidad de secuenciación predeterminado, comprende ajustar los datos de secuenciación contados en (i). El umbral de profundidad de secuenciación predeterminado puede estar entre 15 y 20. Ajustar los datos de secuenciación contados en (i) comprende: establecer un umbral de etiquetas moleculares de la diana para determinar etiquetas moleculares verdaderas y etiquetas moleculares falsas asociadas con la diana en los datos de secuenciación obtenidos en (b). Establecer un umbral para las etiquetas moleculares de la diana comprende realizar un análisis estadístico de las etiquetas moleculares de la diana. Realizar el análisis estadístico comprende: ajustar la distribución de las etiquetas moleculares del objetivo y sus apariciones a dos distribuciones binomiales negativas; determinar el número de etiquetas moleculares verdaderas n usando las dos distribuciones binomiales negativas; y eliminar las etiquetas moleculares falsas de los datos de secuenciación obtenidos en (b), en el que las etiquetas moleculares falsas comprenden etiquetas moleculares con apariciones inferiores a la aparición del enésimo marcador molecular más abundante, y en el que las etiquetas moleculares verdaderas comprenden etiquetas moleculares con apariciones mayores mayor o igual a la aparición de la enésima etiqueta molecular más abundante.

[0076] En el presente documento se describen sistemas informáticos para determinar el número de objetivos. En algunas formas de realización, el sistema informático comprende: una memoria legible por computadora que almacena instrucciones ejecutables; y uno o más procesadores de computadora en comunicación con la memoria legible por computadora, en donde uno o más procesadores de computadora están programados mediante las instrucciones ejecutables para realizar: (a) codificar de barras estocásticamente una pluralidad de objetivos usando una pluralidad de códigos de barras estocásticos para crear una pluralidad de objetivos con códigos de barras estocásticos, en los que cada uno de la pluralidad de códigos de barras estocásticos comprende una etiqueta molecular; (b) obtener datos de secuenciación de los objetivos con códigos de barras estocásticos; y (c) para una o más de la pluralidad de dianas: (i) contar el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación; (ii) identificar grupos de etiquetas moleculares de la diana usando adyacencia direccional; (iii) colapsar los datos de secuenciación obtenidos en (b) usando los grupos de etiquetas moleculares de la diana identificada en (ii); y (iv) estimar el número de la diana, en donde el número de la diana estimado se correlaciona con el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación contados en (i) después de colapsar los datos de secuenciación en (ii). La pluralidad de dianas comprende dianas de todo el transcriptoma de una célula.

[0077] En algunas formas de realización, las etiquetas moleculares de la diana dentro de un grupo están dentro de un umbral de adyacencia direccional predeterminado entre sí. El umbral de adyacencia direccional es una distancia de Hamming de uno. Las etiquetas moleculares de la diana dentro del grupo comprenden uno o más etiquetas moleculares principales y etiquetas moleculares secundarios de uno o más etiquetas moleculares principales, en donde la aparición del marcador molecular principal es mayor o igual a un umbral de aparición de adyacencia direccional predeterminado. El umbral de aparición de adyacencia direccional predeterminado puede ser el doble de la aparición de una etiqueta molecular secundaria menos uno.

[0078] En algunas formas de realización, colapsar los datos de secuenciación obtenidos en (b) usando los grupos de etiquetas moleculares de la diana identificada en (ii) comprende: atribuir la aparición de la etiqueta molecular secundaria a la etiqueta molecular principal.

[0079] En algunas formas de realización, las instrucciones ejecutables programan además uno o más procesadores de computadora para: determinar una profundidad de secuenciación del objetivo. Estimar el número del objetivo, si la profundidad de secuenciación del objetivo está por encima de un umbral de profundidad de secuenciación predeterminado, comprende ajustar los datos de secuenciación contados en (i). El umbral de profundidad de secuenciación predeterminado puede estar entre 15 y 20. Ajustar los datos de secuenciación contados en (i) comprende: establecer un umbral de etiquetas moleculares de la diana para determinar etiquetas moleculares verdaderas y etiquetas moleculares falsas asociadas con la diana en los datos de secuenciación obtenidos en (b). Establecer un umbral para las etiquetas moleculares de la diana comprende realizar un análisis estadístico de las etiquetas moleculares de la diana. Realizar el análisis estadístico comprende: ajustar la distribución de las etiquetas moleculares del objetivo y sus apariciones a dos distribuciones binomiales negativas; determinar el número de etiquetas moleculares verdaderas n usando las dos distribuciones binomiales negativas; y eliminar las etiquetas moleculares falsas de los datos de secuenciación obtenidos en (b), en el que las etiquetas moleculares falsas comprenden etiquetas moleculares con apariciones inferiores a la aparición del enésimo marcador molecular más abundante, y en el que las etiquetas moleculares verdaderas comprenden etiquetas moleculares con apariciones mayores mayor o igual a la aparición de la enésima etiqueta molecular más abundante.

[0080] En el presente documento se describen métodos para corregir errores de PCR o secuenciación. En algunas formas de realización, el método comprende: (a) obtener datos de secuenciación de los objetivos con códigos de barras estocásticos; y (b) para uno o más de la pluralidad de objetivos: (i) contar el número de etiquetas moleculares con secuencias distintas asociadas con el objetivo en los datos de secuenciación; (ii) determinar una serie de etiquetas moleculares de ruido con secuencias distintas asociadas con la diana en los datos de secuenciación; y (iii) estimar el número de la diana, en donde el número de la diana estimado se correlaciona con el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación contados en (i) ajustados de acuerdo con la cantidad de etiquetas moleculares de ruido determinado en (ii). En algunas formas de realización, el método comprende además determinar un estado de secuenciación del objetivo en los datos de secuenciación. El estado de secuenciación del objetivo en los datos de secuenciación es secuenciación saturada, subsecuenciación o sobresecuenciación. En algunas formas de realización, el método se puede utilizar para determinar el número de objetivos. El método puede comprender además (c) codificar de barras estocásticamente la pluralidad de objetivos usando la pluralidad de códigos de barras estocásticos para crear la pluralidad de objetivos con códigos de barras estocásticos; y (d) secuenciar los objetivos con códigos de barras estocásticos para generar los datos de secuenciación de los objetivos con códigos de barras estocásticos recibidos.

[0081] En algunas formas de realización, el estado de secuenciación saturada se determina porque la diana tiene una cantidad de etiquetas moleculares con secuencias distintas mayores que un umbral de saturación predeterminado. El umbral de saturación predeterminado es aproximadamente 6557 si los códigos de barras estocásticos comprenden aproximadamente 6561 etiquetas moleculares con secuencias distintas. El umbral de saturación predeterminado es aproximadamente 65532 si los códigos de barras estocásticos comprenden aproximadamente 65536 etiquetas moleculares con secuencias distintas. Si el estado de secuenciación del objetivo en los datos de secuenciación es el estado de secuenciación saturado, el número de etiquetas moleculares de ruido determinado en (ii) es cero.

[0082] En algunas formas de realización, el estado de subsecuenciación puede ser determinado por el objetivo que tiene una profundidad (por ejemplo, una profundidad promedio, mínima o máxima) menor que un umbral de subsecuenciación predeterminado. El umbral de secuenciación insuficiente es de aproximadamente cuatro. El umbral de secuenciación insuficiente puede ser independiente del número de etiquetas moleculares con secuencias distintas. Si el estado de secuenciación del objetivo en los datos de secuenciación es el estado de secuenciación insuficiente, el número de etiquetas moleculares de ruido determinado en (ii) es cero.

[0083] En algunas formas de realización, el estado de sobresecuenciación se determina porque la diana tiene una cantidad de etiquetas moleculares con secuencias distintas mayores que un umbral de sobresecuenciación predeterminado. Por ejemplo, el umbral de sobresecuenciación puede ser de aproximadamente 250 si los códigos de barras estocásticos comprenden aproximadamente 6561 etiquetas moleculares con secuencias distintas. El método comprende, si el estado de secuenciación de la diana en los datos de secuenciación es el estado de sobresecuenciación: submuestrear el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación hasta aproximadamente el umbral de sobresecuenciación predeterminado.

[0084] En algunas formas de realización, determinar el número de etiquetas moleculares de ruido con distintas secuencias asociadas con la diana en los datos de secuenciación comprende: si se satisface una condición de ajuste de distribución binomial negativa, (iv) ajustar una distribución binomial negativa de señal al número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación contados en (i), en donde la distribución binomial de señal negativa corresponde a una serie de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación contados en (i) que son etiquetas moleculares de señal; (v) ajustar una distribución binomial de ruido negativo al número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación contados en (i), en donde la distribución binomial de ruido negativo corresponde a una cantidad de etiquetas moleculares con secuencias distintas asociadas con el objetivo en los datos de secuenciación contados en (i) son etiquetas moleculares de ruido; y (vi) determinar el número de etiquetas moleculares de ruido utilizando la distribución binomial negativa de señal ajustada en (v) y la distribución binomial negativa de ruido ajustada en (vi).

[0085] En algunas formas de realización, la condición de ajuste de distribución binomial negativa comprende: el estado de secuenciación del objetivo en los datos de secuenciación no es el estado de secuenciación insuficiente o el estado de secuenciación excesiva. Determinación del número de etiquetas moleculares de ruido utilizando la distribución binomial de señal negativa ajustada en (v) y la distribución binomial de ruido negativo ajustada en (vi) comprende: para cada una de las distintas secuencias asociadas con el objetivo en los datos de secuenciación: determinar una probabilidad de señal de que la secuencia distinta esté en la distribución binomial de señal negativa; determinar una probabilidad de ruido de la secuencia distinta de estar en la distribución binomial de ruido negativo; y determinar que la secuencia distinta es una etiqueta molecular de ruido si la probabilidad de la señal es menor que la probabilidad del ruido.

[0086] En algunas formas de realización, determinar el número de etiquetas moleculares de ruido con distintas secuencias asociadas con el objetivo en los datos de secuenciación comprende: agregar pseudopuntos al número de etiquetas moleculares con secuencias distintas asociadas con el objetivo en los datos de secuenciación antes de determinar el número de etiquetas moleculares de ruido con secuencias distintas asociadas con el objetivo en los datos de secuenciación en (ii), si el estado de secuenciación de la diana en los datos de secuenciación no es el estado de secuenciación insuficiente o el estado de secuenciación excesiva y el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación contados en (i) es menor que un umbral de pseudopuntos. El umbral de pseudopuntos es diez.

[0087] En algunas formas de realización, determinar el número de etiquetas moleculares de ruido con distintas secuencias asociadas con la diana en los datos de secuenciación comprende: eliminar etiquetas moleculares no únicas al determinar el número de etiquetas moleculares de ruido con secuencias distintas asociadas con la diana en los datos de secuenciación en (ii), si el estado de secuenciación de la diana en los datos de secuenciación no es el estado de secuenciación insuficiente o el estado de secuenciación excesiva y el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación contados en (i) no es inferior a un umbral de pseudopuntos.

[0088] En algunas formas de realización, eliminar las etiquetas moleculares no únicos comprende eliminar las etiquetas moleculares no únicos al determinar el número de etiquetas moleculares de ruido con secuencias distintas asociadas con la diana en los datos de secuenciación en (ii) si el número de etiquetas moleculares etiquetas con secuencias distintas asociadas con la diana en los datos de secuenciación es mayor que un umbral de etiqueta molecular reciclado predeterminado. Por ejemplo, el umbral de etiquetas moleculares recicladas puede ser aproximadamente 650 si los códigos de barras estocásticos comprenden aproximadamente 6561 etiquetas moleculares con secuencias distintas.

[0089] En algunas formas de realización, eliminar las etiquetas moleculares no únicos comprende: determinar un número teórico de etiquetas moleculares no únicos para el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación; y eliminar una etiqueta molecular con una aparición mayor que el enésimo marcador molecular más abundante de las etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación, en donde n es el número teórico de etiquetas moleculares no únicos.

[0090] En el presente documento se describen sistemas informáticos para determinar el número de objetivos. En algunas formas de realización, el sistema informático comprende: una memoria legible por computadora que almacena instrucciones ejecutables; y uno o más procesadores de computadora en comunicación con la memoria legible por computadora, en donde el uno o más procesadores de computadora son programados por las instrucciones ejecutables para realizar: (a) codificar con barras estocásticamente una pluralidad de objetivos usando una pluralidad de códigos de barras estocásticos para crear una pluralidad de objetivos con códigos de barras estocásticos, en donde cada uno de la pluralidad de códigos de barras estocásticos comprende una etiqueta molecular; (b) obtener datos de secuenciación de los objetivos con códigos de barras estocásticos; y (c) para una o más de la pluralidad de dianas: (i) contar el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación; (ii) determinar una serie de etiquetas moleculares de ruido con secuencias distintas asociadas con la diana en los datos de secuenciación; y (iii) estimar el número de la diana, en donde el número de la diana estimado se correlaciona con el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación contados en (i) ajustados de acuerdo con la cantidad de etiquetas moleculares de ruido determinado en (ii). En algunas formas de realización, el método comprende además determinar un estado de secuenciación del objetivo en los datos de secuenciación. El estado de secuenciación del objetivo en los datos de secuenciación es secuenciación saturada, subsecuenciación o sobresecuenciación.

[0091] En algunas formas de realización, el estado de secuenciación saturada se determina porque la diana tiene una cantidad de etiquetas moleculares con secuencias distintas mayores que un umbral de saturación predeterminado. Por ejemplo, el umbral de saturación predeterminado puede ser aproximadamente 6557 si los códigos de barras estocásticos comprenden aproximadamente 6561 etiquetas moleculares con secuencias distintas. El umbral de saturación predeterminado puede ser de aproximadamente 65532 si los códigos de barras estocásticos comprenden aproximadamente 65536 etiquetas moleculares con secuencias distintas. Si el estado de secuenciación del objetivo en los datos de secuenciación es el estado de secuenciación saturado, el número de etiquetas moleculares de ruido determinado en (ii) es cero.

[0092] En algunas formas de realización, el estado de subsecuenciación puede determinarse cuando el objetivo tiene una profundidad (por ejemplo, una profundidad promedio, mínima o máxima) menor que un umbral de subsecuenciación predeterminado. El umbral de secuenciación insuficiente es de aproximadamente cuatro. El umbral de secuenciación insuficiente puede ser independiente del número de etiquetas moleculares con secuencias distintas. Si el estado de secuenciación del objetivo en los datos de secuenciación es el estado de secuenciación insuficiente, el número de etiquetas moleculares de ruido determinado en (ii) es cero.

[0093] En algunas formas de realización, el estado de sobresecuenciación se determina porque la diana tiene una cantidad de etiquetas moleculares con secuencias distintas mayores que un umbral de sobresecuenciación predeterminado. Por ejemplo, el umbral de sobresecuenciación puede ser de aproximadamente 250 si los códigos de barras estocásticos comprenden aproximadamente 6561 etiquetas moleculares con secuencias distintas. El método comprende, si el estado de secuenciación de la diana en los datos de secuenciación es el estado de sobresecuenciación: submuestrear el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación hasta aproximadamente el umbral de sobresecuenciación predeterminado.

[0094] En algunas formas de realización, determinar el número de etiquetas moleculares de ruido con distintas secuencias asociadas con la diana en los datos de secuenciación comprende: si se satisface una condición de ajuste de distribución binomial negativa, (iv) ajustar una distribución binomial negativa de señal al número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación contados en (i), en donde la distribución binomial de señal negativa corresponde a una serie de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación contados en (i) que son etiquetas moleculares de señal; (v) ajustar una distribución binomial de ruido negativo al número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación contados en (i), en donde la distribución binomial de ruido negativo corresponde a una cantidad de etiquetas moleculares con secuencias distintas asociadas con el objetivo en los datos de secuenciación contados en (i) son etiquetas moleculares de ruido; y (vi) determinar el número de etiquetas moleculares de ruido utilizando la distribución binomial negativa de señal ajustada en (v) y la distribución binomial negativa de ruido ajustada en (vi).

[0095] En algunas formas de realización, la condición de ajuste de distribución binomial negativa comprende: el estado de secuenciación del objetivo en los datos de secuenciación no es el estado de secuenciación insuficiente o el estado de secuenciación excesiva. Determinación del número de etiquetas moleculares de ruido utilizando la distribución binomial de señal negativa ajustada en (v) y la distribución binomial de ruido negativo ajustada en (vi) comprende: para cada una de las distintas secuencias asociadas con el objetivo en los datos de secuenciación: determinar una probabilidad de señal de que la secuencia distinta esté en la distribución binomial de señal negativa; determinar una probabilidad de ruido de la secuencia distinta de estar en la distribución binomial de ruido negativo; y determinar que la secuencia distinta es una etiqueta molecular de ruido si la probabilidad de la señal es menor que la probabilidad del ruido.

[0096] En algunas formas de realización, determinar el número de etiquetas moleculares de ruido con distintas secuencias asociadas con el objetivo en los datos de secuenciación comprende: agregar pseudopuntos al número de etiquetas moleculares con secuencias distintas asociadas con el objetivo en los datos de secuenciación antes de determinar el número de etiquetas moleculares de ruido con secuencias distintas asociadas con el objetivo en los datos

de secuenciación en (ii), si el estado de secuenciación de la diana en los datos de secuenciación no es el estado de secuenciación insuficiente o el estado de secuenciación excesiva y el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación contados en (i) es menor que un umbral de pseudopuntos. El umbral de pseudopuntos es diez.

[0097] En algunas formas de realización, determinar el número de etiquetas moleculares de ruido con distintas secuencias asociadas con la diana en los datos de secuenciación comprende: eliminar etiquetas moleculares no únicas al determinar el número de etiquetas moleculares de ruido con secuencias distintas asociadas con la diana en los datos de secuenciación en (ii), si el estado de secuenciación de la diana en los datos de secuenciación no es el estado de secuenciación insuficiente o el estado de secuenciación excesiva y el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación contados en (i) no es inferior a un umbral de pseudopuntos.

[0098] En algunas formas de realización, eliminar las etiquetas moleculares no únicos comprende eliminar las etiquetas moleculares no únicos al determinar el número de etiquetas moleculares de ruido con secuencias distintas asociadas con la diana en los datos de secuenciación en (ii) si el número de etiquetas moleculares etiquetas con secuencias distintas asociadas con la diana en los datos de secuenciación es mayor que un umbral de etiqueta molecular reciclado predeterminado. Por ejemplo, el umbral de etiquetas moleculares recicladas puede ser aproximadamente 650 si los códigos de barras estocásticos comprenden aproximadamente 6561 etiquetas moleculares con secuencias distintas.

[0099] En el presente documento se divulgan uno o más medios legibles por computadora no transitorios que comprenden códigos ejecutables que, cuando se ejecutan, realizan cualquiera de los métodos divulgados en el presente documento.

BREVE DESCRIPCIÓN DE LOS DIBUJOS

[0100]

FIG. 1 ilustra un código de barras estocástico ejemplar no limitativo.
FIG. 2 muestra un flujo de trabajo ejemplar no limitante de códigos de barras estocásticos y conteo digital.
FIG. 3 es una ilustración esquemática que muestra un proceso ejemplar no limitante para generar una biblioteca indexada de objetivos con códigos de barras estocásticos a partir de una pluralidad de objetivos.
FIG. 4 es una ilustración esquemática que muestra distribuciones ejemplares no limitantes de errores de etiquetas moleculares, errores de etiquetas de muestras y señales de etiquetas moleculares verdaderas.
FIG. 5 es un diagrama de flujo que muestra una forma de realización ejemplar no limitante de corrección de errores de secuenciación y PCR usando etiquetas moleculares.
FIG. 6 es una ilustración esquemática que muestra datos de secuencia obtenidos mediante secuenciación completa y secuenciación incompleta.
FIG. 7 es un diagrama de flujo que muestra una forma de realización ejemplar no limitante de la corrección de errores de secuenciación y PCR usando etiquetas moleculares basados en la adyacencia direccional.
FIG. 8 es un diagrama de flujo que muestra una forma de realización ejemplar no limitante de la corrección de errores de secuenciación y PCR basándose en la corrección de errores de sustitución recursiva y segundas derivadas del cambio de profundidad del marcador molecular.
FIG. 9 es un diagrama de flujo que muestra una forma de realización ejemplar no limitante de la corrección de errores de secuenciación y PCR basándose en la corrección de errores por sustitución recursiva y la corrección de errores basada en la distribución.
FIG. 10 es un diagrama de flujo que muestra una forma de realización ejemplar no limitante de corrección de errores usando dos distribuciones binomiales negativas.
FIG. 11 es un diagrama de flujo que muestra una forma de realización ejemplar no limitante de la corrección de errores de secuenciación y PCR basándose en la corrección de errores por sustitución recursiva y la corrección de errores basada en la distribución mediante submuestreo de placas de micropocillos y mapeo de etiquetas moleculares.
FIG. 12 es un diagrama de flujo que muestra una forma de realización ejemplar no limitante de corrección de errores de secuenciación y PCR basándose en la corrección de errores de sustitución recursiva y la corrección de errores basada en la distribución mediante submuestreo de genes y mapeo de etiquetas moleculares.
FIG. 13 es un diagrama de flujo que muestra una forma de realización ejemplar no limitante de la corrección de errores de secuenciación y PCR basándose en la corrección de errores por sustitución recursiva y la corrección de errores basada en distribución por recursividad.
FIG. 14 es un diagrama de flujo que muestra una forma de realización ejemplar no limitante de la corrección de errores de secuenciación y PCR basándose en la corrección de errores de sustitución recursiva y la corrección de errores basada en la distribución utilizando la segunda etiqueta molecular más alta para las estimaciones de parámetros iniciales.
FIG. 15 muestra un instrumento ejemplar no limitante adecuado para usar en los métodos de la divulgación.
FIG. 16 ilustra una arquitectura ejemplar no limitante de un sistema informático que se puede utilizar en relación con formas de realización de la presente divulgación.
FIG. 17 ilustra una arquitectura ejemplar no limitante que muestra una red con una pluralidad de sistemas informáticos adecuados para su uso en los métodos de la divulgación.

FIG. 18 ilustra una arquitectura ejemplar no limitante de un sistema informático multiprocesador que utiliza un espacio de memoria de direcciones virtuales compartido de acuerdo con los métodos de la divulgación.
FIGS. 19A-19B muestran ejemplos no limitantes de genes secuenciados completa e incompletamente.
FIG. 20 es un gráfico ejemplar no limitante de lecturas de secuenciación frente a sus rangos después de la corrección de errores de secuenciación de una base y umbral para separar códigos de barras verdaderos y de error.
FIG. 21 es una ilustración ejemplar no limitativa del modelo de Poisson truncado en cero.
FIG. 22 muestra gráficos de barras de lecturas de secuenciación totales por pocillo.
FIG. 23 muestra gráficos de barras de % de genes completamente secuenciados, % de etiquetas moleculares (EM) retenidas como códigos de barras verdaderos y % de lecturas retenidas asignadas a aquellas EM retenidas para cada pocillo.
FIG. 24 muestra diagramas de caja del % de lecturas retenidas variadas según los genes para cada pocillo.
FIGS. 25A-25B muestran gráficos de análisis de componentes principales (PCA) del uso de EM sin procesar frente a IM corregido después de aplicar el algoritmo de dos placas.
FIG. 26 es un gráfico ejemplar de cálculo teórico de etiquetas moleculares únicas utilizadas a medida que aumentan las moléculas de entrada.
FIG. 27 es un gráfico ejemplar que muestra la cobertura de etiquetas moleculares de cada etiqueta molecular a través de una placa de micropocillos para un gen de alta expresión - ATCB, donde se observaron distintas distribuciones entre etiquetas moleculares de error y etiquetas moleculares reales.
FIG. 28 es un gráfico ejemplar que muestra el ajuste de dos distribuciones binomiales negativas a la cobertura de etiquetas moleculares de cada etiqueta molecular a través de una placa de micropocillos para un gen de alta expresión: ATCB. El ajuste de dos distribuciones binomiales negativas demuestra que se pueden distinguir estadísticamente los errores de etiqueta molecular con una profundidad de etiqueta molecular más baja y una etiqueta molecular verdadera con una profundidad de etiqueta molecular más alta. El eje x es la profundidad molecular.
FIG. 29 muestra la corrección de la etiqueta molecular, donde la distancia de Hamming por pares de 1 estaba sobrerrepresentada. Después de la corrección de la etiqueta molecular, las etiquetas moleculares con una distancia de Hamming de uno entre sí se agruparon y colapsaron en la misma etiqueta molecular principal.
FIG. 30 muestra la curva del número corregido de etiquetas moleculares frente al número corregido de coberturas de lecturas.
FIG. 31 muestra una ilustración esquemática de un ejemplo de corrección de errores de sustitución recursiva.
FIG. 32, los paneles (a)-(e) muestran resultados ejemplares de corrección de errores de secuenciación y PCR basados en segundas derivadas del cambio de profundidad del marcador molecular.
FIG. 33, los paneles (a)-(c) muestran resultados ejemplares de corrección de PCR y errores de secuenciación basados en dos distribuciones binomiales negativas para CD69.
FIG. 34, los paneles (a)-(c) muestran resultados ejemplares de corrección de PCR y errores de secuenciación basados en dos distribuciones binomiales negativas para CD3E.
FIG. 35, los paneles (a)-(c) muestran resultados ejemplares de corrección de PCR y errores de secuenciación basados en dos distribuciones binomiales negativas para genes de alta expresión.
FIG. 36 muestra resultados ejemplares del reciclaje de etiquetas moleculares ricas en G para genes de alta expresión.
FIG. 37, los paneles (a)-(b) muestran resultados ejemplares del ajuste de datos de entrada para genes de alta expresión antes de ajustar dos distribuciones binomiales negativas.
FIG. 38, los paneles (a)-(j) muestran una validación ejemplar no limitante del conjunto de datos corregido utilizando dos distribuciones binomiales negativas.
FIG. 39, los paneles (a)-(d) muestran visualizaciones ejemplares de incrustación de vecinas t-estocásticas (t-SNE) del ensayo dirigido Precise™ de 96 pocillos de células individuales mixtas de Jurkat y cáncer de mama (BrCa) (86 genes examinados).
FIG. 40, los paneles (a)-(b) son gráficos ejemplares no limitantes que muestran análisis de expresión diferencial entre grupos de células para genes con >0 EM en ambos grupos seleccionados calculados por DBScan y determinados por el nivel de marcador genético en cada grupo.
FIG. 41, los paneles (a)-(d) son gráficos ejemplares no limitantes que muestran la visualización de incrustación vecina t-estocástica (t-SNE) de un ensayo dirigido BD Precise™ de una placa de 96 pocillos de células individuales Jurkat mixtas y cáncer de mama (T47D) con 86 genes examinados.
FIG. 42, los paneles (a)-(b) son mapas de calor ejemplares no limitantes que muestran la expresión génica diferencial mediante recuentos de etiquetas moleculares entre diferentes grupos de células identificados en la FIG. 41 antes de cualquier paso de corrección de errores (EM sin procesar se muestra en la FIG. 42, panel (a)) y después de la corrección RSEC y DBEC (EM ajustado se muestra en la FIG. 42, panel (b)).

DESCRIPCIÓN DETALLADA

[0101] En la siguiente descripción detallada, se hace referencia a los dibujos adjuntos, que forman parte de la misma. En los dibujos, símbolos similares normalmente identifican componentes similares, a menos que el contexto indique lo contrario.

[0102] La cuantificación de pequeñas cantidades de ácidos nucleicos, por ejemplo moléculas de ácido ribonucleótido mensajero (ARNm), es clínicamente importante para determinar, por ejemplo, los genes que se expresan en una célula en diferentes etapas de desarrollo o en diferentes condiciones ambientales. Sin embargo, también puede resultar muy complicado determinar el número absoluto de moléculas de ácido nucleico (p. ej., moléculas de ARNm), especialmente cuando el número de moléculas es muy pequeño. Un método para determinar el número absoluto de moléculas en una muestra es la reacción en hebra de la polimerasa digital (PCR). Idealmente, la PCR produce una copia idéntica de una molécula en cada ciclo. Sin embargo, la PCR puede tener desventajas tales como que cada molécula se replica con una probabilidad estocástica, y esta probabilidad varía según el ciclo de la PCR y la secuencia genética, lo que genera un sesgo de amplificación y mediciones inexactas de la expresión genética. Se pueden utilizar códigos de barras estocásticos con etiquetas moleculares únicas (también denominados índices moleculares (IM)) para contar el número de moléculas y corregir el sesgo de amplificación. Los códigos de barras estocásticos, como el ensayo Precise™ (Cellular Research, Inc. (Palo Alto, CA)) pueden corregir el sesgo inducido por la PCR y los pasos de preparación de la biblioteca mediante el uso de etiquetas moleculares (EM) para etiquetar los ARNm durante la transcripción inversa (TI).

[0103] El ensayo Precise™ puede utilizar un conjunto que no se agota de códigos de barras estocásticos con un gran número, por ejemplo de 6561 a 65536, etiquetas moleculares únicas en oligonucleótidos poli(T) para hibridar con todos los ARNm poli(A) en una muestra durante el paso TI. Además de las etiquetas moleculares, se pueden utilizar etiquetas de muestra (también denominadas índice de muestra (IM)) de códigos de barras estocásticos para identificar cada pocillo de la placa Precise™. Un código de barras estocástico puede comprender un sitio de cebado de PCR universal. Durante la TI, las moléculas del gen objetivo reaccionan aleatoriamente con códigos de barras estocásticos. Cada molécula diana puede hibridarse con un código de barras estocástico, lo que genera moléculas de ácido ribonucleótido complementario (ADNc) con código de barras estocástico. Después del etiquetado, las moléculas de ADNc con código de barras estocásticas de los micropocillos de una placa de micropocillos se pueden agrupar en un solo tubo para la amplificación y secuenciación por PCR. Los datos de secuenciación sin procesar se pueden analizar para producir el número de lecturas, el número de códigos de barras estocásticos con etiquetas moleculares únicas y el número de moléculas de ARNm basándose en una corrección de Poisson o un método de corrección basado en dos distribuciones binomiales negativas.

[0104] Además de la corrección del sesgo, las etiquetas moleculares pueden proporcionar una mejor comprensión de la calidad estadística de los resultados al revelar el número inicial de moléculas de ADNc presentes en las lecturas de secuenciación observadas. Por ejemplo, una gran cantidad de lecturas puede indicar una respuesta estadísticamente precisa, pero si las lecturas se derivan de solo una pequeña cantidad de moléculas de ARNm iniciales, entonces la precisión de la medición puede verse comprometida.

[0105] Aunque el sesgo de amplificación inducido por la PCR y los pasos de preparación de la biblioteca se puede remediar, por ejemplo, mediante etiquetas moleculares, la cuantificación del número absoluto de moléculas aún puede ser un desafío debido a varios otros factores. En primer lugar, la estimación del número de moléculas de ARNm puede estar limitada por la diversidad total de las etiquetas moleculares. Durante la codificación de barras estocástica, las moléculas de ARNm pueden reaccionar aleatoriamente con los códigos de barras estocásticos disponibles. Por tanto, cada molécula de ARNm puede hibridarse con un código de barras estocástico; sin embargo, su etiqueta molecular podría no ser necesariamente única para un gen determinado. Cuando la cantidad de moléculas de ARNm es pequeña en relación con la cantidad de códigos de barras estocásticos, es probable que cada molécula de ARNm se hibride con un código de barras estocástico con una etiqueta molecular única, y contar la cantidad de moléculas puede ser equivalente a contar la cantidad de etiquetas moleculares.

[0106] A medida que aumenta el número de moléculas de ARNm, es cada vez más probable que múltiples moléculas de ARNm se hibriden con códigos de barras estocásticos con las mismas etiquetas moleculares. Por lo tanto, el uso de recuentos de etiquetas moleculares únicas puede subestimar el número de moléculas. En algunos casos, el número de moléculas de ARNm se puede estimar basándose en una corrección de Poisson o una corrección basada en dos distribuciones binomiales negativas del número de etiquetas moleculares únicas observadas en total. Sin embargo, en el extremo donde se observa toda la colección de 6561 códigos de barras estocásticos, es posible que ya no sea posible una corrección de Poisson o una corrección basada en dos distribuciones binomiales negativas. Por ejemplo, independientemente de 65.000 o 100.000 moléculas de ARNm iniciales, en cualquier caso se espera un máximo de 6.561 códigos de barras estocásticos saturados.

[0107] En segundo lugar, los errores de PCR (es decir, errores ocurridos durante la amplificación de PCR) pueden introducir códigos de barras estocásticos artificiales e inflar arbitrariamente los recuentos de etiquetas moleculares. En tercer lugar, el sesgo de amplificación por PCR y la PCR ineficiente pueden generar copias bajas de moléculas con códigos de barras que no se pueden distinguir de los errores. En cuarto lugar, los errores de secuenciación, la llamada inexacta de secuencias de códigos de barras estocásticos, pueden introducir códigos de barras estocásticos artificiales e inflar los recuentos de etiquetas moleculares. Además, la profundidad de la secuenciación puede ser importante, especialmente cuando la secuenciación es demasiado superficial para detectar todos los ARNm con código de barras estocásticos presentes en una biblioteca de muestras.

[0108] En el presente documento se divulgan métodos y sistemas para determinar el número de dianas con uno o más de PCR o errores de secuenciación corregidos o ajustados. En algunas formas de realización, el método comprende: (a)

codificar de forma estocástica una pluralidad de objetivos usando una pluralidad de códigos de barras estocásticos para crear una pluralidad de objetivos con códigos de barras estocásticos, en donde cada uno de la pluralidad de códigos de barras estocásticos comprende una etiqueta molecular; (b) obtener datos de secuenciación de los objetivos con códigos de barras estocásticos; y (c) para una o más de la pluralidad de dianas: (i) contar el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación; (ii) determinar un estado de calidad de la diana en los datos de secuenciación obtenidos en (b); (iii) determinar uno o más errores de datos de secuenciación en los datos de secuenciación obtenidos en (b), en donde determinar uno o más errores de datos de secuenciación en los datos de secuenciación comprende determinar uno o más de: el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación, el estado de calidad de la diana en los datos de secuenciación y el número de etiquetas moleculares con secuencias distintas en la pluralidad de códigos de barras estocásticos; y (iv) estimar el número de la diana, en donde el número de la diana estimado se correlaciona con el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación contados en (i) ajustados de acuerdo con uno o más errores de datos de secuenciación determinados en (iii).

[0109] Se divulgan métodos para determinar el número de dianas con uno o más errores de PCR o de secuenciación corregidos o ajustados en función de la adyacencia direccional. En algunas formas de realización, el método comprende: (a) codificar de forma estocástica una pluralidad de objetivos usando una pluralidad de códigos de barras estocásticos para crear una pluralidad de objetivos con códigos de barras estocásticos, en donde cada uno de la pluralidad de códigos de barras estocásticos comprende una etiqueta molecular; (b) obtener datos de secuenciación de los objetivos con códigos de barras estocásticos; (c) para una o más de la pluralidad de dianas: (i) contar el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación; (ii) identificar grupos de etiquetas moleculares de la diana usando adyacencia direccional; (iii) colapsar los datos de secuenciación obtenidos en (b) usando los grupos de etiquetas moleculares de la diana identificada en (ii); (iv) estimar el número de la diana, en donde el número de la diana estimado se correlaciona con el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación contados en (i) después de colapsar los datos de secuenciación en (ii).

[0110] Los sistemas informáticos para determinar el número de objetivos con uno o más errores de PCR o de secuenciación corregidos o ajustados. Se divulga un medio legible por computadora no transitorio que contiene códigos ejecutables, cuando se ejecuta, hace que uno o más dispositivos informáticos determinen el número de objetivos con uno o más errores de PCR o de secuenciación corregidos o ajustados.

Definiciones

[0111] A menos que se defina lo contrario, los términos técnicos y científicos utilizados en el presente documento tienen el mismo significado que entiende comúnmente un experto en la técnica a la que pertenece la presente divulgación. Véase, por ejemplo, Singleton et al., Dictionary of Microbiology and Molecular Biology, 2ª ed., J. Wiley & Sons (Nueva York, NY 1994); Sambrook et al., Molecular Cloning, A Laboratory Manual, Cold Springs Harbor Press (Cold Springs Harbor, NY 1989). Para los fines de la presente divulgación, los siguientes términos se definen a continuación.

[0112] Como se usa en el presente documento, el término "adaptador" puede significar una secuencia para facilitar la amplificación o secuenciación de ácidos nucleicos asociados. Los ácidos nucleicos asociados pueden comprender ácidos nucleicos diana. Los ácidos nucleicos asociados pueden comprender uno o más de etiquetas espaciales, etiquetas diana, etiquetas de muestra, etiquetas de indexación, códigos de barras, códigos de barras estocásticos o etiquetas moleculares. Los adaptadores pueden ser lineales. Los adaptadores pueden ser adaptadores preadenilados. Los adaptadores pueden ser monocatenarios o bicatenarios. Se pueden ubicar uno o más adaptadores en el extremo 5' o 3' de un ácido nucleico. Cuando los adaptadores comprenden secuencias conocidas en los extremos 5' y 3', las secuencias conocidas pueden ser secuencias iguales o diferentes. Un adaptador situado en los extremos 5' y/o 3' de un polinucleótido puede ser capaz de hibridar con uno o más oligonucleótidos inmovilizados en una superficie. Un adaptador puede, en algunas formas de realización, comprender una secuencia universal. Una secuencia universal puede ser una región de secuencia de nucleótidos que es común a dos o más moléculas de ácido nucleico. Las dos o más moléculas de ácido nucleico también pueden tener regiones de secuencia diferente. Así, por ejemplo, los adaptadores 5' pueden comprender secuencias de ácido nucleico idénticas y/o universales y los adaptadores 3' pueden comprender secuencias idénticas y/o universales. Una secuencia universal que puede estar presente en diferentes miembros de una pluralidad de moléculas de ácido nucleico puede permitir la replicación o amplificación de múltiples secuencias diferentes usando un único cebador universal que es complementario a la secuencia universal. De manera similar, al menos una, dos (por ejemplo, un par) o más secuencias universales que pueden estar presentes en diferentes miembros de una colección de moléculas de ácido nucleico pueden permitir la replicación o amplificación de múltiples secuencias diferentes usando al menos uno, dos (por ejemplo, un par) o más cebadores universales únicos que son complementarios a las secuencias universales. Por tanto, un cebador universal incluye una secuencia que puede hibridarse con dicha secuencia universal. Las moléculas que portan la secuencia de ácido nucleico diana pueden modificarse para unir adaptadores universales (por ejemplo, secuencias de ácido nucleico no diana) a uno o ambos extremos de las diferentes secuencias de ácido nucleico diana. Uno o más cebadores universales unidos al ácido nucleico diana pueden proporcionar sitios para la hibridación de cebadores universales. Uno o más cebadores universales unidos al ácido nucleico diana pueden ser iguales o diferentes entre sí.

[0113] Como se usa en el presente documento, el término "asociado" o "asociado con" puede significar que dos o más especies son identificables por estar coubicadas en un momento determinado. Una asociación puede significar que dos o más especies están o estuvieron dentro de un contenedor similar. Una asociación puede ser una asociación informática, en la que, por ejemplo, se almacena información digital sobre dos o más especies y se puede utilizar para determinar que una o más de las especies estuvieron coubicadas en un momento determinado. Una asociación también puede ser una asociación física. En algunas formas de realización, dos o más especies asociadas están "atadas", "unidas" o "inmovilizadas" entre sí o a una superficie sólida o semisólida común. Una asociación puede referirse a medios covalentes o no covalentes para unir etiquetas a soportes sólidos o semisólidos tales como perlas. Una asociación puede ser un enlace covalente entre un objetivo y una etiqueta.

[0114] Como se utiliza en el presente documento, el término "complementario" puede referirse a la capacidad de emparejamiento preciso entre dos nucleótidos. Por ejemplo, si un nucleótido en una posición determinada de un ácido nucleico es capaz de formar enlaces de hidrógeno con un nucleótido de otro ácido nucleico, entonces se considera que los dos ácidos nucleicos son complementarios entre sí en esa posición. La complementariedad entre dos moléculas de ácido nucleico monocatenario puede ser "parcial", en la que sólo se unen algunos de los nucleótidos, o puede ser completa cuando existe una complementariedad total entre las moléculas monocatenarias. Se puede decir que una primera secuencia de nucleótidos es el "complemento" de una segunda secuencia si la primera secuencia de nucleótidos es complementaria a la segunda secuencia de nucleótidos. Se puede decir que una primera secuencia de nucleótidos es el "complemento inverso" de una segunda secuencia, si la primera secuencia de nucleótidos es complementaria de una secuencia que es inversa (es decir, el orden de los nucleótidos se invierte) de la segunda secuencia. Tal como se utilizan en el presente documento, los términos "complemento", "complementario" y "complemento inverso" se pueden utilizar indistintamente. Se entiende a partir de la divulgación que si una molécula puede hibridarse con otra molécula puede ser el complemento de la molécula que se está hibridando.

[0115] Como se utiliza en el presente documento, el término "recuento digital" puede referirse a un método para estimar un número de moléculas diana en una muestra. El recuento digital puede incluir el paso de determinar una cantidad de etiquetas únicas que se han asociado con objetivos en una muestra. Esta metodología estocástica transforma el problema de contar moléculas de uno de localizar e identificar moléculas idénticas a una serie de preguntas digitales de sí/no sobre la detección de un conjunto de etiquetas predefinidas.

[0116] Como se usa en el presente documento, el término "etiqueta" o "etiquetas" puede referirse a códigos de ácido nucleico asociados con una diana dentro de una muestra. Un marcador puede ser, por ejemplo, un marcador de ácido nucleico. Una etiqueta puede ser una etiqueta total o parcialmente amplificable. Una etiqueta puede ser total o parcialmente secuenciable. Un marcador puede ser una porción de un ácido nucleico nativo que sea identificable como distinto. Una etiqueta puede ser una secuencia conocida. Un marcador puede comprender una unión de secuencias de ácidos nucleicos, por ejemplo una unión de una secuencia nativa y no nativa. Tal como se utiliza en el presente documento, el término "etiqueta" se puede utilizar indistintamente con los términos "índice", "marcador" o "etiqueta-marcador". Las etiquetas pueden transmitir información. Por ejemplo, en diversas formas de realización, se pueden usar etiquetas para determinar la identidad de una muestra, una fuente de una muestra, una identidad de una célula y/o un objetivo.

[0117] Como se utiliza en el presente documento, el término "depósitos que no se agotan" puede referirse a un conjunto de códigos de barras estocásticos formados por muchas etiquetas diferentes. Un depósito que no se agota puede comprender una gran cantidad de códigos de barras estocásticos diferentes, de modo que cuando el depósito que no se agota está asociado con un conjunto de objetivos, es probable que cada objetivo esté asociado con un código de barras estocástico único. La unicidad de cada molécula objetivo marcada puede determinarse mediante estadísticas de elección aleatoria y depende del número de copias de moléculas objetivo idénticas en la colección en comparación con la diversidad de etiquetas. El tamaño del conjunto resultante de moléculas objetivo marcadas puede determinarse mediante la naturaleza estocástica del proceso de codificación de barras, y el análisis del número de códigos de barras estocásticos detectados permite calcular el número de moléculas objetivo presentes en la colección o muestra original. Cuando la relación entre el número de copias de una molécula diana presente y el número de códigos de barras estocásticos únicos es baja, las moléculas diana marcadas son altamente únicas (es decir, existe una probabilidad muy baja de que más de una molécula diana haya sido etiquetada con una etiqueta dada).

[0118] Como se usa en el presente documento, el término "ácido nucleico" se refiere a una secuencia de polinucleótidos, o fragmento de la misma. Un ácido nucleico puede comprender nucleótidos. Un ácido nucleico puede ser exógeno o endógeno a una célula. Un ácido nucleico puede existir en un entorno libre de células. Un ácido nucleico puede ser un gen o un fragmento del mismo. Un ácido nucleico puede ser ADN. Un ácido nucleico puede ser ARN. Un ácido nucleico puede comprender uno o más análogos (por ejemplo, hebra principal alterada, azúcar o nucleobase). Algunos ejemplos no limitantes de análogos incluyen: 5-bromouracilo, ácido peptídico nucleico, ácido xenonucleico, morfolinos, ácidos nucleicos bloqueados, ácidos nucleicos de glicol, ácidos nucleicos de treosa, didesoxinucleótidos, cordicepina, 7-deaza-GTP, fluoróforos (por ejemplo, rodamina o fluoresceína unida al azúcar), nucleótidos que contienen tiol, nucleótidos unidos a biotina, análogos de bases fluorescentes, islas CpG, metil-7-guanosina, nucleótidos metilados, inosina, tiouridina, pseudouridina, dihidrouridina, queuosina y wyosina. "Ácido nucleico", "polinucleótido", "polinucleótido diana" y "ácido nucleico diana" se pueden utilizar indistintamente.

[0119] Un ácido nucleico puede comprender una o más modificaciones (por ejemplo, una modificación de base, una modificación del esqueleto), para proporcionar al ácido nucleico una característica nueva o mejorada (por ejemplo, estabilidad mejorada). Un ácido nucleico puede comprender una etiqueta de afinidad de ácido nucleico. Un nucleósido puede ser una combinación de base y azúcar. La porción de base del nucleósido puede ser una base heterocíclica. Las dos clases más comunes de tales bases heterocíclicas son las purinas y las pirimidinas. Los nucleótidos pueden ser nucleósidos que incluyen además un grupo fosfato unido covalentemente a la porción de azúcar del nucleósido. Para aquellos nucleósidos que incluyen un azúcar pentofuranosilo, el grupo fosfato puede estar unido al resto hidroxilo 2', 3' o 5' del azúcar. Al formar ácidos nucleicos, los grupos fosfato pueden unir covalentemente nucleósidos adyacentes entre sí para formar un compuesto polimérico lineal. A su vez, los extremos respectivos de este compuesto polimérico lineal se pueden unir además para formar un compuesto circular; sin embargo, generalmente son adecuados compuestos lineales. Además, los compuestos lineales pueden tener complementariedad de bases de nucleótidos internas y, por lo tanto, pueden plegarse de manera que se produzca un compuesto total o parcialmente bicatenario. Dentro de los ácidos nucleicos, comúnmente se puede decir que los grupos fosfato forman la columna vertebral internucleosídica del ácido nucleico. El enlace o hebra principal puede ser un enlace fosfodiéster de 3' a 5'.

[0120] Un ácido nucleico puede comprender una estructura principal modificada y/o enlaces internucleosídicos modificados. Las cadenas principales modificadas pueden incluir aquellas que retienen un átomo de fósforo en la hebra principal y aquellas que no tienen un átomo de fósforo en la hebra principal. Las cadenas principales de ácido nucleico modificadas adecuadas que contienen un átomo de fósforo en ellas pueden incluir, por ejemplo, fosforotioatos, fosforotioatos quirales, fosforoditioatos, fosfotriésteres, aminoalquilfosfotriésteres, metil y otros alquilfosfonatos tales como 3'-alquilenfosfonatos, 5'-alquilenfosfonatos, fosfonatos quirales, fosfinatos, fosforamidatos que incluyen 3'-amino fosforamidato y aminoalquil fosforamidatos, fosforodiamidatos, tionofosforamidatos, tionoalquilfosfonatos, tionoalquilfosfotriésteres, selenofosfatos y boranofosfatos que tienen enlaces 3'-5' normales, análogos unidos 2'-5' y aquellos que tienen polaridad invertida en los que uno o más enlaces internucleotídicos es un enlace 3' a 3', 5' a 5' o 2' a 2'.

[0121] Un ácido nucleico puede comprender cadenas principales de polinucleótidos que están formadas por enlaces internucleósidos de alquilo o cicloalquilo de hebra corta, enlaces internucleósidos de heteroátomo y alquilo o cicloalquilo mixtos, o uno o más enlaces internucleósidos heteroatómicos o heterocíclicos de hebra corta. Estos pueden incluir aquellos que tienen enlaces morfolino (formados en parte a partir de la porción de azúcar de un nucleósido); cadenas principales de siloxano; cadenas principales de sulfuro, sulfóxido y sulfona; cadenas principales de formacetilo y tioformacetilo; cadenas principales de metilenformacetilo y tioformacetilo; cadenas principales de riboacetilo; cadenas principales que contienen alquenos; cadenas principales de sulfamato; cadenas principales de metilenimino y metilenihidrazino; cadenas principales de sulfonato y sulfonamida; cadenas principales de amida; y otros que tienen partes componentes mezcladas de N, O, S y CH2.

[0122] Un ácido nucleico puede comprender un mimético de ácido nucleico. Se puede pretender que el término "mimético" incluya polinucleótidos en los que solo el anillo de furanosa o tanto el anillo de furanosa como el enlace internucleotídico se reemplazan con grupos no furanosa; el reemplazo de solo el anillo de furanosa también puede denominarse un sustituto de azúcar. El resto de base heterocíclica o un resto de base heterocíclica modificada se puede mantener para la hibridación con un ácido nucleico diana apropiado. Uno de dichos ácidos nucleicos puede ser un ácido peptídico nucleico (PNA). En un PNA, la hebra principal de azúcar de un polinucleótido se puede reemplazar con una hebra principal que contiene amida, en particular una hebra principal de aminoetilglicina. Los nucleótidos pueden retenerse y unirse directa o indirectamente a átomos de nitrógeno aza de la porción amida del esqueleto. La hebra principal de los compuestos de PNA puede comprender dos o más unidades de aminoetilglicina unidas, lo que le da al PNA una hebra principal que contiene amida. Los restos de bases heterocíclicas pueden unirse directa o indirectamente a átomos de nitrógeno aza de la porción amida de la hebra principal.

[0123] Un ácido nucleico puede comprender una estructura principal de morfolino. Por ejemplo, un ácido nucleico puede comprender un anillo de morfolino de 6 miembros en lugar de un anillo de ribosa. En algunas de estas formas de realización, un fosforodiamidato u otro enlace internucleósido no fosfodiéster puede reemplazar un enlace fosfodiéster.

[0124] Un ácido nucleico puede comprender unidades de morfolino unidas (es decir, ácido morfolino nucleico) que tienen bases heterocíclicas unidas al anillo de morfolino. Los grupos de enlace pueden unir las unidades monoméricas de morfolino en un ácido morfolinonucleico. Los compuestos oligoméricos no iónicos basados en morfolino pueden tener menos interacciones no deseadas con proteínas celulares. Los polinucleótidos basados en morfolino pueden ser imitadores no iónicos de ácidos nucleicos. Se puede unir una variedad de compuestos dentro de la clase morfolino usando diferentes grupos de enlace. Se puede hacer referencia a otra clase de miméticos de polinucleótidos como ácidos nucleicos de ciclohexenilo (CeNA). El anillo de furanosa normalmente presente en una molécula de ácido nucleico puede sustituirse por un anillo de ciclohexenilo. Los monómeros de fosforamidita protegidos con CeNA DMT se pueden preparar y utilizar para la síntesis de compuestos oligoméricos utilizando la química de fosforamidita. La incorporación de monómeros de CeNA en una hebra de ácido nucleico puede aumentar la estabilidad de un híbrido de ADN/ARN. Los oligoadenilatos de CeNA pueden formar complejos con complementos de ácidos nucleicos con una estabilidad similar a la de los complejos nativos. Una modificación adicional puede incluir ácidos nucleicos bloqueados (LNA) en los que el grupo 2'-hidroxilo está unido al átomo de carbono 4' del anillo de azúcar formando así un enlace 2'-C, 4'-C-oximetileno, formando así un resto de azúcar bicíclico. El enlace puede ser un grupo metileno (-CH2), que une el átomo de oxígeno 2'

y el átomo de carbono 4' en el que n es 1 o 2. LNA y análogos de LNA pueden mostrar estabilidades térmicas dúplex muy altas con ácido nucleico complementario (Tm=+3 a +10 ºC), estabilidad frente a la degradación exonucleolítica 3' y buenas propiedades de solubilidad.

[0125] Un ácido nucleico también puede incluir modificaciones o sustituciones de nucleobases (a menudo denominadas simplemente "base"). Como se usa en el presente documento, las nucleobases "no modificadas" o "naturales" pueden incluir las bases purínicas (por ejemplo, adenina (A) y guanina (G)) y las bases pirimidínicas (por ejemplo, timina (T), citosina (C) y uracilo (U)). Las nucleobases modificadas pueden incluir otras nucleobases sintéticas y naturales tales como 5-metilcitosina (5-me-C), 5-hidroximetilcitosina, xantina, hipoxantina, 2-aminoadenina, 6-metilo y otros derivados alquílicos de adenina y guanina, 2-propil y otros derivados alquílicos de adenina y guanina, 2-tiouracilo, 2-tiotimina y 2-tiocitosina, 5-halouracilo y citosina, 5-propinil (-C=C-CH3)uracilo y citosina y otros derivados alquinílicos de bases pirimidínicas, 6-azouracilo, citosina y timina, 5-uracilo (pseudouracilo), 4-tiouracilo, 8-halo, 8-amino, 8-tiol, 8-tioalquilo, 8-hidroxilo y otras adeninas y guaninas 8-sustituidas, 5-halo particularmente 5-bromo, 5-trifluorometilo y otros uracilos y citosinas 5-sustituidos, 7-metilguanina y 7-metiladenina, 2-F-adenina, 2-aminoadenina, 8-azaguanina y 8-azaadenina, 7-desazaguanina y 7-desazaadenina y 3-desazaguanina y 3-desazaadenina. Las nucleobases modificadas pueden incluir pirimidinas tricíclicas como la fenoxazina citidina(1H-pirimido(5,4-b)(1,4)benzoxazin-2(3H)-ona), fenotiazina citidina (1H-pirimido(5,4-b)(1,4)benzotiazin-2(3H))-ona), abrazaderas G tales como una fenoxazina citidina sustituida (por ejemplo, 9-(2-aminoetoxi)-H-pirimido(5,4-(b)(1,4)benzoxazin-2(3H)-ona), fenotiazina citidina (1H-pirimido(5,4-b)(1,4)benzotiazin-2(3H)-ona), abrazaderas G tales como una fenoxazina citidina sustituida (por ejemplo, 9-(2-aminoetoxi)-H-pirimido (5,4-(b)(1,4)benzoxazin-2(3H)-ona), carbazol citidina (2H-pirimido(4,5-b)indol-2-ona), piridoindol citidina (H-pirido(3',':4,5)pirrolo[2,3-d]pirimidin-2-ona).

[0126] Como se usa en el presente documento, el término "muestra" puede referirse a una composición que comprende dianas. Las muestras adecuadas para el análisis mediante los métodos, dispositivos y sistemas descritos incluyen células, tejidos, órganos u organismos.

[0127] Como se usa en el presente documento, el término "dispositivo de muestreo" o "dispositivo" puede referirse a un dispositivo que puede tomar una sección de una muestra y/o colocar la sección sobre un sustrato. Un dispositivo de muestra puede referirse, por ejemplo, a una máquina clasificadora de células activadas por fluorescencia (FACS), una máquina clasificadora de células, una aguja de biopsia, un dispositivo de biopsia, un dispositivo de sección de tejido, un dispositivo de microfluidos, una rejilla de cuchillas y/o un micrótomo.

[0128] Como se usa en el presente documento, el término "soporte sólido" puede referirse a superficies sólidas o semisólidas discretas a las que se pueden unir una pluralidad de códigos de barras estocásticos. Un soporte sólido puede abarcar cualquier tipo de esfera, bola, cojinete, cilindro u otra configuración similar sólida, porosa o hueca, compuesta de material plástico, cerámico, metálico o polimérico (por ejemplo, hidrogel) sobre el cual se puede inmovilizar un ácido nucleico (por ejemplo, de forma covalente o no covalente). Un soporte sólido puede comprender una partícula discreta que puede ser esférica (por ejemplo, microesferas) o tener una forma no esférica o irregular, tal como cúbica, cuboide, piramidal, cilíndrica, cónica, oblonga o en forma de disco, y similares. Una pluralidad de soportes sólidos espaciados en una matriz puede no comprender un sustrato. Un soporte sólido puede usarse indistintamente con el término "perla".

[0129] Un soporte sólido puede referirse a un "sustrato". Un sustrato puede ser un tipo de soporte sólido. Un sustrato puede referirse a una superficie sólida o semisólida continua sobre la cual se pueden realizar los métodos de la divulgación. Un sustrato puede referirse a una matriz, un cartucho, un chip, un dispositivo y una diapositiva, por ejemplo.

[0130] Como se usa en este documento, el término "etiqueta espacial" puede referirse a una etiqueta que puede asociarse con una posición en el espacio.

[0131] Como se usa en el presente documento, el término "código de barras estocástico" puede referirse a una secuencia de polinucleótidos que comprende etiquetas. Un código de barras estocástico puede ser una secuencia de polinucleótidos que puede usarse para códigos de barras estocásticos. Se pueden utilizar códigos de barras estocásticos para cuantificar objetivos dentro de una muestra. Los códigos de barras estocásticos se pueden utilizar para controlar los errores que pueden ocurrir después de asociar una etiqueta con un objetivo. Por ejemplo, se puede utilizar un código de barras estocástico para evaluar errores de amplificación o secuenciación. Un código de barras estocástico asociado con un objetivo puede denominarse objetivo de código de barras estocástico o objetivo de etiqueta de código de barras estocástico.

[0132] Como se usa en el presente documento, el término "código de barras estocástico específico de gen" puede referirse a una secuencia de polinucleótidos que comprende etiquetas y una región de unión a diana que es específica de gen. Un código de barras estocástico puede ser una secuencia de polinucleótidos que puede usarse para códigos de barras estocásticos. Se pueden utilizar códigos de barras estocásticos para cuantificar objetivos dentro de una muestra. Los códigos de barras estocásticos se pueden utilizar para controlar los errores que pueden ocurrir después de asociar una etiqueta con un objetivo. Por ejemplo, se puede utilizar un código de barras estocástico para evaluar errores de amplificación o secuenciación. Un código de barras estocástico asociado con un objetivo puede denominarse objetivo de código de barras estocástico o objetivo de etiqueta de código de barras estocástico.

[0133] Como se utiliza en el presente documento, el término "código de barras estocástico" puede referirse al etiquetado aleatorio (por ejemplo, código de barras) de ácidos nucleicos. Los códigos de barras estocásticos pueden utilizar una estrategia recursiva de Poisson para asociar y cuantificar etiquetas asociadas con objetivos. Tal como se utiliza en el presente documento, el término "códigos de barras estocásticos" se puede utilizar indistintamente con "códigos de barras estocásticos específicos de genes".

[0134] Como se usa en este documento, el término "objetivo" puede referirse a una composición que puede asociarse con un código de barras estocástico. Los objetivos adecuados a modo de ejemplo para el análisis mediante los métodos, dispositivos y sistemas divulgados incluyen oligonucleótidos, ADN, ARN, ARNm, microARN, ARNt y similares. Los objetivos pueden ser monocatenarios o bicatenarios. En algunas formas de realización, las dianas pueden ser proteínas. En algunas formas de realización, los objetivos son lípidos.

[0135] Como se usa en el presente documento, el término "transcriptasas inversas" puede referirse a un grupo de enzimas que tienen actividad transcriptasa inversa (es decir, que catalizan la síntesis de ADN a partir de una plantilla de ARN). En general, tales enzimas incluyen, pero no se limitan a transcriptasa inversa retroviral, transcriptasa inversa de retrotransposón, transcriptasas inversas de retroplasmídicos, transcriptasas inversas de retrones, transcriptasas inversas bacterianas, transcriptasa inversa derivada de intrones del grupo II y mutantes, variantes o derivados de los mismos. Las transcriptasas inversas no retrovirales incluyen transcriptasas inversas de retrotransposones no LTR, transcriptasas inversas de retroplasmidos, transciptasas inversas de retrones y transcriptasas inversas de intrones del grupo II. Ejemplos de transcriptasas inversas de intrones del grupo II incluyen la transcriptasa inversa de intrones *Lactococcus lactis* LI.LtrB, la transcriptasa inversa de intrones *Thermosynechococcus* alarga Tel4c o la transcriptasa inversa de intrones *Geobacillus stearothermophilus GsI-IIC.* Otras clases de transcriptasas inversas pueden incluir muchas clases de transcriptasas inversas no retrovirales (es decir, retrones, intrones del grupo II y retroelementos generadores de diversidad, entre otros).

[0136] Los términos "cebador adaptador universal", "adaptador de cebador universal" o "secuencia adaptadora universal" se usan indistintamente para referirse a una secuencia de nucleótidos que se puede usar para hibridar códigos de barras estocásticos para generar códigos de barras estocásticos específicos de genes. Una secuencia de adaptador universal puede ser, por ejemplo, una secuencia conocida que es universal en todos los códigos de barras estocásticos utilizados en los métodos de la divulgación. Por ejemplo, cuando se marcan múltiples objetivos utilizando los métodos descritos en el presente documento, cada una de las secuencias específicas del objetivo puede unirse a la misma secuencia adaptadora universal. En algunas formas de realización, se pueden usar más de una secuencia adaptadora universal en los métodos divulgados en este documento. Por ejemplo, cuando se marcan múltiples objetivos utilizando los métodos descritos en el presente documento, al menos dos de las secuencias específicas del objetivo están unidas a diferentes secuencias adaptadoras universales. Se pueden incluir un cebador adaptador universal y su complemento en dos oligonucleótidos, uno de los cuales comprende una secuencia específica de la diana y el otro comprende un código de barras estocástico. Por ejemplo, una secuencia adaptadora universal puede ser parte de un oligonucleótido que comprende una secuencia específica de la diana para generar una secuencia de nucleótidos que es complementaria a un ácido nucleico diana. Un segundo oligonucleótido que comprende un código de barras estocástico y una secuencia complementaria de la secuencia adaptadora universal puede hibridarse con la secuencia de nucleótidos y generar un código de barras estocástico específico de la diana. En algunas formas de realización, un cebador adaptador universal tiene una secuencia que es diferente de un cebador de PCR universal usado en los métodos de esta divulgación.

[0137] En el presente documento se describen métodos y sistemas para detectar y/o corregir errores ocurridos durante la PCR y/o secuenciación. Los tipos de errores pueden variar, por ejemplo, incluir, entre otros, errores de sustitución (una o más bases) y errores de no sustitución. Entre los errores de sustitución, los errores de sustitución de una base pueden ocurrir con mucha más frecuencia que aquellos que están separados por más de una base. Los métodos y sistemas se pueden utilizar, por ejemplo, para proporcionar un recuento preciso de objetivos moleculares mediante códigos de barras estocásticos.

Códigos de barras estocásticos

[0138] Los códigos de barras estocásticos se han descrito, por ejemplo, en US20150299784, WO2015031691 y Fu et al, Proc Natl Acad Sci EE. UU. 31 de mayo de 2011;108(22):9026-31. Brevemente, un código de barras estocástico puede ser una secuencia de polinucleótidos que puede usarse para etiquetar estocásticamente (por ejemplo, código de barras, etiqueta) un objetivo. Un código de barras estocástico puede comprender una o más etiquetas. Los marcadores ejemplares pueden incluir un marcador universal, un marcador celular, una etiqueta molecular, un marcador de muestra, un marcador de placa, un marcador espacial y/o un marcador preespacial. FIG. 1 ilustra un código de barras estocástico ejemplar 104 con una etiqueta espacial. El código de barras estocástico 104 puede comprender una amina 5' que puede unir el código de barras estocástico a un soporte sólido 105. El código de barras estocástico puede comprender una etiqueta universal, una etiqueta de dimensión, una etiqueta espacial, una etiqueta de célula y/o una etiqueta molecular. El orden de las diferentes etiquetas (incluidas, entre otras, la etiqueta universal, la etiqueta de dimensión, la etiqueta espacial, la etiqueta de célula y la etiqueta de molécula) en el código de barras estocástico puede variar. Por ejemplo, como se muestra en la FIG. 1, el marcador universal puede ser el marcador situado más en 5' y el marcador molecular puede ser el marcador situado más en 3'. La etiqueta espacial, la etiqueta de dimensión y la etiqueta de célula pueden estar en cualquier orden. En algunas formas de realización, la etiqueta universal, la etiqueta espacial, la etiqueta dimensional, la etiqueta celular y la etiqueta molecular están en cualquier orden.

[0139] Una etiqueta, por ejemplo la etiqueta celular, puede comprender un conjunto único de subsecuencias de ácido nucleico de longitud definida, por ejemplo, siete nucleótidos cada una (equivalente al número de bits utilizados en algunos códigos de corrección de errores de Hamming), que pueden ser diseñado para proporcionar capacidad de corrección de errores. El conjunto de subsecuencias de corrección de errores comprende siete secuencias de nucleótidos que pueden diseñarse de manera que cualquier combinación por pares de secuencias en el conjunto exhiba una "distancia genética" definida (o número de bases no coincidentes), por ejemplo, un conjunto de subsecuencias de corrección de errores. Se pueden diseñar secuencias para exhibir una distancia genética de tres nucleótidos. En este caso, la revisión de las secuencias de corrección de errores en el conjunto de datos de secuencia para moléculas de ácido nucleico diana marcadas (descritas más completamente a continuación) puede permitir detectar o corregir errores de amplificación o secuenciación. En algunas formas de realización, la longitud de las subsecuencias de ácido nucleico utilizadas para crear códigos de corrección de errores puede variar, por ejemplo, pueden ser o ser aproximadamente 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 30, 31, 40, 50, o un número o rango entre dos de estos valores, nucleótidos de longitud. En algunas formas de realización, se pueden usar subsecuencias de ácidos nucleicos de otras longitudes para crear códigos de corrección de errores.

[0140] El código de barras estocástico puede comprender una región de unión al objetivo. La región de unión al objetivo puede interactuar con un objetivo en una muestra. La diana puede ser, o comprender, ácidos ribonucleicos (ARN), ARN mensajeros (ARNm), microARN, pequeños ARN interferentes (ARNip), productos de degradación de ARN, ARN que comprenden cada uno una cola poli(A), o cualquier combinación de los mismos. En algunas formas de realización, la pluralidad de objetivos puede incluir ácidos desoxirribonucleicos (ADN).

[0141] En algunas formas de realización, una región de unión a diana puede comprender una secuencia oligo(dT) que puede interactuar con colas poli(A) de ARNm. Una o más de las etiquetas del código de barras estocástico (p. ej., la etiqueta universal, la etiqueta de dimensión, la etiqueta espacial, la etiqueta de célula y la etiqueta molecular) pueden separarse mediante un espaciador de otra una o dos de las etiquetas restantes de el código de barras estocástico. El espaciador puede ser, por ejemplo, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19 o 20 o más nucleótidos. En algunas formas de realización, ninguna de las etiquetas del código de barras estocástico está separada por un espaciador.

Etiquetas universales

[0142] Un código de barras estocástico puede comprender una o más etiquetas universales. En algunas formas de realización, una o más etiquetas universales pueden ser las mismas para todos los códigos de barras estocásticos en el conjunto de códigos de barras estocásticos unidos a un soporte sólido determinado. En algunas formas de realización, una o más etiquetas universales pueden ser las mismas para todos los códigos de barras estocásticos unidos a una pluralidad de cuentas. En algunas formas de realización, un marcador universal puede comprender una secuencia de ácido nucleico que es capaz de hibridarse con un cebador de secuenciación. Los cebadores de secuenciación se pueden utilizar para secuenciar códigos de barras estocásticos que comprenden una etiqueta universal. Los cebadores de secuenciación (por ejemplo, cebadores de secuenciación universales) pueden comprender cebadores de secuenciación asociados con plataformas de secuenciación de alto rendimiento. En algunas formas de realización, un marcador universal puede comprender una secuencia de ácido nucleico que es capaz de hibridarse con un cebador de PCR. En algunas formas de realización, el marcador universal puede comprender una secuencia de ácido nucleico que es capaz de hibridarse con un cebador de secuenciación y un cebador de PCR. La secuencia de ácido nucleico del marcador universal que es capaz de hibridarse con un cebador de secuenciación o de PCR puede denominarse sitio de unión del cebador. Una etiqueta universal puede comprender una secuencia que puede usarse para iniciar la transcripción del código de barras estocástico. Una etiqueta universal puede comprender una secuencia que puede usarse para la extensión del código de barras estocástico o una región dentro del código de barras estocástico. Una etiqueta universal puede ser, o ser aproximadamente, 1, 2, 3, 4, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, o un número o rango entre dos de estos valores., nucleótidos de longitud. Por ejemplo, un marcador universal puede comprender al menos aproximadamente 10 nucleótidos. Un marcador universal puede tener al menos, o como máximo, 1, 2, 3, 4, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 100, 200 o 300 nucleótidos de longitud. En algunas formas de realización, un conector escindible o un nucleótido modificado puede ser parte de la secuencia marcadora universal para permitir que el código de barras estocástico se escinda del soporte.

Etiquetas de dimensiones

[0143] Un código de barras estocástico puede comprender una o más etiquetas de dimensión. En algunas formas de realización, una etiqueta de dimensión puede comprender una secuencia de ácido nucleico que proporciona información sobre una dimensión en la que ocurrió el marcaje estocástico. Por ejemplo, una etiqueta de dimensión puede proporcionar información sobre el momento en que un objetivo recibió un código de barras estocástico. Una etiqueta de dimensión se puede asociar con un tiempo de código de barras estocástico en una muestra. Se puede activar una etiqueta de dimensión en el momento del etiquetado estocástico. Se pueden activar diferentes etiquetas de dimensiones en diferentes momentos. La etiqueta de dimensión proporciona información sobre el orden en el que se codificaron estocásticamente los objetivos, grupos de objetivos y/o muestras. Por ejemplo, una población de células puede tener un código de barras estocástico en la fase GO del ciclo celular. Las células se pueden pulsar nuevamente con códigos de barras estocásticos

en la fase G1 del ciclo celular. Las células pueden ser pulsadas nuevamente con códigos de barras estocásticos en la fase S del ciclo celular, y así sucesivamente. Los códigos de barras estocásticos en cada pulso (por ejemplo, cada fase del ciclo celular) pueden comprender etiquetas de diferentes dimensiones. De esta manera, la etiqueta de dimensión proporciona información sobre qué objetivos se etiquetaron en qué fase del ciclo celular. Las etiquetas de dimensiones pueden interrogar muchos tiempos biológicos diferentes. Los tiempos biológicos ejemplares pueden incluir, entre otros, el ciclo celular, la transcripción (por ejemplo, el inicio de la transcripción) y la degradación de la transcripción. En otro ejemplo, una muestra (por ejemplo, una célula, una población de células) puede marcarse estocásticamente antes y/o después del tratamiento con un fármaco y/o terapia. Los cambios en el número de copias de objetivos distintos pueden ser indicativos de la respuesta de la muestra al fármaco y/o a la terapia.

[0144] Se puede activar una etiqueta de dimensión. Una etiqueta de dimensión activable se puede activar en un momento específico. La etiqueta activable puede estar, por ejemplo, activada constitutivamente (por ejemplo, no apagada). La etiqueta de dimensión activable puede activarse, por ejemplo, de forma reversible (por ejemplo, la etiqueta de dimensión activable puede activarse y desactivarse). La etiqueta de dimensión puede ser, por ejemplo, activable de forma reversible al menos 1, 2, 3, 4, 5, 6, 7, 8, 9 o 10 o más veces. La etiqueta de dimensión puede activarse de forma reversible, por ejemplo, al menos 1, 2, 3, 4, 5, 6, 7, 8, 9 o 10 o más veces. En algunas formas de realización, la etiqueta de dimensión se puede activar con fluorescencia, luz, un evento químico (por ejemplo, escisión, ligadura de otra molécula, adición de modificaciones (por ejemplo, pegilado, sumoilado, acetilado, metilado, desacetilado, desmetilado), un evento fotoquímico. (por ejemplo, fotoenjaulamiento) e introducción de un nucleótido no natural.

[0145] La etiqueta de dimensión puede, en algunas formas de realización, ser idéntica para todos los códigos de barras estocásticos unidos a un soporte sólido dado (por ejemplo, cuenta), pero diferente para diferentes soportes sólidos (por ejemplo, cuentas). En algunas formas de realización, al menos el 60 %, 70 %, 80 %, 85 %, 90 %, 95 %, 97 %, 99 % o 100 % de los códigos de barras estocásticos en el mismo soporte sólido pueden comprender la misma etiqueta de dimensión. En algunas formas de realización, al menos el 60 % de los códigos de barras estocásticos sobre el mismo soporte sólido pueden comprender la misma etiqueta de dimensión. En algunas formas de realización, al menos el 95 % de los códigos de barras estocásticos sobre el mismo soporte sólido pueden comprender la misma etiqueta de dimensión.

[0146] Puede haber hasta 10^6 o más secuencias de etiquetas de dimensiones únicas representadas en una pluralidad de soportes sólidos (por ejemplo, perlas). Una etiqueta de dimensión puede ser aproximadamente 1, 2, 3, 4, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, o un número o rango entre dos de estos valores. nucleótidos de longitud. Una etiqueta de dimensión puede tener al menos, o como máximo, 1, 2, 3, 4, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 100, 200 o 300 nucleótidos de longitud. Una etiqueta de dimensión puede comprender entre aproximadamente 5 y aproximadamente 200 nucleótidos. Una etiqueta de dimensión puede comprender entre aproximadamente 10 y aproximadamente 150 nucleótidos. Una etiqueta de dimensión puede comprender entre aproximadamente 20 y aproximadamente 125 nucleótidos de longitud.

Etiquetas espaciales

[0147] Un código de barras estocástico puede comprender una o más etiquetas espaciales. En algunas formas de realización, una etiqueta espacial puede comprender una secuencia de ácido nucleico que proporciona información sobre la orientación espacial de una molécula diana que está asociada con el código de barras estocástico. Una etiqueta espacial se puede asociar con una coordenada en una muestra. La coordenada puede ser una coordenada fija. Por ejemplo, se puede fijar una coordenada con referencia a un sustrato. Una etiqueta espacial puede hacer referencia a una cuadrícula bidimensional o tridimensional. Una coordenada se puede fijar en referencia a un punto de referencia. El hito puede ser identificable en el espacio. Un punto de referencia puede ser una estructura de la que se pueden visualizar imágenes. Un hito puede ser una estructura biológica, por ejemplo un hito anatómico. Un punto de referencia puede ser un punto de referencia celular, por ejemplo un orgánulo. Un punto de referencia puede ser un punto de referencia no natural, como una estructura con un identificador identificable, como un código de color, un código de barras, una propiedad magnética, fluorescentes, radiactividad o un tamaño o forma únicos. Una etiqueta espacial puede asociarse con una partición física (por ejemplo, un pocillo, un contenedor o una gota). En algunas formas de realización, se usan múltiples etiquetas espaciales juntas para codificar una o más posiciones en el espacio.

[0148] La etiqueta espacial puede ser idéntica para todos los códigos de barras estocásticos unidos a un soporte sólido determinado (por ejemplo, cuentas), pero diferente para diferentes soportes sólidos (por ejemplo, cuentas). En algunas formas de realización, el porcentaje de códigos de barras estocásticos en el mismo soporte sólido que comprende la misma etiqueta espacial puede ser, o ser aproximadamente, 60 %, 70 %, 80 %, 85 %, 90 %, 95 %, 97 %, 99 %. 100 %, o un número o rango entre dos de estos valores. En algunas formas de realización, el porcentaje de códigos de barras estocásticos sobre el mismo soporte sólido que comprende la misma etiqueta espacial puede ser al menos, o como máximo, 60 %, 70 %, 80 %, 85 %, 90 %, 95 %, 97 %, 99 %, o 100 %. En algunas formas de realización, al menos el 60 % de los códigos de barras estocásticos sobre el mismo soporte sólido pueden comprender la misma etiqueta espacial. En algunas formas de realización, al menos el 95 % de los códigos de barras estocásticos sobre el mismo soporte sólido pueden comprender la misma etiqueta espacial.

[0149] Puede haber hasta 10^6 o más secuencias de etiquetas espaciales únicas representadas en una pluralidad de soportes sólidos (por ejemplo, perlas). Una etiqueta espacial puede ser, o ser aproximadamente, 1, 2, 3, 4, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, o un número o rango entre dos de estos valores., nucleótidos de longitud. Una etiqueta espacial

puede tener al menos o como máximo 1, 2, 3, 4, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 100, 200 o 300 nucleótidos de longitud. Una etiqueta espacial puede comprender entre aproximadamente 5 y aproximadamente 200 nucleótidos. Una etiqueta espacial puede comprender entre aproximadamente 10 y aproximadamente 150 nucleótidos. Una etiqueta espacial puede comprender entre aproximadamente 20 y aproximadamente 125 nucleótidos de longitud.

Etiquetas de célula

[0150] Un código de barras estocástico puede comprender una o más etiquetas de célula. En algunas formas de realización, un marcador celular puede comprender una secuencia de ácido nucleico que proporciona información para determinar qué ácido nucleico diana se originó a partir de qué célula. En algunas formas de realización, la etiqueta de la célula es idéntica para todos los códigos de barras estocásticos unidos a un soporte sólido determinado (por ejemplo, cuentas), pero diferente para diferentes soportes sólidos (por ejemplo, cuentas). En algunas formas de realización, el porcentaje de códigos de barras estocásticos en el mismo soporte sólido que comprende la misma etiqueta celular puede ser, o ser aproximadamente 60 %, 70 %, 80 %, 85 %, 90 %, 95 %, 97 %, 99 %, 100 %, o un número o rango entre dos de estos valores. En algunas formas de realización, el porcentaje de códigos de barras estocásticos en el mismo soporte sólido que comprende la misma etiqueta celular puede ser, o ser aproximadamente 60 %, 70 %, 80 %, 85 %, 90 %, 95 %, 97 %, 99 %, o 100 %. Por ejemplo, al menos el 60 % de los códigos de barras estocásticos sobre el mismo soporte sólido pueden comprender la misma etiqueta de célula. Como otro ejemplo, al menos el 95 % de los códigos de barras estocásticos sobre el mismo soporte sólido pueden comprender la misma etiqueta de célula.

[0151] Puede haber hasta 10^6 o más secuencias de etiquetas celulares únicas representadas en una pluralidad de soportes sólidos (por ejemplo, perlas). Una etiqueta de célula puede ser, o ser aproximadamente, 1, 2, 3, 4, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, o un número o rango entre dos de estos valores., nucleótidos de longitud. Un marcador celular puede tener al menos, o como máximo, 1, 2, 3, 4, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 100, 200 o 300 nucleótidos de longitud. Por ejemplo, un marcador celular puede comprender entre aproximadamente 5 y aproximadamente 200 nucleótidos. Como otro ejemplo, un marcador celular puede comprender entre aproximadamente 10 y aproximadamente 150 nucleótidos. Como otro ejemplo más, un marcador celular puede comprender entre aproximadamente 20 y aproximadamente 125 nucleótidos de longitud.

Etiquetas moleculares

[0152] Un código de barras estocástico puede comprender una o más etiquetas moleculares. En algunas formas de realización, una etiqueta molecular puede comprender una secuencia de ácido nucleico que proporciona información de identificación para el tipo específico de especie de ácido nucleico diana hibridada con el código de barras estocástico. Una etiqueta molecular puede comprender una secuencia de ácido nucleico que proporciona un contador para la aparición específica de la especie de ácido nucleico objetivo hibridada con el código de barras estocástico (por ejemplo, región de unión al objetivo).

[0153] En algunas formas de realización, un conjunto diverso de etiquetas moleculares se une a un soporte sólido determinado (por ejemplo, una perla). En algunas formas de realización, puede haber, o haber aproximadamente, 102, 103, 104, 105, 106, 107, 108, 109, o un número o rango de secuencias de etiquetas moleculares únicas. Por ejemplo, una pluralidad de códigos de barras estocásticos puede comprender aproximadamente 6561 etiquetas moleculares con secuencias distintas. Como otro ejemplo, una pluralidad de códigos de barras estocásticos puede comprender aproximadamente 65536 etiquetas moleculares con secuencias distintas. En algunas formas de realización, puede haber al menos, o como máximo, 102, 103, 104, 105, 10^6, 10^7, 10^8 o 10^9 secuencias de etiquetas moleculares únicas. Las secuencias de etiquetas moleculares únicas unidas a un soporte sólido determinado (p. ej., perla).

[0154] Una etiqueta molecular puede ser, o ser aproximadamente, 1, 2, 3, 4, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, o un número o un rango entre dos cualesquiera. de estos valores, nucleótidos de longitud. Una etiqueta molecular puede tener al menos, o como máximo, 1, 2, 3, 4, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 100, 200 o 300 nucleótidos de longitud.

Región de enlace objetivo

[0155] Un código de barras estocástico puede comprender una o más regiones de unión objetivo. En algunas formas de realización, una región de unión a diana puede hibridarse con una diana de interés. En algunas formas de realización, las regiones de unión a diana pueden comprender una secuencia de ácido nucleico que se hibrida específicamente con una diana (por ejemplo, ácido nucleico diana, molécula diana, por ejemplo, un ácido nucleico celular a analizar), por ejemplo con una secuencia genética específica. En algunas formas de realización, una región de unión diana puede comprender una secuencia de ácido nucleico que puede unirse (por ejemplo, hibridarse) a una ubicación específica de un ácido nucleico diana específico. En algunas formas de realización, la región de unión diana puede comprender una secuencia de ácido nucleico que es capaz de hibridación específica con un sitio saliente de enzima de restricción (por ejemplo, un saliente de extremo pegajoso de EcoRI). El código de barras estocástico puede entonces ligarse a cualquier molécula de ácido nucleico que comprenda una secuencia complementaria al saliente del sitio de restricción.

[0156] En algunas formas de realización, una región de unión diana puede comprender una secuencia de ácido nucleico diana no específica. Una secuencia de ácido nucleico diana no específica puede referirse a una secuencia que puede

unirse a múltiples ácidos nucleicos diana, independientemente de la secuencia específica del ácido nucleico diana. Por ejemplo, la región de unión diana puede comprender una secuencia multimérica aleatoria o una secuencia oligo(dT) que se hibrida con la cola poli(A) de las moléculas de ARNm. Una secuencia de multímero aleatorio puede ser, por ejemplo, un dímero, trímero, cuatrámero, pentámero, hexámero, septámero, octámero, nonámero, decámero o secuencia de multímero superior aleatorio de cualquier longitud. En algunas formas de realización, la región de unión objetivo es la misma para todos los códigos de barras estocásticos unidos a una perla determinada. En algunas formas de realización, las regiones de unión a diana para la pluralidad de códigos de barras estocásticos unidos a una perla determinada pueden comprender dos o más secuencias de unión a diana diferentes. Una región de unión diana puede ser, o ser aproximadamente, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, o un número o rango entre dos cualesquiera de estos valores, nucleótidos de longitud. Una región de unión diana puede tener como máximo aproximadamente 5, 10, 15, 20, 25, 30, 35, 40, 45, 50 o más nucleótidos de longitud.

[0157] En algunas formas de realización, una región de unión a diana puede comprender un oligo(dT) que puede hibridarse con ARNm que comprenden extremos poliadenilados. Una región de unión a diana puede ser específica de un gen. Por ejemplo, una región de unión a diana se puede configurar para hibridar con una región específica de una diana. Una región de unión a la diana puede ser, o ser aproximadamente, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26 27, 28, 29, 30, o un número o rango entre dos de estos valores cualesquiera, nucleótidos de longitud. Una región de unión a diana puede ser al menos, o como máximo, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26 27, 28, 29 o 30 nucleótidos de longitud. Una región de unión al objetivo puede tener una longitud de aproximadamente 5 a 30 nucleótidos. Cuando un código de barras estocástico comprende una región de unión a diana específica de un gen, el código de barras estocástico puede denominarse código de barras estocástico específico de un gen.

Propiedad de orientación

[0158] Un código de barras estocástico puede comprender una o más propiedades de orientación que pueden usarse para orientar (por ejemplo, alinear) los códigos de barras estocásticos. Un código de barras estocástico puede comprender un resto para enfoque isoeléctrico. Diferentes códigos de barras estocásticos pueden comprender diferentes puntos de enfoque isoeléctrico. Cuando estos códigos de barras estocásticos se introducen en una muestra, la muestra puede someterse a un enfoque isoeléctrico para orientar los códigos de barras estocásticos de una manera conocida. De esta manera, la propiedad de orientación se puede utilizar para desarrollar un mapa conocido de códigos de barras estocásticos en una muestra. Las propiedades de orientación ejemplares pueden incluir movilidad electroforética (por ejemplo, basada en el tamaño del código de barras estocástico), punto isoeléctrico, espín, conductividad y/o autoensamblaje. Por ejemplo, los códigos de barras estocásticos con una propiedad de orientación de autoensamblaje pueden autoensamblarse en una orientación específica (por ejemplo, nanoestructura de ácido nucleico) tras la activación.

Propiedad de afinidad

[0159] Un código de barras estocástico puede comprender una o más propiedades de afinidad. Por ejemplo, una etiqueta espacial puede comprender una propiedad de afinidad. Una propiedad de afinidad puede incluir un resto químico y/o biológico que puede facilitar la unión del código de barras estocástico a otra entidad (por ejemplo, un receptor celular). Por ejemplo, una propiedad de afinidad puede comprender un anticuerpo, por ejemplo, un anticuerpo específico para un resto específico (por ejemplo, receptor) en una muestra. En algunas formas de realización, el anticuerpo puede guiar el código de barras estocástico a un tipo de célula o molécula específica. Los objetivos en y/o cerca del tipo de célula o molécula específica pueden marcarse estocásticamente. La propiedad de afinidad puede, en algunas formas de realización, proporcionar información espacial además de la secuencia de nucleótidos del marcador espacial porque el anticuerpo puede guiar el código de barras estocástico a una ubicación específica. El anticuerpo puede ser un anticuerpo terapéutico, por ejemplo un anticuerpo monoclonal o un anticuerpo policlonal. El anticuerpo puede ser humanizado o quimérico. El anticuerpo puede ser un anticuerpo desnudo o un anticuerpo de fusión.

[0160] El anticuerpo puede ser una molécula de inmunoglobulina de longitud completa (es decir, de origen natural o formada mediante procesos recombinatorios de fragmentos de genes de inmunoglobulina normales) (por ejemplo, un anticuerpo IgG) o una porción inmunológicamente activa (es decir, que se une específicamente) de una molécula de inmunoglobulina, como un fragmento de anticuerpo.

[0161] El fragmento de anticuerpo puede ser, por ejemplo, una porción de un anticuerpo tal como F(ab')2, Fab', Fab, Fv, sFv y similares. En algunas formas de realización, el fragmento de anticuerpo puede unirse con el mismo antígeno que reconoce el anticuerpo de longitud completa. El fragmento de anticuerpo puede incluir fragmentos aislados que consisten en regiones variables de anticuerpos, tales como los fragmentos "Fv" que consisten en las regiones variables de las cadenas pesada y ligera y moléculas polipeptídicas de hebra sencilla recombinantes en las que las regiones variables ligeras y pesadas están conectadas por un conector peptídico ("proteínas scFv"). Los anticuerpos ejemplares pueden incluir, entre otros, anticuerpos para células cancerosas, anticuerpos para virus, anticuerpos que se unen a receptores de la superficie celular (CD8, CD34, CD45) y anticuerpos terapéuticos.

Cebador del adaptador universal

[0162] Un código de barras estocástico puede comprender uno o más cebadores adaptadores universales. Por ejemplo, un código de barras estocástico específico de un gen puede comprender un cebador adaptador universal. Un cebador adaptador universal puede referirse a una secuencia de nucleótidos que es universal en todos los códigos de barras estocásticos. Se puede utilizar un cebador adaptador universal para crear códigos de barras estocásticos específicos de genes. Un cebador adaptador universal puede ser, o ser aproximadamente, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26 27, 28, 29, 30, o un número o rango entre dos cualesquiera de estos nucleótidos en longitud. Un cebador adaptador universal puede ser al menos, o como máximo, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19., 20, 21, 22, 23, 24, 25, 26 27, 28, 29 o 30 nucleótidos de longitud. Un cebador adaptador universal puede tener una longitud de 5 a 30 nucleótidos.

Soportes sólidos

[0163] Los códigos de barras estocásticos descritos en el presente documento pueden, en algunas formas de realización, asociarse con un soporte sólido. El soporte sólido puede ser, por ejemplo, una partícula sintética. En algunas formas de realización, algunos o todos las etiquetas moleculares (por ejemplo, los primeras etiquetas moleculares) de una pluralidad de códigos de barras estocásticos (por ejemplo, la primera pluralidad de códigos de barras estocásticos) sobre un soporte sólido difieren en al menos un nucleótido. Las etiquetas de las células de los códigos de barras estocásticos sobre un mismo soporte sólido pueden ser las mismas. Las etiquetas celulares de los códigos de barras estocásticos sobre diferentes soportes sólidos pueden diferir en al menos un nucleótido. Por ejemplo, las primeras etiquetas de célula de una primera pluralidad de códigos de barras estocásticos sobre un primer soporte sólido pueden tener la misma secuencia, y las segundas etiquetas de célula de una segunda pluralidad de códigos de barras estocásticos sobre un segundo soporte sólido pueden tener la misma secuencia. Las primeras etiquetas de célula de la primera pluralidad de códigos de barras estocásticos en el primer soporte sólido y las segundas etiquetas de célula de la segunda pluralidad de códigos de barras estocásticos en el segundo soporte sólido pueden diferir en al menos un nucleótido. Una etiqueta celular puede tener, por ejemplo, aproximadamente 5-20 nucleótidos de longitud. Una etiqueta molecular puede tener, por ejemplo, una longitud de aproximadamente 5 a 20 nucleótidos. La partícula sintética puede ser, por ejemplo, una perla.

[0164] La perla puede ser, por ejemplo, una perla de gel de sílice, una perla de vidrio de poro controlado, una perla magnética, una perla Dynabead, una perla de Sephadex/Sepharose, una perla de celulosa, una perla de poliestireno o cualquier combinación de las mismas. La perla puede comprender un material tal como polidimetilsiloxano (PDMS), poliestireno, vidrio, polipropileno, agarosa, gelatina, hidrogel, paramagnético, cerámico, plástico, vidrio, metilestireno, polímero acrílico, titanio, látex, sefarosa, celulosa, nailon, silicona o cualquier combinación de los mismos.

[0165] En algunas formas de realización, la perla puede ser una perla polimérica, por ejemplo una perla deformable o una perla de gel, funcionalizada con códigos de barras estocásticos (tales como perlas de gel de 10X Genomics (San Francisco, CA). En alguna implementación, una perla de gel puede comprender geles a base de polímeros. Las perlas de gel se pueden generar, por ejemplo, encapsulando uno o más precursores poliméricos en gotitas. Tras la exposición de los precursores poliméricos a un acelerador (por ejemplo, tetrametiletilendiamina (TEMED)), se puede generar una perla de gel.

[0166] En algunas formas de realización, la perla polimérica puede disolverse, fundirse o degradarse, por ejemplo, bajo una condición deseada. La condición deseada puede incluir una condición ambiental. La condición deseada puede dar como resultado que la perla polimérica se disuelva, funda o degrade de manera controlada. Una perla de gel puede disolverse, derretirse o degradarse debido a un estímulo químico, un estímulo físico, un estímulo biológico, un estímulo térmico, un estímulo magnético, un estímulo eléctrico, un estímulo luminoso o cualquier combinación de los mismos.

[0167] Los analitos y/o reactivos, tales como códigos de barras de oligonucleótidos, por ejemplo, pueden acoplarse/inmovilizarse a la superficie interior de una perla de gel (por ejemplo, el interior accesible mediante difusión de un código de barras de oligonucleótidos y/o materiales usados para generar un código de barras de oligonucleótido) y/o la superficie exterior de una perla de gel o cualquier otra microcápsula descrita en el presente documento. El acoplamiento/inmovilización puede realizarse mediante cualquier forma de enlace químico (por ejemplo, enlace covalente, enlace iónico) o fenómenos físicos (por ejemplo, fuerzas de Van der Waals, interacciones dipolo-dipolo, etc.). En algunos casos, el acoplamiento/inmovilización de un reactivo a una perla de gel o cualquier otra microcápsula descrita en el presente documento puede ser reversible, tal como, por ejemplo, mediante un resto lábil (por ejemplo, mediante un entrecruzante químico, incluidos los entrecruzadores químicos descritos en este documento). Tras la aplicación de un estímulo, la fracción lábil puede escindirse y el reactivo inmovilizado puede liberarse. En algunos casos, el resto lábil es un enlace disulfuro. Por ejemplo, en el caso en el que un código de barras de oligonucleótido se inmoviliza en una perla de gel mediante un enlace disulfuro, la exposición del enlace disulfuro a un agente reductor puede escindir el enlace disulfuro y liberar el código de barras de oligonucleótido de la perla. El resto lábil puede incluirse como parte de una perla de gel o microcápsula, como parte de un conector químico que une un reactivo o analito a una perla de gel o microcápsula, y/o como parte de un reactivo o analito.

[0168] En algunas formas de realización, una perla de gel puede comprender una amplia gama de polímeros diferentes que incluyen, entre otros: polímeros, polímeros sensibles al calor, polímeros fotosensibles, polímeros magnéticos,

polímeros sensibles al pH, polímeros sensibles a las sales, polímeros químicamente sensibles, polielectrolitos, polisacáridos, péptidos, proteínas y/o plásticos. Los polímeros pueden incluir, entre otros, materiales tales como poli(N-isopropilacrilamida) (PNIPAAm), poli(sulfonato de estireno) (PSS), poli(alilamina) (PAAm), ácido poli(acrílico) (PAA), poli(etilenimina) (PEI), poli(cloruro de dialildimetilamonio) (PDADMAC), poli(pirolo) (PPy), poli(vinilpirrolidona) (PVPON), poli(vinilpiridina) (PVP), ácido poli(metacrílico) (PMAA), poli(metacrilato de metilo) (PMMA), poliestireno (PS), poli(tetrahidrofurano) (PTHF), poli(ftaladehído) (PTHF), poli(hexil viológeno) (PHV), poli(L-lisina) (PLL), poli(L-arginina) (PARG), ácido poli(láctico-co-glicólico) (PLGA)

[0169] Se pueden utilizar numerosos estímulos químicos para desencadenar la alteración o degradación de las perlas. Los ejemplos de estos cambios químicos pueden incluir, entre otros, cambios mediados por el pH en la pared de la perla, desintegración de la pared de la perla mediante escisión química de enlaces cruzados, despolimerización desencadenada de la pared de la perla y reacciones de cambio de la pared de la perla. También se pueden utilizar cambios masivos para provocar la alteración de las perlas.

[0170] Los cambios físicos o de volumen en la microcápsula a través de diversos estímulos también ofrecen muchas ventajas en el diseño de cápsulas para liberar reactivos. Los cambios físicos o de masa ocurren a escala macroscópica, en la que la ruptura de la perla es el resultado de fuerzas mecanofísicas inducidas por un estímulo. Estos procesos pueden incluir, entre otros, ruptura inducida por presión, fusión de la pared del cordón o cambios en la porosidad de la pared del cordón.

[0171] También se pueden usar estímulos biológicos para desencadenar la alteración o degradación de las perlas. Generalmente, los desencadenantes biológicos se parecen a los desencadenantes químicos, pero muchos ejemplos utilizan biomoléculas o moléculas que se encuentran comúnmente en sistemas vivos, como enzimas, péptidos, sacáridos, ácidos grasos, ácidos nucleicos y similares. Por ejemplo, las perlas pueden comprender polímeros con entrecruzamientos peptídicos que son sensibles a la escisión por proteasas específicas. Más específicamente, un ejemplo puede comprender una microcápsula que comprende enlaces cruzados peptídicos GFLGK. Tras la adición de un desencadenante biológico como la proteasa catepsina B, los enlaces cruzados peptídicos de la cubierta se escinden y se libera el contenido de las perlas. En otros casos, las proteasas pueden activarse mediante calor. En otro ejemplo, las perlas comprenden una pared de cubierta que comprende celulosa. La adición de la enzima hidrolítica quitosano sirve como desencadenante biológico para la ruptura de enlaces celulósicos, la despolimerización de la pared de la cáscara y la liberación de su contenido interno.

[0172] También se puede inducir a las perlas a liberar su contenido tras la aplicación de un estímulo térmico. Un cambio de temperatura puede provocar diversos cambios en las cuentas. Un cambio de calor puede provocar la fusión de una perla de manera que la pared de la perla se desintegre. En otros casos, el calor puede aumentar la presión interna de los componentes internos del cordón de modo que el cordón se rompa o explote. En otros casos más, el calor puede transformar la perla en un estado encogido y deshidratado. El calor también puede actuar sobre los polímeros sensibles al calor dentro de la pared de una perla para provocar la rotura de la perla.

[0173] La inclusión de nanopartículas magnéticas en la pared de perlas de las microcápsulas puede permitir la ruptura provocada de las perlas, así como guiar las perlas en una matriz. Un dispositivo de esta divulgación puede comprender perlas magnéticas para cualquier propósito. En un ejemplo, la incorporación de nanopartículas de Fe_3O_4 en perlas que contienen polielectrolitos provoca la ruptura en presencia de un estímulo de campo magnético oscilante.

[0174] Una perla también puede romperse o degradarse como resultado de una estimulación eléctrica. De manera similar a las partículas magnéticas descritas en la sección anterior, las perlas eléctricamente sensibles pueden permitir tanto la ruptura provocada de las perlas como otras funciones como la alineación en un campo eléctrico, la conductividad eléctrica o reacciones redox. En un ejemplo, se alinean perlas que contienen material eléctricamente sensible en un campo eléctrico de manera que se pueda controlar la liberación de reactivos internos. En otros ejemplos, los campos eléctricos pueden inducir reacciones redox dentro de la propia pared de la perla que pueden aumentar la porosidad.

[0175] Un estímulo ligero se puede utilizar para romper las perlas. Son posibles numerosos activadores de luz y pueden incluir sistemas que utilizan diversas moléculas, como nanopartículas y cromóforos, capaces de absorber fotones de rangos específicos de longitudes de onda. Por ejemplo, se pueden utilizar recubrimientos de óxido metálico como activadores de cápsulas. La irradiación UV de cápsulas de polielectrolito recubiertas con SiO_2 puede provocar la desintegración de la pared de la perla. En otro ejemplo más, se pueden incorporar materiales fotoconmutables tales como grupos azobenceno en la pared de la perla. Tras la aplicación de luz ultravioleta o luz visible, productos químicos como estos sufren una isomerización cis a trans reversible tras la absorción de fotones. En este aspecto, la incorporación de interruptores de fotones da como resultado una pared de perlas que puede desintegrarse o volverse más porosa al aplicar un disparador de luz.

[0176] Por ejemplo, en un ejemplo no limitante de código de barras estocástico ilustrado en la FIG. 2, después de introducir células tales como células individuales en una pluralidad de micropocillos de una matriz de micropocillos en el bloque 208, se pueden introducir perlas en la pluralidad de micropocillos de la matriz de micropocillos en el bloque 212. Cada micropocillo puede comprender una perla. Las perlas pueden comprender una pluralidad de códigos de barras estocásticos. Un código de barras estocástico puede comprender una región de amina 5' unida a una perla. El código de

barras estocástico puede comprender una etiqueta universal, una etiqueta molecular, una región de unión a diana o cualquier combinación de los mismos.

[0177] Los códigos de barras estocásticos divulgados en el presente documento pueden asociarse con (por ejemplo, unirse a) un soporte sólido (por ejemplo, una cuenta). Cada uno de los códigos de barras estocásticos asociados con un soporte sólido puede comprender una etiqueta molecular seleccionada de un grupo que comprende al menos 100 o 1000 etiquetas moleculares con secuencias únicas. En algunas formas de realización, diferentes códigos de barras estocásticos asociados con un soporte sólido pueden comprender etiquetas moleculares de diferentes secuencias. En algunas formas de realización, un porcentaje de códigos de barras estocásticos asociados con un soporte sólido comprende la misma etiqueta de célula. Por ejemplo, el porcentaje puede ser, o ser aproximadamente 60 %, 70 %, 80 %, 85 %, 90 %, 95 %, 97 %, 99 %, 100 %, o un número o un rango entre dos de estos valores. Como otro ejemplo, el porcentaje puede ser al menos o como máximo 60 %, 70 %, 80 %, 85 %, 90 %, 95 %, 97 %, 99 % o 100 %. En algunas formas de realización, los códigos de barras estocásticos asociados con un soporte sólido pueden tener la misma etiqueta de célula. Los códigos de barras estocásticos asociados con diferentes soportes sólidos pueden tener diferentes etiquetas de células seleccionadas de un grupo que comprende al menos 100 o 1000 etiquetas de células con secuencias únicas.

[0178] Los códigos de barras estocásticos divulgados en el presente documento pueden asociarse (por ejemplo, unirse a) un soporte sólido (por ejemplo, una cuenta). En algunas formas de realización, la codificación de barras estocástica de la pluralidad de objetivos en la muestra se puede realizar con un soporte sólido que incluye una pluralidad de partículas sintéticas asociadas con la pluralidad de códigos de barras estocásticos. En algunas formas de realización, el soporte sólido puede incluir una pluralidad de partículas sintéticas asociadas con la pluralidad de códigos de barras estocásticos. Las etiquetas espaciales de varios códigos de barras estocásticos sobre diferentes soportes sólidos pueden diferir en al menos un nucleótido. El soporte sólido puede, por ejemplo, incluir la pluralidad de códigos de barras estocásticos en dos dimensiones o en tres dimensiones. Las partículas sintéticas pueden ser perlas. Las perlas pueden ser perlas de gel de sílice, perlas de vidrio de poro controlado, perlas magnéticas, perlas Dynabeads, perlas de Sephadex/Sepharose, perlas de celulosa, perlas de poliestireno o cualquier combinación de las mismas. El soporte sólido puede incluir un polímero, una matriz, un hidrogel, un dispositivo de conjunto de agujas, un anticuerpo o cualquier combinación de los mismos. En algunas formas de realización, los soportes sólidos pueden flotar libremente. En algunas formas de realización, los soportes sólidos pueden estar incrustados en una matriz sólida o semisólida. Los códigos de barras estocásticos no podrán estar asociados a soportes sólidos. Los códigos de barras estocásticos pueden ser nucleótidos individuales. Los códigos de barras estocásticos se pueden asociar con un sustrato.

[0179] Como se usan en el presente documento, los términos "atado", "unido" e "inmovilizado" se usan indistintamente y pueden referirse a medios covalentes o no covalentes para unir códigos de barras estocásticos a un soporte sólido. Se puede utilizar cualquiera de una variedad de soportes sólidos diferentes como soportes sólidos para unir códigos de barras estocásticos presintetizados o para la síntesis en fase sólida *in situ* de códigos de barras estocásticos.

[0180] En algunas formas de realización, el soporte sólido es una perla. La perla puede comprender uno o más tipos de esfera, bola, cojinete, cilindro u otra configuración similar sólida, porosa o hueca en la que se pueda inmovilizar un ácido nucleico (por ejemplo, de forma covalente o no covalente). La perla puede estar compuesta, por ejemplo, de plástico, cerámica, metal, material polimérico o cualquier combinación de los mismos. Una perla puede ser, o comprender, una partícula discreta que es esférica (por ejemplo, microesferas) o que tiene una forma no esférica o irregular, tal como cúbica, cuboide, piramidal, cilíndrica, cónica, oblonga o en forma de disco, y similar. En algunas formas de realización, una perla puede tener forma no esférica.

[0181] Las perlas pueden comprender una variedad de materiales que incluyen, entre otros, materiales paramagnéticos (por ejemplo, magnesio, molibdeno, litio y tantalio), materiales superparamagnéticos (por ejemplo, nanopartículas de ferrita (Fe_3O_4; magnetita)), materiales ferromagnéticos (por ejemplo, hierro, níquel, cobalto, algunas de sus aleaciones y algunos compuestos de metales de tierras raras), cerámica, plástico, vidrio, poliestireno, sílice, metilestireno, polímeros acrílicos, titanio, látex, sefarosa, agarosa, hidrogel, polímero, celulosa, nailon o cualquier combinación de los mismos.

[0182] En algunas formas de realización, la perla (por ejemplo, la perla a la que se unen las etiquetas estocásticas) es una perla de hidrogel. En algunas formas de realización, la perla comprende hidrogel.

[0183] Algunas formas de realización descritas en el presente documento incluyen una o más partículas (por ejemplo, perlas). Cada una de las partículas puede comprender una pluralidad de oligonucleótidos (por ejemplo, códigos de barras estocásticos). Cada uno de la pluralidad de oligonucleótidos puede comprender una secuencia marcadora molecular, una secuencia marcadora celular y una región de unión a la diana (por ejemplo, una secuencia oligo dT, una secuencia específica de gen, un multímero aleatorio o una combinación de los mismos). La secuencia marcadora celular de cada uno de la pluralidad de oligonucleótidos puede ser la misma. Las secuencias marcadoras celulares de oligonucleótidos en diferentes partículas pueden ser diferentes de modo que se puedan identificar los oligonucleótidos en diferentes partículas. El número de secuencias de etiquetas celulares diferentes puede ser diferente en diferentes implementaciones. En algunas formas de realización, el número de secuencias de marcador celular puede ser, o aproximadamente 10, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000., 9000, 10000, 20000, 30000, 40000, 50000, 60000, 70000, 80000, 90000, 100000, 10^6, 10^7, 10^8, 10^9, un número o un rango entre dos de estos valores, o más. En algunas formas de realización, el número de secuencias de marcador celular puede ser al menos o como

máximo 10, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, 2000, 3000, 4000, 5000, 6000. 7000, 8000, 9000, 10000, 20000, 30000, 40000, 50000, 60000, 70000, 80000, 90000, 100000, 10^6, 10^7, 10^8 o 10^9. En algunas formas de realización, no más de 1, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000 o más de la pluralidad de partículas incluyen oligonucleótidos con la misma secuencia celular. En alguna forma de realización, la pluralidad de partículas que incluyen oligonucleótidos con la misma secuencia celular puede ser como máximo 0,1 %, 0,2 %, 0,3 %, 0,4 %, 0,5 %, 0,6 %, 0,7 %, 0,8 %, 0,9 %, 1 %, 2 %, 3 %, 4 %, 5 %, 6 %, 7 %, 8 %, 9 %, 10 % o más. En algunas formas de realización, ninguna de la pluralidad de partículas tiene la misma secuencia de marcador celular.

[0184] La pluralidad de oligonucleótidos en cada partícula puede comprender diferentes secuencias de etiquetas moleculares. En algunas formas de realización, el número de secuencias de etiquetas moleculares puede ser de aproximadamente 10, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000., 9000, 10000, 20000, 30000, 40000, 50000, 60000, 70000, 80000, 90000, 100000, 10^6, 10^7, 10^8, 10^9, o un número o rango entre dos de estos valores. En algunas formas de realización, el número de secuencias de etiquetas moleculares puede ser al menos o como máximo 10, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, 2000, 3000, 4000, 5000, 6000. 7000, 8000, 9000, 10000, 20000, 30000, 40000, 50000, 60000, 70000, 80000, 90000, 100000, 10^6, 10^7, 10^8 o 10^9. Por ejemplo, al menos 100 de la pluralidad de oligonucleótidos comprenden diferentes etiquetas moleculares secuencias. Como otro ejemplo de ejemplo, en una sola partícula, al menos 100, 500, 1000, 5000, 10000, 15000, 20000, 50000, un número o un rango entre dos cualesquiera de estos valores, o más de la pluralidad de oligonucleótidos comprenden diferentes secuencias de etiquetas moleculares. Algunas formas de realización proporcionan una pluralidad de partículas que comprenden códigos de barras estocásticos. En algunas formas de realización, la proporción de una aparición (o una copia o un número) de un objetivo a etiquetar y las diferentes secuencias de etiquetas moleculares pueden ser al menos 1:1, 1:2, 1:3, 1:4, 1:5, 1:6, 1:7, 1:8, 1:9, 1:10, 1:11, 1:12, 1:13, 1:14, 1:15, 1:16, 1:17, 1:18, 1:19, 1:20, 1:30, 1:40, 1:50, 1:60, 1: 70, 1:80, 1:90 o más. En algunas formas de realización, cada uno de la pluralidad de oligonucleótidos comprende además un marcador de muestra, un marcador universal o ambos. La partícula puede ser, por ejemplo, una nanopartícula o micropartícula.

[0185] El tamaño de las perlas puede variar. Por ejemplo, el diámetro de la perla puede oscilar entre 0,1 micrómetros y 50 micrómetros. En algunas formas de realización, los diámetros de las perlas pueden ser, o ser aproximadamente, 0,1, 0,5, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50 micrómetros o un número o un rango entre dos de estos valores.

[0186] Los diámetros de la perla pueden estar relacionados con el diámetro de los pocillos del sustrato. En algunas formas de realización, los diámetros de la perla pueden ser, o ser aproximadamente, 10 %, 20 %, 30 %, 40 %, 50 %, 60 %, 70 %, 80 %, 90 %, 100 %, o un número o un rango entre dos de estos valores, más largo o más corto que el diámetro del pocillo. El diámetro de las perlas puede estar relacionado con el diámetro de una célula (por ejemplo, una única célula atrapada por un pocillo del sustrato). En algunas formas de realización, los diámetros de las perlas pueden ser, o ser aproximadamente, 10 %, 20 %, 30 %, 40 %, 50 %, 60 %, 70 %, 80 %, 90 %, 100 %, 150 %, 200 %, 250 %, 300 %, o un número o un rango entre dos cualesquiera de estos valores, más largo o más corto que el diámetro de la célula.

[0187] Se puede unir una perla a un sustrato y/o incrustarla en él. Una perla puede unirse y/o incrustarse en un gel, hidrogel, polímero y/o matriz. La posición espacial de una perla dentro de un sustrato (por ejemplo, gel, matriz, armazón o polímero) se puede identificar usando la etiqueta espacial presente en el código de barras estocástico de la perla que puede servir como dirección de ubicación.

[0188] Los ejemplos de perlas pueden incluir, pero no se limitan a perlas de estreptavidina, perlas de agarosa, perlas magnéticas, Dynabeads®, microperlas MACS®, perlas conjugadas con anticuerpos (por ejemplo, microperlas antiinmunoglobulinas), perlas conjugadas con proteína A, perlas conjugadas con proteína G. perlas conjugadas, perlas conjugadas de proteína A/G, perlas conjugadas de proteína L, perlas conjugadas de oligo(dT), perlas de sílice, perlas similares a la sílice, microperlas anti-biotina, microperlas anti-fluorocromo y perlas magnéticas terminadas en carboxilo BcMag™.

[0189] Una perla puede asociarse con (por ejemplo, impregnarse con) puntos cuánticos o tintes fluorescentes para hacerla fluorescente en un canal óptico de fluorescencia o en múltiples canales ópticos. Una perla se puede asociar con óxido de hierro u óxido de cromo para volverla paramagnética o ferromagnética. Las cuentas pueden ser identificables. Por ejemplo, se puede fotografiar una cuenta con una cámara. Una cuenta puede tener un código detectable asociado con la cuenta. Por ejemplo, una cuenta puede comprender un código de barras estocástico. Una perla puede cambiar de tamaño, por ejemplo debido al hinchamiento en una solución orgánica o inorgánica. Una perla puede ser hidrófoba. Una perla puede ser hidrófila. Una perla puede ser biocompatible.

[0190] Se puede visualizar un soporte sólido (por ejemplo, una perla). El soporte sólido puede comprender una etiqueta de visualización (por ejemplo, tinte fluorescente). Un soporte sólido (por ejemplo, una cuenta) se puede grabar con un identificador (por ejemplo, un número). El identificador se puede visualizar a través de imágenes de las cuentas.

Sustratos y matriz de micropocillos

[0191] Como se usa en el presente documento, un sustrato puede referirse a un tipo de soporte sólido. Un sustrato puede referirse a un soporte sólido que puede comprender códigos de barras estocásticos de la divulgación. Un sustrato puede comprender, por ejemplo, varios micropocillos. Por ejemplo, un sustrato puede ser una matriz de pocillos que comprende dos o más micropocillos. En algunas formas de realización, un micropocillo puede comprender una pequeña cámara de reacción de volumen definido. En algunas formas de realización, un micropocillo puede atrapar una o más células. En algunas formas de realización, un micropocillo puede atrapar solo una célula. En algunas formas de realización, un micropocillo puede atrapar uno o más soportes sólidos. En algunas formas de realización, un micropocillo puede atrapar solo un soporte sólido. En algunas formas de realización, un micropocillo atrapa una única célula y un único soporte sólido (por ejemplo, una perla).

Métodos de códigos de barras estocásticos

[0192] La divulgación proporciona métodos para estimar el número de dianas distintas en distintas ubicaciones en una muestra física (por ejemplo, tejido, órgano, tumor, célula). Los métodos pueden comprender colocar los códigos de barras estocásticos muy cerca de la muestra, lisar la muestra, asociar distintos objetivos con los códigos de barras estocásticos, amplificar los objetivos y/o contar digitalmente los objetivos. El método puede comprender además analizar y/o visualizar la información obtenida de las etiquetas espaciales en los códigos de barras estocásticos. En algunas formas de realización, un método comprende visualizar la pluralidad de objetivos en la muestra. Mapear la pluralidad de objetivos en el mapa de la muestra puede incluir generar un mapa bidimensional o un mapa tridimensional de la muestra. El mapa bidimensional y el mapa tridimensional se pueden generar antes o después de codificar estocásticamente con barras la pluralidad de objetivos en la muestra. Visualizar la pluralidad de objetivos en la muestra puede incluir mapear la pluralidad de objetivos en un mapa de la muestra. Mapear la pluralidad de objetivos en el mapa de la muestra puede incluir generar un mapa bidimensional o un mapa tridimensional de la muestra. El mapa bidimensional y el mapa tridimensional se pueden generar antes o después de codificar estocásticamente con barras la pluralidad de objetivos en la muestra. En algunas formas de realización, el mapa bidimensional y se puede generar un mapa tridimensional antes o después de lisar la muestra. Lisar la muestra antes o después de generar el mapa bidimensional o el mapa tridimensional puede incluir calentar la muestra, poner en contacto la muestra con un detergente, cambiar el pH de la muestra o cualquier combinación de los mismos.

[0193] En algunas formas de realización, codificar de forma estocástica la pluralidad de objetivos comprende hibridar una pluralidad de códigos de barras estocásticos con una pluralidad de objetivos para crear objetivos con códigos de barras estocásticos. Codificar de forma estocástica la pluralidad de objetivos puede comprender la generación de una biblioteca indexada de los objetivos con código de barras estocástico. La generación de una biblioteca indexada de objetivos con códigos de barras estocásticos se puede realizar con un soporte sólido que comprende la pluralidad de códigos de barras estocásticos.

Contacto de una muestra y un código de barras estocástico

[0194] La divulgación proporciona métodos para poner en contacto una muestra (por ejemplo, células) con un sustrato de la divulgación. Una muestra que comprende, por ejemplo, una sección delgada de célula, órgano o tejido puede ponerse en contacto con códigos de barras estocásticos. Las células pueden ponerse en contacto, por ejemplo, mediante flujo por gravedad en el que las células pueden sedimentarse y crear una monocapa. La muestra puede ser una sección delgada de tejido. La sección delgada se puede colocar sobre el sustrato. La muestra puede ser unidimensional (por ejemplo, formar una superficie plana). La muestra (por ejemplo, células) se puede esparcir por el sustrato, por ejemplo, haciendo crecer/cultivar las células en el sustrato.

[0195] Cuando los códigos de barras estocásticos están muy cerca de los objetivos, los objetivos pueden hibridarse con el código de barras estocástico. Los códigos de barras estocásticos se pueden contactar en una proporción no agotable de modo que cada objetivo distinto pueda asociarse con un código de barras estocástico distinto de la divulgación. Para garantizar una asociación eficiente entre el objetivo y el código de barras estocástico, los objetivos se pueden vincular al código de barras estocástico.

Lisis celular

[0196] Tras la distribución de las células y los códigos de barras estocásticos, las células se pueden lisar para liberar las moléculas diana. La lisis celular se puede lograr mediante cualquiera de una variedad de medios, por ejemplo, mediante medios químicos o bioquímicos, mediante choque osmótico o mediante lisis térmica, lisis mecánica u lisis óptica. Las células se pueden lisar mediante la adición de un tampón de lisis celular que comprende un detergente (por ejemplo, SDS, Li dodecil sulfato, Triton X-100, Tween-20 o NP-40), un disolvente orgánico (por ejemplo, metanol o acetona) o enzimas digestivas (por ejemplo, proteinasa K, pepsina o tripsina), o cualquier combinación de los mismos. Para aumentar la asociación de una diana y un código de barras estocástico, la velocidad de difusión de las moléculas diana se puede alterar, por ejemplo, reduciendo la temperatura y/o aumentando la viscosidad del lisado.

[0197] En algunas formas de realización, la muestra se puede lisar usando un papel de filtro. El papel de filtro se puede empapar con un tampón de lisis encima del papel de filtro. El papel de filtro se puede aplicar a la muestra con presión, lo que puede facilitar la lisis de la muestra y la hibridación de los objetivos de la muestra con el sustrato.

[0198] En algunas formas de realización, la lisis se puede realizar mediante lisis mecánica, lisis por calor, lisis óptica y/o lisis química. La lisis química puede incluir el uso de enzimas digestivas como la proteinasa K, la pepsina y la tripsina. La lisis se puede realizar mediante la adición de un tampón de lisis al sustrato. Un tampón de lisis puede comprender Tris HCl. Un tampón de lisis puede comprender al menos aproximadamente 0,01, 0,05, 0,1, 0,5 o 1 M o más de Tris HCl. Un tampón de lisis puede comprender como máximo aproximadamente 0,01, 0,05, 0,1, 0,5 o 1 M o más Tris HCL. Un tampón de lisis puede comprender aproximadamente Tris HCl 0,1 M. El pH del tampón de lisis puede ser al menos aproximadamente 1, 2, 3, 4, 5, 6, 7, 8, 9 o 10 o más. El pH del tampón de lisis puede ser como máximo aproximadamente 1, 2, 3, 4, 5, 6, 7, 8, 9 o 10 o más. En algunas formas de realización, el pH del tampón de lisis es aproximadamente 7,5. El tampón de lisis puede comprender una sal (por ejemplo, LiCl). La concentración de sal en el tampón de lisis puede ser al menos aproximadamente 0,1, 0,5 o 1 M o más. La concentración de sal en el tampón de lisis puede ser como máximo aproximadamente 0,1, 0,5 o 1 M o más. En algunas formas de realización, la concentración de sal en el tampón de lisis es aproximadamente 0,5 M. El tampón de lisis puede comprender un detergente (por ejemplo, SDS, Li dodecil sulfato, triton X, tween, NP-40). La concentración del detergente en el tampón de lisis puede ser al menos aproximadamente 0,0001 %, 0,0005 %, 0,001 %, 0,005 %, 0,01 %, 0,05 %, 0,1 %, 0,5 %, 1 %, 2 %, 3 %, 4 %, 5 %, 6 % o 7 % o más. La concentración del detergente en el tampón de lisis puede ser como máximo aproximadamente 0,0001 %, 0,0005 %, 0,001 %, 0,005 %, 0,01 %, 0,05 %, 0,1 %, 0,5 %, 1 %, 2 %, 3 %, 4 %. 5 %, 6 % o 7 % o más. En algunas formas de realización, la concentración del detergente en el tampón de lisis es aproximadamente 1 % de Li dodecil sulfato. El tiempo utilizado en el método de lisis puede depender de la cantidad de detergente utilizado. En algunas formas de realización, cuanto más detergente se use, menos tiempo se necesitará para la lisis. El tampón de lisis puede comprender un agente quelante (por ejemplo, EDTA, EGTA). La concentración de un agente quelante en el tampón de lisis puede ser al menos aproximadamente 1, 5, 10, 15, 20, 25 o 30 mM o más. La concentración de un agente quelante en el tampón de lisis puede ser como máximo aproximadamente 1, 5, 10, 15, 20, 25 o 30 mM o más. En algunas formas de realización, la concentración de agente quelante en el tampón de lisis es aproximadamente 10 mM. El tampón de lisis puede comprender un reactivo reductor (por ejemplo, beta-mercaptoetanol, DTT). La concentración del reactivo reductor en el tampón de lisis puede ser al menos aproximadamente 1, 5, 10, 15 o 20 mM o más. La concentración del reactivo reductor en el tampón de lisis puede ser como máximo aproximadamente 1, 5, 10, 15 o 20 mM o más. En algunas formas de realización, la concentración de reactivo reductor en el tampón de lisis es aproximadamente 5 mM. En algunas formas de realización, un tampón de lisis puede comprender aproximadamente TrisHCl 0,1 M, aproximadamente pH 7,5, aproximadamente 0,5 M LiCl, aproximadamente 1 % de dodecilsulfato de litio, aproximadamente EDTA 10 mM y aproximadamente 5 mM de DTT.

[0199] La lisis se puede realizar a una temperatura de aproximadamente 4, 10, 15, 20, 25 o 30 ºC. La lisis se puede realizar durante aproximadamente 1, 5, 10, 15 o 20 o más minutos. Una célula lisada puede comprender al menos aproximadamente 100.000, 200.000, 300.000, 400.000, 500.000, 600.000 o 700.000 o más moléculas de ácido nucleico diana. Una célula lisada puede comprender como máximo aproximadamente 100.000, 200.000, 300.000, 400.000, 500.000, 600.000 o 700.000 o más moléculas de ácido nucleico diana.

Adjunción de códigos de barras estocásticos a moléculas de ácido nucleico objetivo

[0200] Después de la lisis de las células y la liberación de moléculas de ácido nucleico de las mismas, las moléculas de ácido nucleico pueden asociarse aleatoriamente con los códigos de barras estocásticos del soporte sólido colocalizado. La asociación puede comprender la hibridación de una región de reconocimiento objetivo de un código de barras estocástico con una porción complementaria de la molécula de ácido nucleico objetivo (por ejemplo, el oligo(dT) del código de barras estocástico puede interactuar con una cola poli(A) de un objetivo). Las condiciones de ensayo utilizadas para la hibridación (por ejemplo, pH del tampón, fuerza iónica, temperatura, etc.) pueden elegirse para promover la formación de híbridos estables específicos. En algunas formas de realización, las moléculas de ácido nucleico liberadas de las células lisadas pueden asociarse con la pluralidad de sondas en el sustrato (por ejemplo, hibridarse con las sondas en el sustrato). Cuando las sondas comprenden oligo(dT), las moléculas de ARNm pueden hibridarse con las sondas y transcribirse de forma inversa. La porción oligo(dT) del oligonucleótido puede actuar como cebador para la síntesis de la primera hebra de la molécula de ADNc. Por ejemplo, en un ejemplo no limitante de código de barras estocástico ilustrado en la FIG. 2, en el bloque 216, las moléculas de ARNm pueden hibridarse con códigos de barras estocásticos en cuentas. Por ejemplo, los fragmentos de nucleótidos monocatenarios pueden hibridarse con las regiones de unión al objetivo de los códigos de barras estocásticos.

[0201] La unión puede comprender además la ligación de una región de reconocimiento objetivo de un código de barras estocástico y una porción de la molécula de ácido nucleico objetivo. Por ejemplo, la región de unión diana puede comprender una secuencia de ácido nucleico que puede ser capaz de hibridación específica con un saliente del sitio de restricción (por ejemplo, un saliente del extremo adhesivo de EcoRI). El procedimiento de ensayo puede comprender además tratar los ácidos nucleicos diana con una enzima de restricción (por ejemplo, EcoRI) para crear un saliente del sitio de restricción. El código de barras estocástico puede entonces ligarse a cualquier molécula de ácido nucleico que comprenda una secuencia complementaria al saliente del sitio de restricción. Puede usarse una ligasa (por ejemplo, ADN ligasa T4) para unir los dos fragmentos.

[0202] Por ejemplo, en un ejemplo no limitante de código de barras estocástico ilustrado en la FIG. 2, en el bloque 220, las dianas marcadas de una pluralidad de células (o una pluralidad de muestras) (por ejemplo, moléculas de código de barras diana) se pueden agrupar posteriormente, por ejemplo, en un tubo. Las dianas marcadas se pueden agrupar, por ejemplo, recuperando los códigos de barras estocásticos y/o las perlas a las que están unidas las moléculas de código de barras diana.

[0203] La recuperación de colecciones basadas en soporte sólido de moléculas de código de barras objetivo adjuntas se puede implementar mediante el uso de perlas magnéticas y un campo magnético aplicado externamente. Una vez que se han agrupado las moléculas del código de barras objetivo, todo el procesamiento posterior puede realizarse en un único recipiente de reacción. El procesamiento adicional puede incluir, por ejemplo, reacciones de transcripción inversa, reacciones de amplificación, reacciones de escisión, reacciones de disociación y/o reacciones de extensión de ácidos nucleicos. Se pueden realizar reacciones de procesamiento adicionales dentro de los micropocillos, es decir, sin agrupar primero las moléculas de ácido nucleico diana marcadas de una pluralidad de células.

Transcripción inversa

[0204] La divulgación proporciona un método para crear un conjugado objetivo-código de barras estocástico usando transcripción inversa (por ejemplo, en el bloque 224 de la FIG. 2). El conjugado objetivo estocástico-código de barras puede comprender el código de barras estocástico y una secuencia complementaria de todo o una parte del ácido nucleico objetivo (es decir, una molécula de ADNc con código de barras estocástico). La transcripción inversa de la molécula de ARN asociada puede ocurrir mediante la adición de un cebador de transcripción inversa junto con la transcriptasa inversa. El cebador de transcripción inversa puede ser un cebador oligo(dT), un cebador hexanucleotídico aleatorio o un cebador oligonucleotídico específico de la diana. Los cebadores oligo(dT) pueden tener, o pueden tener, aproximadamente, 12-18 nucleótidos de longitud y unirse a la cola endógena poli(A) en el extremo 3' del ARNm de mamífero. Los cebadores de hexanucleótidos aleatorios pueden unirse al ARNm en una variedad de sitios complementarios. Los cebadores oligonucleotídicos específicos de la diana normalmente ceban selectivamente el ARNm de interés.

[0205] En algunas formas de realización, la transcripción inversa de la molécula de ARN marcada puede ocurrir mediante la adición de un cebador de transcripción inversa. En algunas formas de realización, el cebador de transcripción inversa es un cebador oligo(dT), un cebador de hexanucleótido aleatorio o un cebador de oligonucleótido específico de la diana. Generalmente, los cebadores oligo(dT) tienen entre 12 y 18 nucleótidos de longitud y se unen a la cola endógena poli(A)+ en el extremo 3' del ARNm de mamíferos. Los cebadores de hexanucleótidos aleatorios pueden unirse al ARNm en una variedad de sitios complementarios. Los cebadores oligonucleotídicos específicos de la diana normalmente ceban selectivamente el ARNm de interés.

[0206] La transcripción inversa puede ocurrir repetidamente para producir múltiples moléculas de ADNc marcadas. Los métodos descritos en el presente documento pueden comprender realizar al menos aproximadamente 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19 o 20 reacciones de transcripción inversa. El método puede comprender realizar al menos aproximadamente 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95 o 100 reacciones de transcripción inversa.

Amplificación

[0207] Una o más reacciones de amplificación de ácidos nucleicos (p. ej., en bloque 228 de la FIG. 2) para crear múltiples copias de las moléculas de ácidos nucleicos diana marcadas. La amplificación se puede realizar de forma multiplexada, en la que se amplifican simultáneamente múltiples secuencias de ácidos nucleicos diana. La reacción de amplificación se puede utilizar para agregar adaptadores de secuenciación a las moléculas de ácido nucleico. Las reacciones de amplificación pueden comprender amplificar al menos una porción de una etiqueta de muestra, si está presente. Las reacciones de amplificación pueden comprender amplificar al menos una porción del marcador celular y/o molecular. Las reacciones de amplificación pueden comprender amplificar al menos una porción de una etiqueta de muestra, una etiqueta celular, una etiqueta espacial, una etiqueta molecular, un ácido nucleico diana o una combinación de los mismos. Las reacciones de amplificación pueden comprender amplificar 0,5 %, 1 %, 2 %, 3 %, 4 %, 5 %, 6 %, 7 %, 8 %, 9 %, 10 %, 15 %, 20 %, 25 %, 30 %, 35 %, 40 %, 45 %, 50 %, 55 %, 60 %, 65 %, 70 %, 75 %, 80 %, 85 %, 90 %, 95 %, 97 %, 100 %, o un rango o un número entre dos cualesquiera de estos valores, de la pluralidad de ácidos nucleicos. El método puede comprender además realizar una o más reacciones de síntesis de ADNc para producir una o más copias de ADNc de moléculas de código de barras objetivo que comprenden un marcador de muestra, un marcador celular, un marcador espacial y/o una etiqueta molecular.

[0208] En algunas formas de realización, la amplificación se puede realizar usando una reacción en hebra de la polimerasa (PCR). Como se usa en el presente documento, PCR puede referirse a una reacción para la amplificación in vitro de secuencias de ADN específicas mediante la extensión simultánea del cebador de hebras complementarias de ADN. Como se usa en el presente documento, la PCR puede abarcar formas derivadas de la reacción, incluidas, entre otras, TI-PCR, PCR en tiempo real, PCR anidada, PCR cuantitativa, PCR digital multiplexada y PCR de ensamblaje.

[0209] La amplificación de los ácidos nucleicos marcados puede comprender métodos no basados en PCR. Ejemplos de métodos no basados en PCR incluyen, entre otros, amplificación por desplazamiento múltiple (MDA), amplificación mediada por transcripción (TMA), amplificación basada en secuencia de ácido nucleico (NASBA), amplificación por desplazamiento de hebra (SDA), amplificación en tiempo real SDA, amplificación de círculo rodante o amplificación de círculo a círculo. Otros métodos de amplificación no basados en PCR incluyen ciclos múltiples de amplificación de la transcripción de ARN impulsada por ARN polimerasa dependiente de ADN o síntesis y transcripción de ADN dirigida por ARN para amplificar objetivos de ADN o ARN, una reacción en hebra de la ligasa (LCR) y una replicasa Qβ (Qβ), uso de sondas palindrómicas, amplificación por desplazamiento de hebra, amplificación impulsada por oligonucleótidos usando una endonucleasa de restricción, un método de amplificación en el que un cebador se hibrida con una secuencia de ácido nucleico y el dúplex resultante se escinde antes de la reacción de extensión y amplificación, amplificación por desplazamiento de hebra utilizando una polimerasa de ácido nucleico que carece de actividad exonucleasa 5', amplificación por círculo rodante y amplificación por extensión de ramificación (RAM). En algunas formas de realización, la amplificación no produce transcritos circularizados.

[0210] En formas de realización, los métodos descritos en el presente documento comprenden además realizar una reacción en hebra de la polimerasa en el ácido nucleico marcado (por ejemplo, ARN marcado, ADN marcado, ADNc marcado) para producir un amplicón marcado estocásticamente. El amplicón marcado puede ser una molécula de doble hebra. La molécula de doble hebra puede comprender una molécula de ARN de doble hebra, una molécula de ADN de doble hebra o una molécula de ARN hibridada con una molécula de ADN. Una o ambas cadenas de la molécula bicatenaria pueden comprender un marcador de muestra, un marcador espacial, un marcador celular y/o una etiqueta molecular. El amplicón marcado estocásticamente puede ser una molécula monocatenaria. La molécula monocatenaria puede comprender ADN, ARN o una combinación de los mismos. Los ácidos nucleicos de la divulgación pueden comprender ácidos nucleicos sintéticos o alterados.

[0211] La amplificación puede comprender el uso de uno o más nucleótidos no naturales. Los nucleótidos no naturales pueden comprender nucleótidos fotolábiles o desencadenables. Los ejemplos de nucleótidos no naturales pueden incluir, entre otros, ácido peptídico nucleico (PNA), morfolino y ácido nucleico bloqueado (LNA), así como ácido nucleico de glicol (GNA) y ácido nucleico treosa (TNA). Se pueden añadir nucleótidos no naturales a uno o más ciclos de una reacción de amplificación. La adición de nucleótidos no naturales se puede utilizar para identificar productos como ciclos específicos o puntos de tiempo en la reacción de amplificación.

[0212] La forma de realización de una o más reacciones de amplificación puede comprender el uso de uno o más cebadores. Uno o más cebadores pueden comprender, por ejemplo, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14 o 15 o más nucleótidos. Uno o más cebadores pueden comprender al menos 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14 o 15 o más nucleótidos. Uno o más cebadores pueden comprender menos de 12-15 nucleótidos. Uno o más cebadores pueden hibridarse con al menos una porción de la pluralidad de objetivos marcados estocásticamente. Uno o más cebadores pueden hibridarse con el extremo 3' o el extremo 5' de la pluralidad de objetivos marcados estocásticamente. Uno o más cebadores pueden hibridarse con una región interna de la pluralidad de objetivos marcados estocásticamente. La región interna puede ser al menos aproximadamente 50, 100, 150, 200, 220, 230, 240, 250, 260, 270, 280, 290, 300, 310, 320, 330, 340, 350, 360, 370, 380. 390, 400, 410, 420, 430, 440, 450, 460, 470, 480, 490, 500, 510, 520, 530, 540, 550, 560, 570, 580, 590, 600, 650, 700, 0, 800, 850, 900 o 1000 nucleótidos de los extremos 3' de la pluralidad de objetivos marcados estocásticamente. Uno o más cebadores pueden comprender un panel fijo de cebadores. El uno o más cebadores pueden comprender al menos uno o más cebadores personalizados. El uno o más cebadores pueden comprender al menos uno o más cebadores de control. El uno o más cebadores pueden comprender al menos uno o más cebadores específicos de gen.

[0213] Uno o más cebadores pueden comprender un cebador universal. El cebador universal puede hibridarse con un sitio de unión del cebador universal. Uno o más cebadores personalizados pueden hibridarse con un primer marcador de muestra, un segundo marcador de muestra, un marcador espacial, un marcador celular, una etiqueta molecular, una diana o cualquier combinación de los mismos. El uno o más cebadores pueden comprender un cebador universal y un cebador personalizado. El cebador personalizado puede diseñarse para amplificar uno o más objetivos. Las dianas pueden comprender un subconjunto de los ácidos nucleicos totales en una o más muestras. Los objetivos pueden comprender un subconjunto del total de objetivos etiquetados estocásticamente en una o más muestras. Uno o más cebadores pueden comprender al menos 96 o más cebadores personalizados. Uno o más cebadores pueden comprender al menos 960 o más cebadores personalizados. Uno o más cebadores pueden comprender al menos 9600 o más cebadores personalizados. Uno o más cebadores personalizados pueden hibridarse con dos o más ácidos nucleicos marcados diferentes. Los dos o más ácidos nucleicos marcados diferentes pueden corresponder a uno o más genes.

[0214] Se puede utilizar cualquier esquema de amplificación en los métodos de la presente divulgación. Por ejemplo, en un esquema, la primera ronda de PCR puede amplificar moléculas unidas a la perla usando un cebador específico de gen y un cebador contra la secuencia del cebador 1 de secuenciación universal de Illumina. La segunda ronda de PCR puede amplificar los primeros productos de PCR utilizando un cebador específico de gen anidado flanqueado por la secuencia del cebador 2 de secuenciación de Illumina y un cebador contra la secuencia del cebador 1 de secuenciación universal de Illumina. La tercera ronda de PCR agrega P5 y P7 y un índice de muestra para convertir los productos de PCR en una biblioteca de secuenciación de Illumina. La secuenciación mediante secuenciación de 150 pb x 2 puede

revelar la etiqueta celular y la etiqueta molecular en la lectura 1, el gen en la lectura 2 y el índice de muestra en la lectura del índice 1.

[0215] En algunas formas de realización, los ácidos nucleicos se pueden eliminar del sustrato mediante escisión química. Por ejemplo, se puede utilizar un grupo químico o una base modificada presente en un ácido nucleico para facilitar su eliminación de un soporte sólido. Por ejemplo, se puede usar una enzima para eliminar un ácido nucleico de un sustrato. Por ejemplo, un ácido nucleico puede eliminarse de un sustrato mediante una digestión con endonucelasa de restricción. Por ejemplo, se puede usar el tratamiento de un ácido nucleico que contiene un dUTP o ddUTP con uracil-d-glicosilasa (UDG) para eliminar un ácido nucleico de un sustrato. Por ejemplo, se puede eliminar un ácido nucleico de un sustrato usando una enzima que realiza la escisión de nucleótidos, tal como una enzima reparadora de escisión de bases, tal como una endonucleasa apurínica/apirimidínica (AP). En algunas formas de realización, un ácido nucleico puede eliminarse de un sustrato usando un grupo fotoescindible y luz. En algunas formas de realización, se puede usar un conector escindible para eliminar un ácido nucleico del sustrato. Por ejemplo, el conector escindible puede comprender al menos uno de biotina/avidina, biotina/estreptavidina, biotina/neutravidina, proteína Ig A, un conector fotolábil, un grupo conector lábil a ácido o base, o un aptámero.

[0216] Cuando son sondas son específicas de genes, las moléculas pueden hibridarse con las sondas y transcribirse de forma inversa y/o amplificarse. En algunas formas de realización, después de que el ácido nucleico se haya sintetizado (por ejemplo, transcrito de forma inversa), se puede amplificar. La amplificación se puede realizar de forma múltiple, en la que se amplifican simultáneamente múltiples secuencias de ácidos nucleicos diana. La amplificación puede agregar adaptadores de secuenciación al ácido nucleico.

[0217] En algunas formas de realización, la amplificación se puede realizar en el sustrato, por ejemplo, con amplificación en puente. Los ADNc pueden tener una cola de homopolímero para generar un extremo compatible para la amplificación del puente utilizando sondas oligo(dT) en el sustrato. En la amplificación en puente, el cebador que es complementario al extremo 3' del ácido nucleico molde puede ser el primer cebador de cada par que está unido covalentemente a la partícula sólida. Cuando una muestra que contiene el ácido nucleico plantilla se pone en contacto con la partícula y se realiza un único ciclo térmico, la molécula plantilla se puede hibridar con el primer cebador y el primer cebador se alarga en dirección hacia adelante mediante la adición de nucleótidos para formar una molécula dúplex que consiste en la molécula plantilla y una hebra de ADN recién formada que es complementaria a la plantilla. En el paso de calentamiento del siguiente ciclo, la molécula dúplex se puede desnaturalizar, liberando la molécula plantilla de la partícula y dejando la hebra de ADN complementaria unida a la partícula a través del primer cebador. En la etapa de hibridación de la etapa de hibridación y elongación que sigue, la hebra complementaria puede hibridarse con el segundo cebador, que es complementario a un segmento de la hebra complementaria en una ubicación eliminada del primer cebador. Esta hibridación puede hacer que la hebra complementaria forme un puente entre el primer y el segundo cebador fijado al primer cebador mediante un enlace covalente y al segundo cebador mediante hibridación. En la etapa de elongación, el segundo cebador se puede alargar en dirección inversa mediante la adición de nucleótidos en la misma mezcla de reacción, convirtiendo así el puente en un puente bicatenario. A continuación, comienza el siguiente ciclo y el puente bicatenario puede desnaturalizarse para producir dos moléculas de ácido nucleico monocatenario, cada una con un extremo unido a la superficie de la partícula a través del primer y segundo cebador, respectivamente, con el otro extremo de cada uno sin unir. En la etapa de hibridación y alargamiento de este segundo ciclo, cada hebra puede hibridarse con un cebador complementario adicional, no utilizado previamente, en la misma partícula, para formar nuevos puentes monocatenarios. Los dos cebadores no utilizados anteriormente que ahora se hibridan se alargan para convertir los dos nuevos puentes en puentes de doble hebra.

[0218] Las reacciones de amplificación pueden comprender amplificar al menos 1 %, 2 %, 3 %, 4 %, 5 %, 6 %, 7 %, 8 %, 9 %, 10 %, 15 %, 20 %, 25 %, 30 %, 35 %, 40 %, 45 %, 50 %, 55 %, 60 %, 65 %, 70 %, 75 %, 80 %, 85 %, 90 %, 95 %, 97 % o 100 % de la pluralidad de ácidos nucleicos.

[0219] La amplificación de los ácidos nucleicos marcados puede comprender métodos basados en PCR o métodos no basados en PCR. La amplificación de los ácidos nucleicos marcados puede comprender una amplificación exponencial de los ácidos nucleicos marcados. La amplificación de los ácidos nucleicos marcados puede comprender una amplificación lineal de los ácidos nucleicos marcados. La amplificación se puede realizar mediante reacción en hebra de la polimerasa (PCR). La PCR puede referirse a una reacción para la amplificación in vitro de secuencias de ADN específicas mediante la extensión simultánea del cebador de hebras complementarias de ADN. La PCR puede abarcar formas derivadas de la reacción, incluidas, entre otras, TI-PCR, PCR en tiempo real, PCR anidada, PCR cuantitativa, PCR digital multiplexada, PCR de supresión, PCR semisupresora y PCR de ensamblaje.

[0220] En algunas formas de realización, la amplificación de los ácidos nucleicos marcados comprende métodos no basados en PCR. Ejemplos de métodos no basados en PCR incluyen, entre otros, amplificación por desplazamiento múltiple (MDA), amplificación mediada por transcripción (TMA), amplificación basada en secuencia de ácido nucleico (NASBA), amplificación por desplazamiento de hebra (SDA), SDA en tiempo real, amplificación de círculo rodante o amplificación de círculo a círculo. Otros métodos de amplificación no basados en PCR incluyen múltiples ciclos de amplificación de la transcripción de ARN impulsada por ARN polimerasa dependiente de ADN o síntesis y transcripción de ADN dirigida por ARN para amplificar objetivos de ADN o ARN, una reacción en hebra de la ligasa (LCR), una replicasa Qβ (Qβ), uso de sondas palindrómicas, amplificación por desplazamiento de hebra, amplificación impulsada por oligonucleótidos utilizando una endonucleasa de restricción, un método de amplificación en el que un cebador se hibrida

con una secuencia de ácido nucleico y el dúplex resultante se escinde antes de la reacción de extensión y amplificación, amplificación por desplazamiento de hebra utilizando una polimerasa de ácido nucleico que carece de actividad exonucleasa 5', amplificación por círculo rodante y/o amplificación por extensión de ramificación (RAM).

[0221] En algunas formas de realización, los métodos descritos en el presente documento comprenden además realizar una reacción en hebra de la polimerasa anidada en el amplicón amplificado (por ejemplo, diana). El amplicón puede ser una molécula de doble hebra. La molécula de doble hebra puede comprender una molécula de ARN de doble hebra, una molécula de ADN de doble hebra o una molécula de ARN hibridada con una molécula de ADN. Una o ambas hebras de la molécula bicatenaria pueden comprender una etiqueta de muestra o una etiqueta de identificación molecular. Alternativamente, el amplicón puede ser una molécula monocatenaria. La molécula monocatenaria puede comprender ADN, ARN o una combinación de los mismos. Los ácidos nucleicos de la presente invención pueden comprender ácidos nucleicos sintéticos o alterados.

[0222] En algunas formas de realización, el método comprende amplificar repetidamente el ácido nucleico marcado para producir múltiples amplicones. Los métodos descritos en el presente documento pueden comprender realizar al menos aproximadamente 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19 o 20 reacciones de amplificación. Alternativamente, el método comprende realizar al menos aproximadamente 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95 o 100 reacciones de amplificación.

[0223] La amplificación puede comprender además agregar uno o más ácidos nucleicos de control a una o más muestras que comprenden una pluralidad de ácidos nucleicos. La amplificación puede comprender además agregar uno o más ácidos nucleicos de control a una pluralidad de ácidos nucleicos. Los ácidos nucleicos de control pueden comprender una etiqueta de control.

[0224] La amplificación puede comprender el uso de uno o más nucleótidos no naturales. Los nucleótidos no naturales pueden comprender nucleótidos fotolábiles y/o activables. Los ejemplos de nucleótidos no naturales incluyen, entre otros, ácido peptídico nucleico (PNA), morfolino y ácido nucleico bloqueado (LNA), así como ácido nucleico de glicol (GNA) y ácido treosa nucleico (TNA). Se pueden añadir nucleótidos no naturales a uno o más ciclos de una reacción de amplificación. La adición de nucleótidos no naturales se puede utilizar para identificar productos como ciclos específicos o puntos de tiempo en la reacción de amplificación.

[0225] La forma de realización de una o más reacciones de amplificación puede comprender el uso de uno o más cebadores. El uno o más cebadores pueden comprender uno o más oligonucleótidos. El uno o más oligonucleótidos pueden comprender al menos aproximadamente 7-9 nucleótidos. El uno o más oligonucleótidos pueden comprender menos de 12-15 nucleótidos. Uno o más cebadores pueden hibridarse con al menos una porción de la pluralidad de ácidos nucleicos marcados. Uno o más cebadores pueden hibridarse con el extremo 3' y/o el extremo 5' de la pluralidad de ácidos nucleicos marcados. Uno o más cebadores pueden hibridarse con una región interna de la pluralidad de ácidos nucleicos marcados. La región interna puede ser al menos aproximadamente 50, 100, 150, 200, 220, 230, 240, 250, 260, 270, 280, 290, 300, 310, 320, 330, 340, 350, 360, 370, 380, 390, 400, 410, 420, 430, 440, 450, 460, 470, 480, 490, 500, 510, 520, 530, 540, 550, 560, 570, 580, 590, 600, 650, 700, 0, 800, 850, 900 o 1000 nucleótidos de los extremos 3' de la pluralidad de ácidos nucleicos marcados. Uno o más cebadores pueden comprender un panel fijo de cebadores. El uno o más cebadores pueden comprender al menos uno o más cebadores personalizados. El uno o más cebadores pueden comprender al menos uno o más cebadores de control. El uno o más cebadores pueden comprender al menos uno o más cebadores de genes constitutivos. El uno o más cebadores pueden comprender un cebador universal. El cebador universal puede hibridarse con un sitio de unión del cebador universal. Uno o más cebadores personalizados pueden hibridarse con la primera etiqueta de muestra, la segunda etiqueta de muestra, la etiqueta de identificación molecular, el ácido nucleico o un producto del mismo. Uno o más cebadores pueden comprender un cebador universal y un cebador personalizado. El cebador personalizado puede diseñarse para amplificar uno o más ácidos nucleicos diana. Los ácidos nucleicos diana pueden comprender un subconjunto de los ácidos nucleicos totales en una o más muestras. En algunas formas de realización, los cebadores son las sondas unidas a la matriz de la divulgación.

[0226] En algunas formas de realización, codificar de forma estocástica la pluralidad de objetivos en la muestra comprende además generar una biblioteca indexada de los fragmentos con código de barras estocástica. Las etiquetas moleculares de diferentes códigos de barras estocásticos pueden ser diferentes entre sí. Generar una biblioteca indexada de objetivos con códigos de barras estocásticos incluye generar una pluralidad de polinucleótidos indexados a partir de la pluralidad de objetivos en la muestra. Por ejemplo, para una biblioteca indexada de dianas con códigos de barras estocásticos que comprende una primera diana indexada y una segunda diana indexada, la región marcadora del primer polinucleótido indexado puede diferir de la región marcadora del segundo polinucleótido indexado en, aproximadamente, en al menos, o por como máximo, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50, o un número o un rango entre dos cualesquiera de estos valores, nucleótidos. En algunas formas de realización, generar una biblioteca indexada de dianas con códigos de barras estocásticos incluye poner en contacto una pluralidad de dianas, por ejemplo moléculas de ARNm, con una pluralidad de oligonucleótidos que incluyen una región poli(T) y una región marcadora; y realizar una síntesis de la primera hebra usando una transcriptasa inversa para producir moléculas de ADNc marcadas de una sola hebra, comprendiendo cada una una región de ADNc y una región marcadora, en donde la pluralidad de dianas incluye al menos dos moléculas de ARNm de secuencias diferentes y la pluralidad de oligonucleótidos incluye al menos dos oligonucleótidos de secuencias diferentes. Generar una biblioteca indexada de objetivos con códigos de barras estocásticos puede

comprender además amplificar las moléculas de ADNc marcadas de una sola hebra para producir moléculas de ADNc marcadas de doble hebra; y realizar una PCR anidada en las moléculas de ADNc marcadas de doble hebra para producir amplicones marcados. En algunas formas de realización, el método puede incluir generar un amplicón marcado con adaptador.

[0227] Los códigos de barras estocásticos utilizan códigos de barras o etiquetas de ácidos nucleicos para etiquetar ácidos nucleicos individuales (por ejemplo, ADN o ARN). En algunas formas de realización, implica agregar códigos de barras o etiquetas de ADN a moléculas de ADNc a medida que se generan a partir de ARNm. Se puede realizar una PCR anidada para minimizar el sesgo de amplificación de la PCR. Se pueden agregar adaptadores para secuenciar usando, por ejemplo, secuenciación de próxima generación (NGS). Los resultados de la secuenciación se pueden usar para determinar etiquetas celulares, etiquetas moleculares y secuencias de fragmentos de nucleótidos de una o más copias de las dianas, por ejemplo en el bloque 232 de la FIG. 2.

[0228] FIG. 3 es una ilustración esquemática que muestra un proceso ejemplar no limitante de generar una biblioteca indexada de objetivos con códigos de barras estocásticos, por ejemplo, ARNm. Como se muestra en el paso 1, el proceso de transcripción inversa puede codificar cada molécula de ARNm con una etiqueta molecular única, una etiqueta celular y un sitio de PCR universal. En particular, las moléculas de ARN 302 se pueden transcribir de forma inversa para producir moléculas de ADNc marcadas 304, incluida una región de ADNc 306, mediante la hibridación estocástica de un conjunto de etiquetas de identificación molecular 310 con la región de cola poli(A) 308 de las moléculas de ARN 302. Cada una de las etiquetas de identificación molecular 310 puede comprender una región de unión a diana, por ejemplo una región poli(dT) 312, una región de etiqueta 314 y una región de PCR universal 316.

[0229] En algunas formas de realización, el marcador celular puede incluir de 3 a 20 nucleótidos. En algunas formas de realización, el marcador molecular puede incluir de 3 a 20 nucleótidos. En algunas formas de realización, cada uno de la pluralidad de códigos de barras estocásticos comprende además uno o más de una etiqueta universal y una etiqueta de célula, en donde las etiquetas universales son las mismas para la pluralidad de códigos de barras estocásticos en el soporte sólido y las etiquetas de célula son las mismas para la pluralidad de códigos de barras estocásticos sobre el soporte sólido. En algunas formas de realización, el marcador universal puede incluir de 3 a 20 nucleótidos. En algunas formas de realización, el marcador celular comprende de 3 a 20 nucleótidos.

[0230] En algunas formas de realización, la región marcadora 314 puede incluir una etiqueta molecular 318 y una etiqueta celular 320. En algunas formas de realización, la región marcadora 314 puede incluir uno o más de un marcador universal, un marcador dimensional y un marcador celular. El marcador molecular 318 puede ser, puede ser aproximadamente, puede ser al menos o puede ser como máximo, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, o un número o rango entre cualquiera de estos valores, de nucleótidos de longitud. La etiqueta de célula 320 puede ser, puede ser aproximadamente, puede ser al menos o puede ser como máximo, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, o un número o rango entre cualquiera de estos valores, de nucleótidos de longitud. La etiqueta universal puede ser, puede ser aproximadamente, puede ser al menos o puede ser como máximo 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, o un número o un rango entre cualquiera de estos valores, de nucleótidos de longitud. Las etiquetas universales pueden ser las mismas para la pluralidad de códigos de barras estocásticos sobre el soporte sólido y las etiquetas de célula son las mismas para la pluralidad de códigos de barras estocásticos sobre el soporte sólido. La etiqueta de dimensión puede ser, puede ser aproximadamente, puede ser al menos o puede ser como máximo 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, o un número o rango entre cualquiera de estos valores, de nucleótidos de longitud.

[0231] En algunas formas de realización, la región de etiqueta 314 puede comprender, comprender aproximadamente, comprender al menos, o comprender como máximo, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, o un número o rango entre cualquiera de estos valores, diferentes etiquetas, como por ejemplo etiqueta molecular 318 y una etiqueta celular 320. Cada etiqueta puede ser, puede ser aproximadamente, puede ser al menos o puede ser como máximo 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, o un número o un rango entre cualquiera de estos valores, de nucleótidos de longitud. Un conjunto de etiquetas de identificación molecular 310 puede contener, contener aproximadamente, contener al menos o puede ser como máximo 10, 20, 40, 50, 70, 80, 90, 10^2, 10^3, 10^4, 10^5, 10^6, 10^7, 10^8, 10^9, 10^{10}, 10^{11}, 10^{12}, 10^{13}, 10^{14}, 10^{15}, 10^{20}, o un número o un rango entre cualquiera de estos valores, etiquetas de identificador molecular 310. Y el conjunto de etiquetas de identificador molecular 310 puede, por ejemplo, contener cada una una región marcadora única 314. Las moléculas de ADNc marcadas 304 se pueden purificar para eliminar el exceso de etiquetas identificadoras moleculares 310. La purificación puede comprender la purificación con perlas Ampure.

[0232] Como se muestra en el paso 2, los productos del proceso de transcripción inversa en el paso 1 se pueden agrupar en 1 tubo y amplificar por PCR con un primer conjunto de cebadores de PCR y un primer cebador de PCR universal. La combinación es posible debido a la región marcadora única 314. En particular, las moléculas de ADNc marcadas 304 pueden amplificarse para producir amplicones anidados marcados por PCR 322. La amplificación puede comprender una amplificación por PCR múltiple. La amplificación puede comprender una amplificación por PCR múltiple con 96 cebadores múltiples en un único volumen de reacción. En algunas formas de realización, la amplificación por PCR múltiple puede utilizar, utilizar aproximadamente, utilizar al menos o utilizar como máximo 10, 20, 40, 50, 70, 80, 90, 10^2, 10^3, 10^4, 10^5, 10^6, 10^7, 10^8, 10^9, 10^{10}, 10^{11}, 10^{12}, 10^{13}, 10^{14}, 10^{15}, 10^{20}, o un número o un rango entre cualquiera de estos valores,

multiplexan los cebadores en un único volumen de reacción. La amplificación puede comprender el primer conjunto de cebadores de PCR 324 de cebadores personalizados 326A-C dirigidos a genes específicos y un cebador universal 328. Los cebadores personalizados 326 pueden hibridarse con una región dentro de la porción de ADNc 306' de la molécula de ADNc marcada 304. El cebador universal 328 puede hibridar con la región de PCR universal 316 de la molécula de ADNc marcada 304.

[0233] Como se muestra en el paso 3 de la FIG. 3, los productos de la amplificación por PCR en el paso 2 se pueden amplificar con un conjunto de cebadores de PCR anidados y un segundo cebador de PCR universal. La PCR anidada puede minimizar el sesgo de amplificación de la PCR. En particular, los amplicones anidados 322 marcados con PCR se pueden amplificar adicionalmente mediante PCR anidada. La PCR anidada puede comprender una PCR múltiple con un conjunto de cebadores de PCR anidados 330 de cebadores de PCR anidados 332a-c y un segundo cebador de PCR universal 328' en un único volumen de reacción. El conjunto de cebadores de PCR anidados 328 puede contener, contener aproximadamente, contener al menos o contener como máximo 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, o un número o rango entre cualquiera de estos valores, diferentes cebadores de PCR anidados 330. Los cebadores de PCR anidados 332 puede contener un adaptador 334 e hibridarse con una región dentro de la porción de ADNc 306" del amplicón marcado 322. El cebador universal 328' puede contener un adaptador 336 e hibridarse con la región de PCR universal 316 del amplicón marcado 322. Por lo tanto, el paso 3 produce el amplicón 338 marcado con adaptador. En algunas formas de realización, los cebadores de PCR anidados 332 y el segundo cebador de PCR universal 328' pueden no contener los adaptadores 334 y 336. En cambio, los adaptadores 334 y 336 pueden ligarse a los productos de PCR anidados para producir amplicón marcado con adaptador 338.

[0234] Como se muestra en el paso 4, los productos de PCR del paso 3 se pueden amplificar por PCR para secuenciación usando cebadores de amplificación de biblioteca. En particular, los adaptadores 334 y 336 pueden usarse para realizar uno o más ensayos adicionales en el amplicón marcado con el adaptador 338. Los adaptadores 334 y 336 pueden hibridarse con los cebadores 340 y 342. El uno o más cebadores 340 y 342 pueden ser cebadores de amplificación por PCR. Uno o más cebadores 340 y 342 pueden ser cebadores de secuenciación. Uno o más adaptadores 334 y 336 pueden usarse para amplificación adicional de los amplicones etiquetados con adaptador 338. Uno o más adaptadores 334 y 336 pueden usarse para secuenciar el amplicón etiquetado con adaptador 338. El cebador 342 puede contener un índice de placa 344 de modo que los amplicones generados usando el mismo conjunto de etiquetas de identificación molecular 318 puedan secuenciarse en una reacción de secuenciación usando secuenciación de próxima generación (NGS).

Corrección de errores de secuenciación y PCR

[0235] En el presente documento se describen métodos para determinar el número de objetivos. En algunas formas de realización, el método comprende: (a) codificar de barras estocásticamente una pluralidad de objetivos usando una pluralidad de códigos de barras estocásticos para crear una pluralidad de objetivos con códigos de barras estocásticos, en donde cada uno de la pluralidad de códigos de barras estocásticos comprende una etiqueta molecular; (b) obtener datos de secuenciación de los objetivos con códigos de barras estocásticos; y (c) para una o más de la pluralidad de dianas: (i) contar el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación; (ii) determinar un estado de calidad de la diana en los datos de secuenciación; (iii) determinar uno o más errores de datos de secuenciación en los datos de secuenciación, en donde determinar uno o más errores de datos de secuenciación en los datos de secuenciación comprende determinar uno o más de: el número de etiquetas moleculares con secuencias distintas asociadas con la diana en el datos de secuenciación, el estado de calidad de la diana en los datos de secuenciación y el número de etiquetas moleculares con secuencias distintas en la pluralidad de códigos de barras estocásticos; y (iv) estimar el número de la diana, en donde el número de la diana estimado se correlaciona con el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación contados ajustados de acuerdo con uno o más errores de datos de secuenciación. Los pasos (i), (ii), (iii) y (iv) se pueden realizar para cada uno de la pluralidad de objetivos. El método puede ser multiplexado.

[0236] En algunas formas de realización, los métodos comprenden además: colapsar los datos de secuenciación antes de determinar uno o más errores de datos de secuenciación. Colapsar los datos de secuenciación comprende: atribuir copias de la diana con etiquetas moleculares similares y con ocurrencias menores que un umbral de ocurrencia de colapso predeterminado como si tuvieran la misma etiqueta molecular para la pluralidad de dianas, en donde dos copias de una diana tienen etiquetas moleculares similares si las etiquetas diana de las dos copias de la diana difieren en al menos una base en la secuencia.

[0237] El porcentaje de etiquetas moleculares en los datos de secuenciación retenidos después de que los datos de secuenciación se ajustan de acuerdo con uno o más errores de datos de secuenciación puede variar. En algunas formas de realización, el porcentaje de las etiquetas moleculares en los datos de secuenciación retenidos después de que los datos de secuenciación se ajustan de acuerdo con uno o más errores de datos de secuenciación puede ser, o ser aproximadamente, 50 %, 60 %, 70 %, 80 %, 90 %, 95 %, 99 %, 99,9 %, o un número o rango entre dos de estos valores. En algunas formas de realización, el porcentaje de las etiquetas moleculares en los datos de secuenciación retenidos después de que los datos de secuenciación se ajustan de acuerdo con uno o más errores de datos de secuenciación puede ser al menos, o como máximo, 50 %, 60 %, 70 %, 80 %, 90 %, 95 %, 99 % o 99,9 %.

Determinar los recuentos de etiquetas moleculares

[0238] FIG. 5 es un diagrama de flujo que muestra una forma de realización ejemplar 500 no limitante de corrección de errores de secuenciación y PCR usando etiquetas moleculares. La forma de realización 500 comienza en el bloque inicial 504 después de codificar de forma estocástica una pluralidad de objetivos usando una pluralidad de códigos de barras estocásticos para crear una pluralidad de objetivos con códigos de barras estocásticos, en donde cada uno de la pluralidad de códigos de barras estocásticos comprende una etiqueta molecular, y después de obtener datos de secuenciación de la objetivos con códigos de barras estocásticos.

[0239] Para una diana, por ejemplo un gen que se origina a partir de una célula en un micropocillo de una matriz de micropocillos, el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación se puede contar en el bloque 508. En los datos de secuenciación, dos copias de la diana pueden tener etiquetas moleculares similares; por ejemplo, las etiquetas moleculares de las dos copias de la diana pueden diferir en una base en la secuencia. Las dos copias del objetivo pueden ser ambas verdaderas, una copia del objetivo puede ser verdadera y la otra copia del objetivo puede ser el resultado de un error de secuenciación o un error de PCR, o ambas copias del objetivo pueden ser resultados de errores de secuenciación o errores de PCR.

Colapso de datos de secuenciación

[0240] En el bloque 512, los datos de secuenciación se pueden contraer. Colapsar los datos de secuenciación puede comprender atribuir copias de la diana con etiquetas moleculares similares y con ocurrencias menores que un umbral de ocurrencia de colapso predeterminado como si tuvieran la misma etiqueta molecular para la pluralidad de dianas. El umbral de ocurrencia de colapso predeterminado puede variar, oscilando de 1 a 100. En algunas formas de realización, el umbral de ocurrencia de colapso predeterminado puede ser, o ser aproximadamente, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10. 20, 30, 40, 50, 60, 70, 80, 90, 100, o un número o rango entre dos de estos valores si los códigos de barras estocásticos comprenden aproximadamente 6561 etiquetas moleculares con secuencias distintas. En algunas formas de realización, el umbral de ocurrencia de colapso predeterminado puede ser al menos, o como máximo, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50, 60, 70. 80, 90 o 100, si los códigos de barras estocásticos comprenden alrededor de 6561 etiquetas moleculares con secuencias distintas. Por ejemplo, las etiquetas moleculares pueden tener 8 nucleótidos de longitud y cada posición de nucleótido puede tener tres posibilidades, como adenina (A), citosina (C), guanina (G); C, G, timina (T); A, G, T; o A, C, T, dando lugar a 3^8 = 6561 etiquetas moleculares únicas.

[0241] En algunas formas de realización, el umbral de ocurrencia de colapso predeterminado puede ser, o ser aproximadamente, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 17, 20, 30, 40, 50, 60., 70, 80, 90, 100, o un número o un rango entre dos de estos valores cualesquiera si los códigos de barras estocásticos comprenden aproximadamente 65536 etiquetas moleculares con secuencias distintas. En algunas formas de realización, el umbral de ocurrencia de colapso predeterminado puede ser al menos, o como máximo, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 17, 20, 30, 40, 50, 60, 70, 80, 90 o 100, si los códigos de barras estocásticos comprenden aproximadamente 65536 etiquetas moleculares con secuencias distintas. Por ejemplo, las etiquetas moleculares pueden tener 8 nucleótidos de longitud y cada posición de nucleótido puede tener cuatro posibilidades A, C, G, T, lo que da como resultado 3^4 = 65536 etiquetas moleculares únicos.

[0242] Por ejemplo, puede haber cinco copias del objetivo. Las cinco copias de la diana pueden tener etiquetas moleculares de TGTGCGTG, TGTGCGCG, TGTGCGGG, GGTGCGTG y TGGGCGTG con un número de lecturas por etiqueta molecular de 261, 2, 2, 1 y 1 respectivamente. Las etiquetas moleculares TGTGCGCG, TGTGCGGG, GGTGCGTG y TGGGCGTG son similares al marcador molecular TGTGCGTG porque difieren del marcador molecular TGTGCGTG en un nucleótido (subrayado). Si hay 6561 etiquetas moleculares con secuencias distintas, y el umbral de aparición de colapso predeterminado es 7, entonces las apariciones de las etiquetas moleculares TTGCGCG, TTGCGGG, GGGTGCGTG y TGGCGTG se pueden atribuir a la etiqueta molecular TGGCGTG.

[0243] Como otro ejemplo, puede haber siete copias del objetivo. Las siete copias de la diana pueden tener etiquetas moleculares de CGTGTCTG, GGGGGCGA, GCTGCTGG, TCGGGCGA, CGCGTTCA, CGCGTTTA y TGGGCTTG con un número de lecturas por etiqueta molecular de 10, 7, 5, 4, 1, 1 y 1 respectivamente. La etiqueta molecular CGCGTTTA es similar a la etiqueta molecular CGCGTTCA porque se diferencian entre sí en un nucleótido (subrayado). Si hay 6561 etiquetas moleculares con secuencias distintas y el umbral de aparición de colapso predeterminado es 7, entonces la aparición de la etiqueta molecular CGCGTTTA se puede atribuir a la etiqueta molecular CGCGTTCA.

Errores de datos de secuenciación

[0244] Los métodos descritos en el presente documento se pueden usar para identificar y/o corregir errores de datos de secuenciación, por ejemplo, los errores que ocurren en los métodos para contar uno o más ácidos nucleicos diana. En algunas formas de realización, un error de datos de secuenciación puede comprender, o ser, un error introducido por PCR, un error introducido por secuenciación, un error causado por contaminación de código de barras, un error de preparación de biblioteca o cualquier combinación de los mismos. El error introducido por la PCR puede comprender, o ser el resultado de un error de amplificación por PCR, un sesgo de amplificación por PCR, una amplificación por PCR insuficiente o cualquier combinación de los mismos. El error introducido por la secuenciación puede comprender, o ser, el resultado de una llamada de base inexacta, una secuenciación insuficiente o cualquier combinación de los mismos. El

error puede comprender, o ser, una eliminación de uno o más nucleótidos, una sustitución de uno o más nucleótidos, una adición de uno o más nucleótidos, o cualquier combinación de los mismos.

Determinar el estado de secuenciación

[0245] Como se describió anteriormente, una pluralidad de objetivos puede tener códigos de barras estocásticos usando una pluralidad de códigos de barras estocásticos para crear una pluralidad de objetivos con códigos de barras estocásticos, y cada uno de la pluralidad de códigos de barras estocásticos puede comprender una etiqueta molecular y obtener datos de secuenciación de los códigos de barras estocásticos. objetivos con códigos de barras. Para una diana, por ejemplo un gen que se origina a partir de una célula en un micropocillo de una matriz de micropocillos, se puede contar el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación. Los datos de secuenciación contados se pueden contraer, por ejemplo, atribuyendo copias de la diana con etiquetas moleculares similares y con apariciones menores que un umbral de aparición de colapso predeterminado como si tuvieran la misma etiqueta molecular para la pluralidad de dianas. Después de colapsar los datos de secuenciación, se puede determinar el estado de calidad del objetivo.

[0246] Con referencia a la FIG. 5, en algunas formas de realización, en el bloque 516, se puede determinar que un estado de calidad del objetivo en los datos de secuenciación es secuenciación completa, secuenciación incompleta o secuenciación saturada. El estado de calidad del objetivo puede depender de si se han observado todas las etiquetas moleculares verdaderas o reales en la profundidad de la secuenciación. Las etiquetas moleculares verdaderas o reales pueden referirse a etiquetas moleculares que no son etiquetas moleculares erróneas o falsas. Error o etiquetas moleculares falsas pueden referirse a etiquetas moleculares que tienen secuencias resultantes de errores de PCR, artefactos o errores de secuenciación. El estado de calidad de la diana en los datos de secuenciación se puede determinar mediante el número de etiquetas moleculares con secuencias distintas en la pluralidad de códigos de barras estocásticos y la cantidad de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación contados.

[0247] En algunas formas de realización, el estado de calidad de la secuenciación completa se puede determinar mediante un índice de dispersión relativo a la distribución de Poisson mayor o igual a un umbral de dispersión de secuenciación completa predeterminado. El índice de dispersión se puede definir como varianza/media para un objetivo. FIG. 6 es una ilustración esquemática que muestra datos de secuencia obtenidos mediante secuenciación completa y secuenciación incompleta. FIG. 6 muestra tres copias del gen A y seis copias del gen B en la biblioteca (círculo izquierdo). Si las tres copias del gen A tuvieron lecturas de secuenciación de seis veces, cinco veces y una vez en los datos de secuenciación (círculo superior derecho), la varianza es 7, la media es 4 y el índice de dispersión es 1,75. Si las seis copias del gen B tuvieron lecturas de secuenciación de nueve veces, dos veces, dos veces, dos veces, una y una vez en los datos de secuenciación (círculo superior derecho), la varianza es 9,36, la media es 2,83 y el índice de dispersión es 3,31. Con estos datos de secuenciación, se puede considerar que el Gen A y el Gen B tienen el estado de secuenciación completa si el umbral de dispersión de secuenciación completa predeterminado es, por ejemplo, 0,9 para secuenciación completa.

[0248] Si no se observó una copia del Gen A y las otras dos copias del Gen A tuvieron lecturas de secuenciación de dos y tres veces en los datos de secuenciación (círculo inferior derecho), la varianza es 0,5, la media es 2,5 y el índice de dispersión es 0,2. Si no se observaron dos copias del gen B y las otras cuatro copias del gen B tuvieron lecturas de secuenciación cuatro veces, dos veces, una y una vez en los datos de secuenciación (círculo inferior derecho), la varianza es 2, la media es 2, y el índice de dispersión es 2. Con estos datos de secuenciación, se puede considerar que el Gen A y el Gen B tienen el estado de secuenciación incompleta si el umbral de dispersión de secuenciación completa predeterminado es, por ejemplo, 1,1 para secuenciación completa.

[0249] El umbral de dispersión de secuenciación completa predeterminado puede variar, oscilando entre 0,5 y 5. En algunas formas de realización, el umbral de dispersión de secuenciación completa predeterminado puede ser, o ser aproximadamente, 0,5, 0,6, 0,7, 0,8, 0,9, 1, 2, 3, 4, 5, 6, o un número o rango entre dos de estos valores. En algunas formas de realización, el umbral de dispersión de secuenciación completa predeterminado puede ser al menos o como máximo 0,5, 0,6, 0,7, 0,8, 0,9, 1, 2, 3, 4, 5 o 6.

[0250] En algunas formas de realización, el estado de calidad de la secuenciación completa puede determinarse además mediante una etiqueta molecular con una aparición mayor o igual a un umbral de aparición de secuenciación completa predeterminado en los datos de secuenciación. El umbral de aparición de secuenciación completa predeterminado puede variar, oscilando entre 8 y 20. En algunas formas de realización, el umbral de aparición de secuenciación completa puede ser, o ser aproximadamente, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17., 18, 19, 20, o un número o rango entre dos de estos valores. En algunas formas de realización, el umbral de aparición de secuenciación completa puede ser al menos, o como máximo, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19 o 20.

[0251] En algunas formas de realización, el estado de calidad de la secuenciación saturada se puede determinar si la diana tiene un número de etiquetas moleculares con secuencias distintas que han sido mayores que un umbral de saturación predeterminado. El estado de calidad de la secuenciación saturada puede determinarse además mediante otra

diana de la pluralidad de dianas que tiene una serie de etiquetas moleculares con secuencias distintas que sean mayores que el umbral de saturación predeterminado.

[0252] El umbral de saturación predeterminado puede variar. En algunas formas de realización, el umbral de saturación predeterminado puede ser, o ser aproximadamente, 6000, 6100, 6200, 6300, 6400, 6500, 6557, 6558, 6559, 6560, 6561 o un número o un rango entre dos cualesquiera de estos valores si los códigos de barras estocásticos comprenden alrededor de 6561 etiquetas moleculares con secuencias distintas. En algunas formas de realización, el umbral de saturación predeterminado puede ser al menos, o como máximo, 6000, 6100, 6200, 6300, 6400, 6500, 6557, 6558, 6559, 6560 o 6561 si los códigos de barras estocásticos comprenden aproximadamente 6561 etiquetas moleculares con distintas secuencias. En algunas formas de realización, el umbral de saturación predeterminado puede ser, o ser aproximadamente, 64000, 64100, 64200, 64300, 64400, 64500, 64600, 64700, 64800, 64900, 65000, 65100, 65200, 65300. 5400, 65500, 65510, 65520, 65530, 65532, 65533, 65534, 65535, o un número o rango entre dos de estos valores si los códigos de barras estocásticos comprenden aproximadamente 65536 etiquetas moleculares con secuencias distintas. En algunas formas de realización, el umbral de saturación predeterminado puede ser al menos, o como máximo, 64000, 64100, 64200, 64300, 64400, 64500, 64600, 64700, 64800, 64900, 65000, 65100, 65200, 65400, 65500, 65510, 65520, 65530, 65532, 65533, 65534 o 65535 si los códigos de barras estocásticos comprenden aproximadamente 65536 etiquetas moleculares con secuencias distintas.

[0253] En algunas formas de realización, el estado de calidad de la diana en los datos de secuenciación se puede clasificar como secuenciación incompleta si el estado de calidad de la diana en los datos de secuenciación no es secuenciación completa y no secuenciación saturada.

Estado de calidad de secuenciación completa

[0254] Los métodos divulgados en el presente documento pueden proporcionar una estimación del número de un objetivo en una biblioteca de secuenciación si el objetivo tiene el estado de calidad de secuenciación completo. Cuando el objetivo en la biblioteca de secuenciación tiene el estado de calidad de secuenciación completo, se puede establecer un umbral a través de modelos de Poisson separados para secuenciar lecturas de códigos de barras estocásticos verdaderos y de error. El estado de calidad del objetivo puede depender de si se han observado todas las etiquetas moleculares verdaderas o reales en la profundidad de la secuenciación. Las etiquetas moleculares verdaderas o reales pueden referirse a etiquetas moleculares que no son etiquetas moleculares erróneas o falsas. Error o etiquetas moleculares falsas pueden referirse a etiquetas moleculares que tienen secuencias resultantes de errores de PCR, artefactos o errores de secuenciación.

[0255] Con referencia a la FIG. 5, en el estado de decisión 520, si una molécula objetivo tiene el estado de secuenciación completo, la forma de realización 500 continúa con el bloque 524. En el bloque 524, los errores de secuenciación de una base se pueden eliminar mediante los siguientes pasos. Paso (1), seleccione la etiqueta molecular asociada con la secuenciación más abundante leída como la primera etiqueta molecular principal si su lectura de secuenciación es mayor que 25. Por ejemplo, después de contar el número de etiquetas moleculares con secuencias distintas asociadas con el objetivo en el datos de secuenciación, seleccione la etiqueta molecular asociada con el objetivo en los datos de secuenciación con la lectura de secuenciación más alta.

[0256] Paso (2), identificar etiquetas moleculares secundarias: etiquetas moleculares con lecturas de secuenciación ≤ 3 y están separadas por una base de la primera etiqueta molecular principal; Si no se encuentra ninguna etiqueta molecular secundaria o no se encuentran etiquetas moleculares secundarias de una base, vaya al paso (5). Paso (3), realice múltiples pruebas binomiales en todas las etiquetas moleculares secundarias y en las etiquetas moleculares principales, y elimine aquellas etiquetas moleculares secundarias cuya hipótesis nula se acepte y atribuya sus lecturas de secuenciación a sus principales. Si no se acepta ninguna de las hipótesis nulas, lo que implica que todas las etiquetas moleculares secundarias no son errores de secuenciación de una base de la etiqueta molecular principal, no es necesario realizar ninguna corrección de lectura. Paso (4), actualice las secuencias de etiquetas moleculares y las lecturas de secuenciación. Por ejemplo, la aparición de la etiqueta molecular secundaria se puede atribuir a la etiqueta molecular principal si se acepta la hipótesis nula de la prueba binomial múltiple. Paso (5), elija la etiqueta molecular con la siguiente secuencia más grande leída como etiqueta molecular principal y repita los pasos anteriores hasta que no quede ninguna etiqueta molecular parental calificada o ninguna etiqueta molecular secundaria calificada.

[0257] En algunas formas de realización, el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación contados se puede ajustar, si la diana tiene el estado de calidad de secuenciación completo, determinando todas las etiquetas moleculares secundarias para una o más etiquetas moleculares parentales; realizar un análisis estadístico tal como una prueba binomial múltiple para al menos una etiqueta molecular secundaria y la etiqueta molecular principal; y atribuir la aparición de la etiqueta molecular secundaria a la etiqueta molecular principal si se acepta la hipótesis nula del análisis estadístico.

[0258] En algunas formas de realización, las etiquetas moleculares secundarias pueden comprender etiquetas moleculares que difieren de la etiqueta molecular principal en una base y tienen apariciones menores o iguales a un umbral secundario de secuenciación completa predeterminado. El umbral secundario de secuenciación completa predeterminado puede variar. En algunas formas de realización, el umbral secundario de secuenciación completa

predeterminado puede ser, o ser aproximadamente, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, o un número o un rango entre dos de estos valores. En algunas formas de realización, el umbral secundario de secuenciación completa predeterminado puede ser al menos, o como máximo, 1, 2, 3, 4, 5, 6, 7, 8, 9 o 10.

[0259] En algunas formas de realización, uno o más etiquetas moleculares originales comprenden etiquetas moleculares con apariciones mayores o iguales a un umbral principal de secuenciación completa predeterminado, en donde el umbral principal de secuenciación completa predeterminado es igual al umbral de aparición de secuenciación completa predeterminado, por ejemplo 8. La hipótesis nula del primer análisis estadístico puede aceptarse si la probabilidad de que la hipótesis nula sea verdadera está por debajo de las tasas de descubrimiento falso. Las tasas de descubrimiento falso pueden variar. En algunas formas de realización, la tasa de descubrimiento falso puede ser, o ser aproximadamente, 1 %, 2 %, 3 %, 4 %, 5 %, 6 %, 7 %, 8 %, 9 %, 10 %, 11 %, 12 %, 13 %, 14 %, 15 %, 16 %, 17 %, 18 %, 19 %, 20 %, o un número o rango entre dos de estos valores. En algunas formas de realización, la tasa de descubrimiento falso puede ser al menos, o como máximo, 1 %, 2 %, 3 %, 4 %, 5 %, 6 %, 7 %, 8 %, 9 %, 10 %, 11 %. 12 %, 13 %, 14 %, 15 %, 16 %, 17 %, 18 %, 19 % o 20 %. El primer análisis estadístico puede ser una prueba binomial múltiple.

[0260] En el bloque 528, se pueden usar modelos de Poisson para establecer umbrales de etiquetas moleculares de la diana para determinar etiquetas moleculares verdaderas y etiquetas moleculares falsas asociadas con la diana en los datos de secuenciación. Por ejemplo, los modelos de Poisson se pueden aplicar a lecturas de secuenciación para distinguir las etiquetas moleculares "probablemente verdaderas" de los artefactos.

[0261] En algunas formas de realización, el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación contados se puede ajustar, si la diana tiene el estado de calidad de secuenciación completo, umbralizando las etiquetas moleculares de la diana para determinar las etiquetas moleculares verdaderas. y etiquetas moleculares falsas asociadas con el objetivo en los datos de secuenciación. Establecer un umbral para las etiquetas moleculares de la diana puede comprender realizar un análisis estadístico de las etiquetas moleculares de la diana.

[0262] En algunas formas de realización, realizar el análisis estadístico comprende: ajustar la distribución de las etiquetas moleculares de la diana y sus apariciones a dos distribuciones de Poisson; determinar el número de etiquetas moleculares verdaderas n usando las dos distribuciones de Poisson; y eliminar las etiquetas moleculares falsas de los datos de secuenciación, en donde las etiquetas moleculares falsas comprenden etiquetas moleculares con apariciones menores que la aparición de la enésima marca molecular más abundante, y en donde las etiquetas moleculares verdaderas comprenden etiquetas moleculares con apariciones mayores o iguales a la aparición de la enésima etiqueta molecular más abundante. Las dos distribuciones de Poisson pueden comprender una primera distribución de Poisson correspondiente a las etiquetas moleculares verdaderas y una segunda distribución de Poisson para las etiquetas moleculares falsas.

[0263] En el bloque 532, se puede estimar el número del objetivo para generar salida después de corregir o ajustar los datos de secuenciación utilizando múltiples pruebas binomiales o las dos distribuciones de Poisson. La forma de realización 500 termina en el bloque final 536.

Estado de calidad de secuenciación saturada

[0264] Es posible que los métodos divulgados en el presente documento no puedan proporcionar una estimación del número de una diana en una biblioteca de secuenciación si la diana tiene el estado de calidad de secuenciación saturada debido a las grandes incertidumbres en la estimación de los recuentos de etiquetas moleculares. Haciendo referencia a la FIG. 5, en algunas formas de realización, en el estado de decisión 520, si el estado de secuenciación no es el estado de secuenciación completo, la forma de realización 500 pasa al estado de decisión 540. En el estado de decisión 540, si el objetivo tiene el estado de secuenciación saturado, la forma de realización 500 procede a el bloque final 536. Con el estado de secuenciación saturado, es posible que no se pueda determinar el número de la diana debido a las grandes incertidumbres en la estimación de los recuentos de etiquetas moleculares.

Estado de calidad de secuenciación incompleta

[0265] Los métodos divulgados en el presente documento pueden proporcionar una estimación del número de una diana en una biblioteca de secuenciación si la diana tiene el estado de calidad de secuenciación incompleta. Cuando la diana en la biblioteca de secuenciación tiene el estado de calidad de secuenciación incompleta, se puede eliminar una diana ruidosa, por ejemplo un gen ruidoso. Un objetivo puede ser ruidoso si su tasa de amplificación (lecturas promedio por etiqueta molecular) es similar a la tasa de amplificación de errores derivados de genes completamente secuenciados en la misma biblioteca que contiene el objetivo. Se puede aplicar un modelo de Poisson truncado en cero a las lecturas de secuenciación del objetivo con un estado de calidad de secuenciación incompleta para extrapolar una estimación del número de códigos de barras estocásticos que comprenden el objetivo con distintas etiquetas moleculares presentes en la biblioteca.

[0266] La forma de realización 500 puede proporcionar una estimación del número de un objetivo en una biblioteca de secuenciación si algunos de los códigos de barras estocásticos verdaderos utilizados para etiquetar el objetivo inicial no

se observaron debido a una profundidad de secuenciación inadecuada. En el estado de decisión 540, si el objetivo no tiene el estado de secuenciación saturado, entonces el objetivo tiene el estado de secuenciación incompleto, y la forma de realización 500 procede al bloque 544 para eliminar un objetivo ruidoso, por ejemplo un gen ruidoso.

[0267] Si el índice de dispersión de un objetivo es > 4 y la lectura de secuenciación máxima para ese objetivo es > 18, el uso del modelado de Poisson para derivar el umbral para separar códigos de barras verdaderos y de error aún puede proporcionar una estimación sensata. Si los datos de secuenciación muestran una sobredispersión moderada, por ejemplo 1,5 <índice de dispersión ≤ 4 y la lectura de secuenciación máxima para ese objetivo es ≤ 18, entonces el uso de modelos de Poisson para derivar el umbral puede subestimar los verdaderos recuentos de etiquetas moleculares. El motivo de la subestimación puede deberse a que las etiquetas moleculares con lecturas bajas probablemente sean una mezcla de etiquetas moleculares verdaderas y falsas. En consecuencia, esas etiquetas moleculares verdaderas con lecturas de secuenciación bajas pueden verse obligadas a incluir errores en el modelo de Poisson, y el modelo de Poisson para etiquetas moleculares verdaderas puede tener menas etiquetas moleculares de las que debería tener. Se puede usar un método ad hoc, por ejemplo usando recuentos de etiquetas moleculares después de eliminar etiquetas moleculares con recuentos bajos como uno. Si el índice de dispersión es cercano a uno, por ejemplo entre 0,9 y 1,5, entonces los recuentos de etiquetas moleculares observados pueden producir una estimación sensata. Si el índice de dispersión está entre 0,1 y 0,9, el modelo de Poisson truncado en cero que presenta el modelo de Poisson poco disperso puede producir estimaciones sensatas; pero si hay errores en los datos de secuenciación, entonces este modelo puede tender a sobreestimar.

[0268] En algunas formas de realización, el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación se puede ajustar, si el estado de calidad de la diana en los datos de secuenciación es el estado de calidad de secuenciación incompleta, determinando si la diana hay ruido en los datos de secuenciación; y eliminar el objetivo ruidoso de los datos de secuenciación. El objetivo puede ser ruidoso si la aparición de las etiquetas moleculares de los objetivos ruidosos es menor o igual que un umbral de objetivo ruidoso de secuenciación incompleta. El umbral del gen ruidoso de secuenciación incompleta puede variar. En algunas formas de realización, el umbral de diana ruidosa de secuenciación incompleta puede igualar la mediana o media de aparición de las etiquetas moleculares de la pluralidad de dianas con estados de calidad de secuenciación completa. En algunas formas de realización, el umbral del gen ruidoso de secuenciación incompleta puede ser, o ser aproximadamente, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, o un número o un rango entre dos de estos valores. En algunas formas de realización, el umbral del gen ruidoso de secuenciación incompleta puede ser al menos, o como máximo, 1, 2, 3, 4, 5, 6, 7, 8, 9 o 10.

[0269] En el bloque 548, se aplica un modelo de Poisson truncado en cero a las lecturas de secuenciación del objetivo con un estado de calidad de secuenciación incompleta para extrapolar una estimación del número de códigos de barras estocásticos que comprenden el objetivo con distintas etiquetas moleculares presentes en la biblioteca. En algunas formas de realización, el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación se ajusta, si el estado de calidad de la diana en los datos de secuenciación obtenidos es el estado de calidad de secuenciación incompleta, determinando si la diana tiene ruido en los datos de secuenciación; y eliminar el objetivo que sea ruidoso.

[0270] En algunas formas de realización, el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación se puede ajustar, si el estado de calidad de la diana en los datos de secuenciación es el estado de calidad de secuenciación incompleta, poniendo un umbral a las etiquetas moleculares del objetivo para determinar etiquetas moleculares verdaderas y etiquetas moleculares falsas en los datos de secuenciación. Establecer un umbral para las etiquetas moleculares de la diana puede comprender realizar un análisis estadístico de las etiquetas moleculares. Realizar el análisis estadístico de las etiquetas moleculares puede comprender: determinar el número de etiquetas moleculares verdaderas n usando un modelo de Poisson truncado en cero; y eliminar las etiquetas moleculares falsas de los datos de secuenciación.

[0271] En algunas formas de realización, las etiquetas moleculares falsas pueden comprender etiquetas moleculares con apariciones inferiores a la aparición del enésimo marcador molecular más abundante. Los verdaderas etiquetas moleculares pueden comprender etiquetas moleculares con apariciones mayores o iguales que la aparición del enésimo marcador molecular más abundante.

Errores de datos de secuenciación

[0272] Los métodos descritos en el presente documento se pueden usar para identificar y/o corregir errores de datos de secuenciación, por ejemplo, los errores que ocurren en los métodos para contar uno o más ácidos nucleicos diana. En algunas formas de realización, el error puede comprender, o ser, una eliminación de uno o más nucleótidos, una sustitución de uno o más nucleótidos, una adición de uno o más nucleótidos, o cualquier combinación de los mismos. El error puede estar presente en una etiqueta molecular (EM), una etiqueta de muestra (SL) u otra etiqueta en un código de barras estocástico. En algunas formas de realización, un error de datos de secuenciación puede comprender, o ser, un error introducido por PCR, un error introducido por secuenciación, un error de contaminación del cebador de transcripción inversa (TI) o cualquier combinación de los mismos. El error introducido por la PCR puede comprender, o ser, el resultado de un error de amplificación por PCR, un sesgo de amplificación por PCR, una amplificación por PCR insuficiente o cualquier combinación de los mismos. El error introducido por la secuenciación puede comprender, o ser, el resultado de

una llamada de base inexacta, una secuenciación insuficiente o cualquier combinación de los mismos. El error de contaminación del cebador TI puede ser un error causado por un cebador de transcripción inversa que ingresa a la PCR.

[0273] Como se usa en el presente documento, el término "cobertura" o "profundidad de secuenciación" puede referirse al número de lecturas de un objetivo con código de barras con una EM particular y un SL particular en datos de secuenciación. Por ejemplo, un objetivo con código de barras puede secuenciarse varias veces. En consecuencia, el objetivo del código de barras con una EM y SL particular se puede observar varias veces. Como otro ejemplo, una célula puede contener múltiples copias de una diana (por ejemplo, múltiples copias de moléculas de ARNm de un gen). Estas múltiples copias del objetivo pueden tener códigos de barras. Después de la amplificación por PCR (por ejemplo, el bloque 28 en la FIG.), puede haber múltiples copias de un objetivo con código de barras con una EM y SL particulares. Durante la secuenciación, se pueden secuenciar algunas o todas las copias múltiples del objetivo con código de barras con las EM y SL particulares. El número de lecturas del objetivo con código de barras con el mismo EM y SL observado en los datos de secuenciación puede denominarse "cobertura" o "profundidad de secuenciación".

[0274] En algunas formas de realización, los errores de datos de secuenciación se pueden identificar y/o corregir. Por ejemplo, las copias de un objetivo de una célula pueden tener códigos de barras con diferentes EM y el mismo SL. El objetivo con código de barras con una EM puede tener múltiples lecturas en los datos de secuenciación. El objetivo con código de barras con una EM diferente puede tener solo unas pocas lecturas (por ejemplo, una lectura). Es más probable que el primer objetivo con código de barras tenga una EM verdadero (o EM real o de señal), en comparación con el último objetivo con código de barras. Este último objetivo con código de barras puede incluir una EM de error (o una EM falso o ruidoso). Esto puede deberse a que se puede esperar que los dos EM tengan coberturas o profundidades de secuencia similares. Este último objetivo con código de barras con solo unas pocas lecturas puede ser un artefacto o error generado durante la secuenciación o la PCR.

[0275] Como otro ejemplo, un código de barras estocástico que ingresa a la PCR puede dar como resultado un error de contaminación del cebador de TI. En algunas formas de realización, después de la transcripción inversa de moléculas de ARNm en moléculas de ADNc (por ejemplo, 24 de la Figura), los códigos de barras estocásticos no incorporados en las moléculas de ADNc se pueden eliminar mediante, por ejemplo, purificación con perlas Ampure. Es posible que el método de eliminación, por ejemplo la purificación con perlas Ampure, no elimine completamente los códigos de barras estocásticos que no se extienden mediante transcripción inversa para incorporarse en moléculas de ADNc con código de barras estocástico. Por ejemplo, 15 %, 10 %, 9 %, 8 %, 7 %, 6 %, 5 %, 4 %, 3 %, 2 %, 1 %, 0,5 %, 0,1 %, o un rango entre cualquiera de estos dos valores de códigos de barras estocásticos que no se extienden mediante transcripción inversa para incorporarse en moléculas de ADNc con código de barras estocástico no se pueden eliminar mediante la purificación con perlas Ampure. Estos códigos de barras estocásticos no eliminados pueden dar lugar a errores en los datos de secuenciación durante la amplificación de moléculas de ADNc (por ejemplo, en el bloque 28 de la FIG.). Los códigos de barras estocásticos entre muestras pueden ser muy similares. Por ejemplo, las etiquetas de muestra de códigos de barras estocásticos pueden ser idénticas para una muestra. Por lo tanto, el cruce de PCR puede ocurrir porque estos códigos de barras estocásticos no eliminados pueden hibridarse con otras moléculas de ácido nucleico de la misma muestra (por ejemplo, las regiones SL de moléculas de ARNm con códigos de barras estocásticos) durante la PCR y pueden resultar en errores de datos de secuenciación conocidos como errores de SL.

[0276] Los EM verdaderos, las EM de error y los errores de SL pueden tener distribuciones distintas. FIG. 4 es una ilustración esquemática que muestra distribuciones ejemplares no limitantes de errores de etiquetas moleculares, errores de etiquetas de muestras y señales de etiquetas moleculares verdaderas. Como se ilustra en la FIG. 4, es más probable que los IM de error tengan una cobertura de EM más baja porque los IM de error pueden ser resultados de errores de PCR o de secuenciación. Por ejemplo, los IM de error pueden ser el resultado principalmente de errores de secuenciación y algunos errores de PCR. Es más probable que los errores de SL tengan una cobertura de EM más baja porque los errores de SL pueden deberse principalmente a códigos de barras estocásticos que ingresan a PCR.

Corrección de errores de secuenciación y PCR según la adyacencia direccional

[0277] En el presente documento se describen métodos para corregir errores de PCR o secuenciación. En algunas formas de realización, el método comprende: (a) recibir datos de secuenciación de objetivos con códigos de barras estocásticos. Los objetivos con códigos de barras estocásticos pueden obtenerse codificando de forma estocástica una pluralidad de objetivos usando una pluralidad de códigos de barras estocásticos para crear una pluralidad de objetivos con códigos de barras estocásticos, en donde cada uno de la pluralidad de códigos de barras estocásticos comprende una etiqueta molecular. En algunas formas de realización, el método comprende: (b) para una o más de la pluralidad de dianas: (i) contar el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación; (ii) identificar grupos de etiquetas moleculares de la diana usando adyacencia direccional; (iii) colapsar los datos de secuenciación recibidos en (b) usando los grupos de etiquetas moleculares de la diana identificada en (ii); y (iv) estimar el número de la diana, en donde el número de la diana estimado se correlaciona con el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación contados en (i) después de colapsar los datos de secuenciación en (ii). La pluralidad de dianas puede comprender dianas de todo el transcriptoma de una célula. En algunas formas de realización, el método comprende además: (c) codificar de forma estocástica la pluralidad de objetivos utilizando la pluralidad de códigos de barras estocásticos para crear la pluralidad de objetivos con

códigos de barras estocásticos; y (d) secuenciar los objetivos con códigos de barras estocásticos para generar los datos de secuenciación de los objetivos con códigos de barras estocásticos recibidos.

[0278] FIG. 7 es un diagrama de flujo que muestra una forma de realización ejemplar no limitante 700 de corrección de errores de secuenciación y PCR usando etiquetas moleculares basadas en la adyacencia direccional. La corrección de errores de PCR y secuenciación usando etiquetas moleculares basadas en la adyacencia direccional puede denominarse corrección de errores de sustitución recursiva (RSEC). El método 700 comienza en el bloque 704 después de recibir datos de secuenciación de una pluralidad de objetivos con códigos de barras estocásticos. En algunas formas de realización, el método 700 comprende además codificar de forma estocástica una pluralidad de objetivos usando una pluralidad de códigos de barras estocásticos para crear la pluralidad de objetivos con códigos de barras estocásticos, en donde cada uno de la pluralidad de códigos de barras estocásticos comprende una etiqueta molecular. En algunas formas de realización, el método 700 comprende además secuenciar la pluralidad de objetivos con códigos de barras estocásticos para obtener los datos de secuenciación.

[0279] En el bloque 708, para uno o más de la pluralidad de objetivos: se puede contar el número de etiquetas moleculares con secuencias distintas asociadas con el objetivo en los datos de secuenciación. En el bloque 712, se pueden identificar grupos de etiquetas moleculares de la diana usando adyacencia direccional. Las etiquetas moleculares del objetivo dentro de un grupo pueden estar dentro de un umbral de adyacencia direccional predeterminado entre sí. El umbral de adyacencia direccional puede variar. En algunas formas de realización, el umbral de adyacencia direccional predeterminado puede ser aproximadamente, ser al menos o ser como máximo una distancia de Hamming de uno o dos.

[0280] En algunas formas de realización, las etiquetas moleculares de la diana dentro del grupo pueden comprender uno o más etiquetas moleculares originales y etiquetas moleculares secundarios de uno o más etiquetas moleculares originales. La aparición del marcador molecular original puede ser mayor o igual a un umbral de aparición de adyacencia direccional predeterminado. En algunas formas de realización, el umbral de aparición de adyacencia direccional predeterminado puede ser aproximadamente, ser al menos o ser como máximo el doble de la aparición de una etiqueta molecular secundaria menos uno. En algunas formas de realización, el umbral de aparición de adyacencia direccional predeterminado puede ser, o ser aproximadamente 1,5 veces, 2 veces, 3 veces, 4 veces, 5 veces, 6 veces, 7 veces, 8 veces, 9 veces, 10 veces, o un número o un rango entre dos de estos valores, la aparición de una etiqueta molecular secundaria. En algunas formas de realización, el umbral de aparición de adyacencia direccional predeterminado puede ser, al menos o como máximo 1,5 veces, 2 veces, 3 veces, 4 veces, 5 veces, 6 veces, 7 veces, 8 veces, 9 veces o 10 veces, el aparición de una etiqueta molecular secundaria.

[0281] En el bloque 720, los datos de secuenciación se colapsan usando los grupos de etiquetas moleculares de la diana. Colapsar los datos de secuenciación puede comprender atribuir la aparición de la etiqueta molecular secundaria a la etiqueta molecular principal. En el bloque 732, se puede estimar el número del objetivo para generar salida después de colapsar los datos de secuenciación. El método 700 termina en el bloque 736.

[0282] En algunas formas de realización, los métodos comprenden además: determinar una profundidad de secuenciación del objetivo. Estimar el número del objetivo, si la profundidad de secuenciación del objetivo está por encima de un umbral de profundidad de secuenciación predeterminado, comprende ajustar los datos de secuenciación contados en (i). El umbral de profundidad de secuenciación predeterminado puede estar entre 15 y 20. Ajustar los datos de secuenciación contados en (i) comprende: establecer un umbral de etiquetas moleculares de la diana para determinar etiquetas moleculares verdaderas y etiquetas moleculares falsas asociadas con la diana en los datos de secuenciación obtenidos en (b). Establecer un umbral para las etiquetas moleculares de la diana comprende realizar un análisis estadístico de las etiquetas moleculares de la diana. Realizar el análisis estadístico comprende: ajustar la distribución de las etiquetas moleculares del objetivo y sus apariciones a dos distribuciones tales como dos distribuciones binomiales negativas; determinar el número de etiquetas moleculares verdaderas n usando las dos distribuciones binomiales negativas; y eliminar las etiquetas moleculares falsas de los datos de secuenciación obtenidos en (b), en el que las etiquetas moleculares falsas comprenden etiquetas moleculares con apariciones inferiores a la aparición del enésimo marcador molecular más abundante, y en el que las etiquetas moleculares verdaderas comprenden etiquetas moleculares con apariciones mayores mayor o igual a la aparición de la enésima etiqueta molecular más abundante.

<u>Corrección de errores de secuenciación y PCR basados en adyacencia direccional y segundas derivadas</u>

[0283] En el presente documento se describen métodos para determinar el número de objetivos. En algunas formas de realización, un método comprende: (a) codificar de barras estocásticamente una pluralidad de objetivos usando una pluralidad de códigos de barras estocásticos para crear una pluralidad de objetivos con códigos de barras estocásticos, en donde cada uno de la pluralidad de códigos de barras estocásticos comprende una etiqueta molecular; (b) obtener datos de secuenciación de los objetivos con códigos de barras estocásticos; y (c) para una o más de la pluralidad de dianas: (i) contar el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación; (ii) identificar grupos de etiquetas moleculares de la diana usando adyacencia direccional; (iii) colapsar los datos de secuenciación obtenidos en (b) usando los grupos de etiquetas moleculares de la diana identificada en (ii); y (iv) estimar el número de la diana, en donde el número de la diana estimado se correlaciona con el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación contados en (i) después de

colapsar los datos de secuenciación en (ii). La pluralidad de dianas puede comprender dianas de todo el transcriptoma de una célula.

[0284] En algunas formas de realización, el método comprende determinar un estado de secuenciación del objetivo en los datos de secuenciación. El estado de secuenciación del objetivo en los datos de secuenciación puede incluir, o ser, secuenciación saturada. En algunas formas de realización, si el estado de secuenciación de la diana en los datos de secuenciación es el estado de secuenciación saturado, el número de la diana estimado en (iv) se correlaciona con el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación contados en (i).

[0285] En algunas formas de realización, el número de la diana estimado se correlaciona con el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación contados en (i) después de corregir los errores de SL. La corrección de errores de SL puede comprender generar un gráfico de suma acumulativa de las etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación; determinar segundas derivadas del gráfico de suma acumulativa; y determinar un límite de profundidad de lectura de EM basado en un mínimo de las segundas derivadas del gráfico de suma acumulativa. En algunas formas de realización, la corrección de los errores de SL puede incluir la eliminación de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación con profundidades de lectura inferiores al límite de profundidad de lectura de EM determinado.

[0286] FIG. 8 es un diagrama de flujo que muestra una forma de realización ejemplar no limitante 800 de corrección de errores de secuenciación y PCR usando etiquetas moleculares basadas en adyacencia direccional y segundas derivadas. El método 800 comienza en el bloque 804 después de recibir datos de secuenciación de una pluralidad de objetivos con códigos de barras estocásticos. En algunas formas de realización, el método 800 comprende además codificar de forma estocástica una pluralidad de objetivos usando una pluralidad de códigos de barras estocásticos para crear la pluralidad de objetivos con códigos de barras estocásticos, en donde cada uno de la pluralidad de códigos de barras estocásticos comprende una etiqueta molecular. En algunas formas de realización, el método 800 comprende además secuenciar la pluralidad de objetivos con códigos de barras estocásticos para obtener los datos de secuenciación.

[0287] En el bloque 808, para uno o más de la pluralidad de objetivos: se puede contar el número de etiquetas moleculares con secuencias distintas asociadas con el objetivo en los datos de secuenciación. En el bloque de decisión 812, se puede determinar si los datos de secuenciación tienen un estado de secuenciación saturado. El estado de secuenciación saturada puede determinarse si la diana tiene una serie de etiquetas moleculares con secuencias distintas mayores que un umbral de saturación predeterminado. El umbral de saturación predeterminado puede ser diferente en diferentes implementaciones. Por ejemplo, el umbral de saturación predeterminado puede ser aproximadamente 6557 si los códigos de barras estocásticos comprenden aproximadamente 6561 etiquetas moleculares con secuencias distintas. Como otro ejemplo, el umbral de saturación predeterminado puede ser aproximadamente 65532 si los códigos de barras estocásticos comprenden aproximadamente 65536 etiquetas moleculares con secuencias distintas.

[0288] Si los datos de secuenciación no tienen un estado de secuenciación saturado en el bloque de decisión 812, el método 800 puede proceder al bloque 816, en el que los recuentos de etiquetas moleculares se pueden ajustar en función de la adyacencia direccional. Por ejemplo, se puede considerar que el objetivo tiene el estado de secuenciación saturada si tiene una cantidad de etiquetas moleculares con secuencias distintas mayores que 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10000, 20000, 30000., 40000, 50000, 60000, 70000, 80000, 90000, 100000, o un número o rango entre dos de estos. Como otro ejemplo, se puede considerar que la diana tiene el estado de secuenciación saturada si tiene una cantidad de etiquetas moleculares con secuencias distintas superiores al 50 %, 60 %, 70 %, 80 %, 90 %, 95 %, 99 %, 99,9 %, o un número o un rango entre dos cualesquiera de estos, de los códigos de barras moleculares de los códigos de barras estocásticos con secuencias distintas. En algunas formas de realización, ajustar los recuentos moleculares basándose en la adyacencia direccional puede ser como se describe con referencia a la FIG 7. Por ejemplo, ajustar los recuentos moleculares basándose en el diccionario puede incluir identificar grupos de etiquetas moleculares de la diana usando adyacencia direccional; colapsar los datos de secuenciación utilizando los grupos de etiquetas moleculares de la diana identificada; y estimar el número de la diana, en donde el número de la diana estimado se correlaciona con el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación contados después de colapsar los datos de secuenciación.

[0289] En el bloque 820, se pueden determinar las segundas derivadas de un gráfico de suma acumulativa. La determinación de las segundas derivadas del gráfico de suma acumulativa puede incluir generar el gráfico de suma acumulativa de las etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación.

[0290] En el bloque 824, las etiquetas moleculares se pueden ajustar en función de un límite de profundidad de lectura de EM. El límite de profundidad de lectura de EM puede basarse en un mínimo (como un mínimo local o un mínimo global) de la segunda derivada del gráfico de suma acumulativa. En algunas formas de realización, la corrección de los errores de SL puede incluir la eliminación de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación con profundidades de lectura inferiores al límite de profundidad de lectura de EM determinado.

[0291] En el bloque 828, se puede estimar el número del objetivo para generar salida después de colapsar los datos de secuenciación y corregir los errores de SL. En el bloque de decisión 812, si los datos de secuenciación tienen el estado de secuenciación saturado, el método 800 puede proceder al bloque 828 para generar salida sin colapsar los datos de secuenciación y corregir errores de SL. El método 800 termina en el bloque 832.

Corrección de errores de secuenciación y PCR según la adyacencia direccional y la corrección de errores basada en la distribución

[0292] En el presente documento se describen métodos para corregir errores de PCR o secuenciación. Los métodos se pueden utilizar para determinar el número de objetivos. En algunas formas de realización, el método comprende: (a) recibir datos de secuenciación de objetivos con códigos de barras estocásticos. Los objetivos con códigos de barras estocásticos se pueden obtener codificando con barras estocásticamente una pluralidad de objetivos usando una pluralidad de códigos de barras estocásticos para crear una pluralidad de objetivos con códigos de barras estocásticos, en donde cada uno de la pluralidad de códigos de barras estocásticos comprende una etiqueta molecular. En algunas formas de realización, el método comprende (b) para una o más de la pluralidad de dianas: (i) contar el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación; (ii) determinar una serie de etiquetas moleculares de ruido con secuencias distintas asociadas con la diana en los datos de secuenciación; y (iii) estimar el número de la diana, en donde el número de la diana estimado se correlaciona con el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación contados en (i) ajustados de acuerdo con la cantidad de etiquetas moleculares de ruido determinado en (ii). En algunas formas de realización, el método comprende determinar un estado de secuenciación del objetivo en los datos de secuenciación. En algunas formas de realización, el método comprende además: (c) codificar de forma estocástica la pluralidad de objetivos utilizando la pluralidad de códigos de barras estocásticos para crear la pluralidad de objetivos con códigos de barras estocásticos; y (d) secuenciar los objetivos con códigos de barras estocásticos para generar los datos de secuenciación de los objetivos con códigos de barras estocásticos recibidos.

[0293] FIG. 9 es un diagrama de flujo que muestra una forma de realización ejemplar no limitante 900 de corrección de errores de secuenciación y PCR basándose en la corrección de errores de sustitución recursiva y la corrección de errores basada en distribución. El método 900 comienza en el bloque 904 después de recibir datos de secuenciación de una pluralidad de objetivos con códigos de barras estocásticos. En algunas formas de realización, el método 900 comprende además codificar de forma estocástica una pluralidad de objetivos usando una pluralidad de códigos de barras estocásticos para crear la pluralidad de objetivos con códigos de barras estocásticos, en donde cada uno de la pluralidad de códigos de barras estocásticos comprende una etiqueta molecular. En algunas formas de realización, el método 900 comprende además secuenciar la pluralidad de objetivos con códigos de barras estocásticos para obtener los datos de secuenciación.

[0294] En el bloque 908, para una o más de la pluralidad de objetivos: se puede contar el número de etiquetas moleculares con secuencias distintas asociadas con el objetivo en los datos de secuenciación. En el bloque de decisión 912, se puede determinar si los datos de secuenciación tienen un estado de secuenciación saturado. Por ejemplo, se puede considerar que el objetivo tiene el estado de secuenciación saturada si tiene una cantidad de etiquetas moleculares con secuencias distintas mayores que 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10000, 20000, 30000, 40000, 50000, 60000, 70000, 80000, 90000, 100000, o un número o rango entre dos de estos. Como otro ejemplo, se puede considerar que la diana tiene el estado de secuenciación saturada si tiene una cantidad de etiquetas moleculares con secuencias distintas superiores al 50 %, 60 %, 70 %, 80 %, 90 %, 95 %, 99 %, 99,9 %, o un número o un rango entre dos cualesquiera de estos, de los códigos de barras moleculares de los códigos de barras estocásticos con secuencias distintas.

[0295] En algunas formas de realización, el estado de secuenciación saturada puede determinarse si la diana tiene una serie de etiquetas moleculares con secuencias distintas mayores que un umbral de saturación predeterminado. El umbral de saturación predeterminado puede ser diferente en diferentes implementaciones. Por ejemplo, el umbral de saturación predeterminado puede ser, o ser aproximadamente, 1000, 2000, 3000, 4000, 5000, 6000, 6557, 7000, 8000, 9000, 10000, 20000, 30000, 40000, 50000, 65532, 70000, 80000, 90000, 100000, o un número o rango entre dos de estos valores. Como otro ejemplo, el umbral de saturación predeterminado puede ser al menos, o como máximo, 1000, 2000, 3000, 4000, 5000, 6000, 6557, 7000, 8000, 9000, 10000, 20000, 30000, 40000, 50000, 60000, 65532, 70000, 80000, 90000 o 100000.

[0296] En algunas formas de realización, el estado de secuenciación saturada puede depender del número de etiquetas moleculares de los códigos de barras estocásticos con secuencias distintas. Por ejemplo, el umbral de saturación predeterminado puede ser aproximadamente 6557 si los códigos de barras estocásticos comprenden aproximadamente 6561 etiquetas moleculares con secuencias distintas. Como otro ejemplo, el umbral de saturación predeterminado puede ser aproximadamente 65532 si los códigos de barras estocásticos comprenden aproximadamente 65536 etiquetas moleculares con secuencias distintas. En algunas formas de realización, el estado de secuenciación saturada puede no depender del número de etiquetas moleculares de los códigos de barras estocásticos con secuencias distintas.

[0297] Si los datos de secuenciación no tienen un estado de secuenciación saturado en el bloque de decisión 912, el método 900 puede proceder al bloque 916, donde los recuentos de etiquetas moleculares se pueden ajustar en función

de la adyacencia direccional. En algunas formas de realización, ajustar los recuentos moleculares basándose en la adyacencia direccional puede ser como se describe con referencia a la FIG 7. Por ejemplo, ajustar los recuentos moleculares basándose en el diccionario puede incluir identificar grupos de etiquetas moleculares de la diana usando adyacencia direccional; colapsar los datos de secuenciación utilizando los grupos de etiquetas moleculares de la diana identificada; y estimar el número de la diana, en donde el número de la diana estimado se correlaciona con el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación contados después de colapsar los datos de secuenciación.

[0298] En el bloque 920, se puede determinar el estado de secuenciación del objetivo en los datos de secuenciación. El estado de secuenciación del objetivo en los datos de secuenciación puede incluir, o estar, bajo secuenciación. En el bloque de decisión 924, se puede determinar si el estado de secuenciación del objetivo en los datos de secuenciación es el estado de secuenciación insuficiente. Por ejemplo, se puede considerar que el objetivo tiene el estado de secuenciación insuficiente si su profundidad (p. ej., una profundidad promedio, mínima o máxima) es menor o menor que aproximadamente 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, o un número o rango entre dos de estos. Como otro ejemplo, se puede considerar que el objetivo tiene el estado de secuenciación insuficiente si su profundidad es menor que al menos, o como máximo, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50, 60, 70, 80, 90 o 100.

[0299] En algunas formas de realización, el estado de subsecuenciación puede determinarse si el objetivo tiene una profundidad (por ejemplo, una profundidad promedio, mínima o máxima) menor que un umbral de subsecuenciación predeterminado. El umbral de secuenciación insuficiente puede ser diferente en diferentes implementaciones. Por ejemplo, el umbral de secuenciación inferior puede ser, o ser aproximadamente, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, o un número o rango entre dos de estos valores. Como otro ejemplo, el umbral de secuenciación insuficiente puede ser al menos o como máximo 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50, 60, 70, 80., 90 o 100.

[0300] En algunas formas de realización, el estado de secuenciación insuficiente puede depender del número de etiquetas moleculares de los códigos de barras estocásticos con secuencias distintas. Por ejemplo, el umbral de secuenciación insuficiente puede ser 10 (u otro número de umbral) si los códigos de barras estocásticos comprenden, o aproximadamente, 1000, 2000, 3000, 4000, 5000, 6000, 6561, 7000, 8000, 9000, 10000, 20000, 30000, 40000, 50000, 60000, 65532, 70000, 80000, 90000, 100000, o un número o rango entre dos de estos valores, etiquetas moleculares con secuencias distintas. Como otro ejemplo, el umbral de secuenciación insuficiente puede ser 10 (u otro número de umbral) si los códigos de barras estocásticos comprenden al menos, o como máximo, 1000, 2000, 3000, 4000, 5000, 6000, 6561, 7000, 8000, 9000, 10000., 20000, 30000, 40000, 50000, 60000, 65532, 70000, 80000, 90000 o 100000. En algunas formas de realización, el estado de secuenciación insuficiente puede no depender del número de etiquetas moleculares de los códigos de barras estocásticos con secuencias distintas.

[0301] En el bloque de decisión 924, si el estado de secuenciación de la diana en los datos de secuenciación no es el estado de secuenciación insuficiente, el método 900 puede proceder al bloque 928 para filtrar los recuentos de etiquetas moleculares. El filtrado de recuentos de etiquetas moleculares puede incluir, en el bloque de decisión 932, determinar el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación que son menores que un umbral de pseudopuntos. El umbral de pseudopuntos puede ser diferente en diferentes implementaciones. Por ejemplo, el umbral de pseudopuntos puede ser, o ser aproximadamente, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, o un número o rango entre dos de estos valores si los códigos de barras estocásticos comprenden aproximadamente 6561 etiquetas moleculares con secuencias distintas. Como otro ejemplo, el umbral de secuenciación de pseudopuntos puede ser al menos o como máximo 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50, 60, 70, 80., 90 o 100, si los códigos de barras estocásticos comprenden aproximadamente 6561 etiquetas moleculares con secuencias distintas.

[0302] En el bloque de decisión 932, si el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación es menor que el umbral de pseudopuntos, el método 900 puede opcionalmente proceder al bloque 936, donde se pueden agregar pseudopuntos al número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación antes de determinar el número de etiquetas moleculares de ruido con secuencias distintas asociadas con la diana en los datos de secuenciación. Los pseudopuntos pueden tener diferentes recuentos de etiquetas moleculares en diferentes implementaciones. Por ejemplo, el recuento de etiquetas moleculares de un pseudopunto puede ser, o ser aproximadamente, 0,0001, 0,001, 0,01, 0,1, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40., 50, 60, 70, 80, 90, 100, o un número o rango entre dos de estos valores. Como otro ejemplo, el recuento de etiquetas moleculares de un pseudopunto puede ser al menos o como máximo 0,0001, 0,001, 0,01, 0,1, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20. 30, 40, 50, 60, 70, 80, 90 o 100. En algunas formas de realización, si el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación es menor que el umbral de pseudopuntos, el método 900 puede continuar para bloquear si el número de etiquetas moleculares con secuencias distintas asociadas con el objetivo en los datos de secuenciación es menor que el umbral de pseudopuntos, el método 944.

[0303] En el bloque de decisión 932, si el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación no es menor que el umbral de pseudopuntos, las etiquetas moleculares no únicas se pueden eliminar en el bloque 940. Las etiquetas se pueden eliminar para determinar el número de etiquetas moleculares de ruido con secuencias distintas asociadas con la diana en los datos de secuenciación en el bloque 944. Las etiquetas moleculares no únicas pueden incluir etiquetas moleculares con secuencias distintas asociadas con la diana en los datos

de secuenciación que son mayores que un umbral de etiqueta molecular reciclado predeterminado. El umbral de la etiqueta molecular reciclada puede ser diferente en diferentes implementaciones. Por ejemplo, el umbral de la etiqueta molecular reciclada puede ser, o ser aproximadamente, 100, 200, 300, 400, 500, 600, 650, 700, 900, 1000, 2000, o un número o un rango entre dos cualesquiera de estos valores. si los códigos de barras estocásticos comprenden alrededor de 6561 etiquetas moleculares con secuencias distintas. Como otro ejemplo, el umbral de etiquetas moleculares recicladas puede ser al menos, o como máximo, 100, 200, 300, 400, 500, 600, 650, 700, 900, 1000 o 2000, si los códigos de barras estocásticos comprenden aproximadamente 6561 etiquetas moleculares con secuencias distintas.

[0304] En algunas formas de realización, eliminar las etiquetas moleculares no únicos comprende: determinar un número teórico de etiquetas moleculares no únicos para el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación. La eliminación de las etiquetas moleculares no únicos puede comprender la eliminación de una etiqueta molecular con una aparición mayor que el enésimo marcador molecular más abundante de las etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación. El número n puede ser el número teórico de etiquetas moleculares no únicas.

[0305] En el bloque 944, los recuentos de etiquetas moleculares se pueden ajustar usando un método de corrección de errores basado en distribución. El método de corrección de errores basado en la distribución puede incluir la determinación del número de etiquetas moleculares de ruido con distintas secuencias asociadas con el objetivo en los datos de secuenciación. La determinación del número de etiquetas moleculares de ruido puede comprender: ajustar dos distribuciones binomiales negativas al número de etiquetas moleculares con secuencias distintas asociadas con el objetivo en los datos de secuenciación. Por ejemplo, determinar el número de etiquetas moleculares de ruido puede comprender: ajustar una distribución binomial negativa de señal (una de las dos distribuciones binomiales negativas) al número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación contados, en donde la distribución binomial negativa de señal corresponde a una serie de etiquetas moleculares con secuencias distintas asociadas con el objetivo en los datos de secuenciación que se cuentan como etiquetas moleculares de señal. La determinación del número de etiquetas moleculares de ruido puede comprender: ajustar una distribución binomial de ruido negativo (la otra de las dos distribuciones binomiales negativas) al número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación contados, en donde la combinación binomial de ruido negativo La distribución corresponde a una serie de etiquetas moleculares con secuencias distintas asociadas con el objetivo en los datos de secuenciación que se cuentan como etiquetas moleculares de ruido. La determinación del número de etiquetas moleculares de ruido puede comprender determinar el número de etiquetas moleculares de ruido utilizando la distribución binomial negativa de señal ajustada y la distribución binomial negativa de ruido ajustada.

[0306] En algunas formas de realización, determinar el número de etiquetas moleculares de ruido utilizando la distribución binomial negativa de señal ajustada y la distribución binomial negativa de ruido ajustada comprende, para cada una de las distintas secuencias asociadas con la diana en los datos de secuenciación: determinar una probabilidad de señal de la secuencia distinta esté en la distribución binomial de señal negativa. Y se puede determinar una probabilidad de ruido de que la secuencia distinta esté en la distribución binomial de ruido negativo. Además, se puede determinar que la secuencia distinta es una etiqueta molecular de ruido si la probabilidad de la señal es menor que la probabilidad del ruido. En algunas formas de realización, ajustar los recuentos de etiquetas moleculares en el bloque 944 puede incluir eliminar singletons (por ejemplo, sustituciones de bases únicas) si se encuentran menos de dos picos (porque pueden ser necesarios dos picos para determinar la distribución binomial negativa de la señal y la distribución binomial negativa del ruido).

[0307] En el bloque 948, se puede estimar el número del objetivo para generar salida después de correcciones de errores basadas en adyacencia y distribución. En el bloque de decisión 912, si el estado de secuenciación del objetivo en los datos de secuenciación es el estado de secuenciación saturado, el método 900 puede pasar al bloque 948 para generar resultados sin ajustar las etiquetas moleculares basándose en la adyacencia direccional y la corrección de errores basada en la distribución. Por ejemplo, el número de marcas moleculares de ruido determinadas puede ser cero.

[0308] En el bloque de decisión 924, si el estado de secuenciación del objetivo en los datos de secuenciación es el estado de secuenciación insuficiente, el método 900 puede proceder al bloque 948 para generar resultados sin ajustar las etiquetas moleculares basándose en la corrección de errores basada en la distribución. Por ejemplo, el número de marcas moleculares de ruido determinadas puede ser cero. El método 900 termina en el bloque 952.

[0309] FIG. 10 es un diagrama de flujo que muestra una forma de realización ejemplar 1000 no limitante de corrección de errores usando dos distribuciones binomiales negativas. Los bloques del método 1000 (tales como los bloques 904-952) se han descrito con referencia a la FIG. 9. Brevemente, el método 1000 comienza en el bloque 904 después de recibir datos de secuenciación de una pluralidad de objetivos con códigos de barras estocásticos. En algunas formas de realización, el método 1000 comprende además codificar de forma estocástica una pluralidad de objetivos usando una pluralidad de códigos de barras estocásticos para crear la pluralidad de objetivos con códigos de barras estocásticos, en donde cada uno de la pluralidad de códigos de barras estocásticos comprende una etiqueta molecular. En algunas formas de realización, el método 1000 comprende además secuenciar la pluralidad de objetivos con códigos de barras estocásticos para obtener los datos de secuenciación.

[0310] En el bloque 908, para uno o más de la pluralidad de objetivos: se puede contar el número de etiquetas moleculares con secuencias distintas asociadas con el objetivo en los datos de secuenciación. En el bloque 916, los recuentos de etiquetas moleculares se pueden ajustar basándose en la adyacencia direccional. En algunas formas de realización, el ajuste de los recuentos moleculares en función de la adyacencia direccional puede ser como se describe con referencia a la FIG 7.

[0311] En el bloque 920, se puede determinar el estado de secuenciación del objetivo en los datos de secuenciación. El estado de secuenciación del objetivo en los datos de secuenciación puede incluir, o estar, bajo secuenciación. En el bloque de decisión 924, se puede determinar si el estado de secuenciación del objetivo en los datos de secuenciación es el estado de secuenciación insuficiente.

[0312] En el bloque de decisión 924, si el estado de secuenciación del objetivo en los datos de secuenciación no es el estado de secuenciación inferior, el método 1000 puede opcionalmente proceder al bloque de decisión 1004. En el bloque de decisión 1004, la profundidad de secuenciación del objetivo puede ser en comparación con un umbral de profundidad de secuenciación predeterminado. El umbral de profundidad de secuenciación puede ser diferente en diferentes implementaciones. Por ejemplo, la profundidad de secuenciación del objetivo puede ser, o ser aproximadamente, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100 o un número o rango entre dos de estos valores. Como otro ejemplo, la profundidad de secuenciación del objetivo puede ser al menos, o como máximo, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50, 60, 70, 80, 90 o 100.

[0313] Si la profundidad de secuenciación del objetivo es mayor que el umbral de profundidad de secuenciación, el método 1000 continúa con el bloque 928. Si la profundidad de secuenciación del objetivo no es mayor que el umbral de profundidad de secuenciación, el método 1000 continúa con el bloque 1008. En el bloque 1008, los singletons (por ejemplo, sustituciones de una sola base) se pueden eliminar antes de generar la salida en el bloque 948.

[0314] En el bloque 928, se pueden filtrar los recuentos de etiquetas moleculares. El filtrado de recuentos de etiquetas moleculares puede incluir, en el bloque de decisión 912, determinar si los datos de secuenciación tienen un estado de secuenciación saturado. Si los datos de secuenciación no tienen un estado de secuenciación saturado en el bloque de decisión 912, el método 1000 puede proceder al bloque de decisión 932. En el bloque de decisión 932, el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación que son menores que un umbral de pseudopuntos puede ser determinado.

[0315] En el bloque de decisión 932, si el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación es menor que el umbral de pseudopuntos, el método 900 puede opcionalmente proceder al bloque 936, donde se pueden agregar pseudopuntos al número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación antes de determinar el número de etiquetas moleculares de ruido con secuencias distintas asociadas con la diana en los datos de secuenciación. En algunas formas de realización, si el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación es menor que el umbral de pseudopuntos, el método 900 puede proceder al bloqueo si la cantidad de etiquetas moleculares con secuencias distintas asociadas con la diana en la Los datos de secuenciación son inferiores al umbral de pseudopuntos, el método 944.

[0316] En el bloque de decisión 932, si el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación no es menor que el umbral de pseudopuntos, las etiquetas moleculares no únicas se pueden eliminar en el bloque 940. Las etiquetas se pueden eliminar para determinar el número de etiquetas moleculares de ruido con secuencias distintas asociadas con la diana en los datos de secuenciación en el bloque 944. Las etiquetas moleculares no únicas pueden incluir etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación que son mayores que un umbral de etiqueta molecular reciclado predeterminado.

[0317] En el bloque 944, los recuentos de etiquetas moleculares se pueden ajustar usando un método de corrección de errores basado en distribución. El método de corrección de errores basado en distribución puede incluir la determinación del número de etiquetas moleculares de ruido con secuencias distintas asociadas con el objetivo en los datos de secuenciación. La determinación del número de etiquetas moleculares de ruido puede comprender: ajustar dos distribuciones binomiales negativas, una distribución binomial negativa de señal y una distribución binomial negativa de ruido, al número de etiquetas moleculares con secuencias distintas asociadas con el objetivo en los datos de secuenciación. La distribución binomial de señal negativa corresponde a una serie de etiquetas moleculares con secuencias distintas asociadas con el objetivo en los datos de secuenciación contados como etiquetas moleculares de señal. La distribución binomial de ruido negativo corresponde a una serie de etiquetas moleculares con secuencias distintas asociadas con el objetivo en los datos de secuenciación que se cuentan como etiquetas moleculares de ruido. La determinación del número de etiquetas moleculares de ruido puede comprender determinar el número de etiquetas moleculares de ruido utilizando la distribución binomial negativa de señal ajustada y la distribución binomial negativa de ruido ajustada.

[0318] En el bloque 948, se puede estimar el número del objetivo para generar salida después de correcciones de errores basadas en adyacencia y distribución. En el bloque de decisión 912, si el estado de secuenciación del objetivo en los datos de secuenciación es el estado de secuenciación saturado, el método 1000 puede proceder al bloque 948 para

generar resultados sin ajustar etiquetas moleculares basándose en la adyacencia direccional y la corrección de errores basada en la distribución.

[0319] En el bloque de decisión 924, si el estado de secuenciación del objetivo en los datos de secuenciación es el estado de secuenciación insuficiente, el método 1000 puede proceder al bloque 948 para generar resultados sin ajustar las etiquetas moleculares basándose en la corrección de errores basada en la distribución. Por ejemplo, el número de marcas moleculares de ruido determinadas puede ser cero. El método 1000 termina en el bloque 952.

Corrección de errores de secuenciación y PCR según la adyacencia direccional, la corrección de errores basada en la distribución y el submuestreo

[0320] FIG. 11 es un diagrama de flujo que muestra una forma de realización ejemplar 1100 no limitante de corrección de errores usando dos distribuciones binomiales negativas. Los bloques del método 1100 (tales como los bloques 904-952) se han descrito con referencia a la FIG. 9. Brevemente, el método 1100 comienza en el bloque 904 después de recibir datos de secuenciación de una pluralidad de objetivos con códigos de barras estocásticos. En algunas formas de realización, el método 1100 comprende además codificar de barras estocásticamente una pluralidad de objetivos usando una pluralidad de códigos de barras estocásticos para crear la pluralidad de objetivos con códigos de barras estocásticos, en donde cada uno de la pluralidad de códigos de barras estocásticos comprende una etiqueta molecular. En algunas formas de realización, el método 1100 comprende además secuenciar la pluralidad de objetivos con códigos de barras estocásticos para obtener los datos de secuenciación.

[0321] En el bloque 908, para uno o más de la pluralidad de objetivos: se puede contar el número de etiquetas moleculares con secuencias distintas asociadas con el objetivo en los datos de secuenciación. En el bloque de decisión 912, se puede determinar si los datos de secuenciación tienen un estado de secuenciación saturado. Si los datos de secuenciación no tienen un estado de secuenciación saturado en el bloque de decisión 912, el método 1100 puede pasar al bloque 916, donde los recuentos de etiquetas moleculares se pueden ajustar en función de la adyacencia direccional. En algunas formas de realización, el ajuste de los recuentos moleculares en función de la adyacencia direccional puede ser como se describe con referencia a la FIG 7.

[0322] En el bloque 920, se puede determinar el estado de secuenciación del objetivo en los datos de secuenciación. El estado de secuenciación del objetivo en los datos de secuenciación puede incluir, o estar, bajo secuenciación. En el bloque de decisión 924, se puede determinar si el estado de secuenciación del objetivo en los datos de secuenciación es el estado de secuenciación insuficiente.

[0323] En el bloque de decisión 924, si el estado de secuenciación del objetivo en los datos de secuenciación no es el estado de secuenciación inferior, el método 1100 puede opcionalmente proceder al bloque de decisión 1104. En el bloque de decisión 1104, si el estado de secuenciación del objetivo en los datos de secuenciación es el estado de sobresecuenciación que se puede determinar. Por ejemplo, se puede considerar que el objetivo tiene el estado de sobresecuenciación o un objetivo de alta expresión, si su profundidad (por ejemplo, una profundidad promedio, mínima o máxima) es mayor que, o mayor que aproximadamente, 50, 100, 200, 250, 300, 400, 500, 600, 700, 800, 900, 1000, o un número o rango entre dos de estos. Como otro ejemplo, se puede considerar que el objetivo tiene el estado de secuenciación insuficiente si su profundidad es mayor que al menos, o como máximo, 50, 100, 200, 250, 300, 400, 500, 600, 700, 800, 900, o 1000.

[0324] En algunas formas de realización, el estado de sobresecuenciación o el objetivo de expresión alta puede determinarse por el objetivo que tiene una profundidad (por ejemplo, una profundidad promedio, mínima o máxima) mayor que un umbral de sobresecuenciación predeterminado. El umbral de sobresecuenciación puede ser diferente en diferentes implementaciones. Por ejemplo, el umbral de sobresecuenciación puede ser, o ser aproximadamente, 50, 100, 200, 250, 300, 400, 500, 600, 700, 800, 900, 1000, o un número o rango entre dos de estos. Como otro ejemplo, el umbral de sobresecuenciación puede ser al menos, o como máximo, 50, 100, 200, 250, 300, 400, 500, 600, 700, 800, 900, 1000.

[0325] En algunas formas de realización, el estado de sobresecuenciación puede depender del número de etiquetas moleculares de los códigos de barras estocásticos con secuencias distintas. Por ejemplo, el umbral de sobresecuenciación puede ser, o ser aproximadamente, 50, 100, 200, 250, 300, 400, 500, 600, 700, 800, 900, 1000, o un número o un rango entre dos de estos valores si los códigos de barras estocásticos comprenden alrededor de 6561 etiquetas moleculares con secuencias distintas. Como otro ejemplo, el umbral de sobresecuenciación puede ser al menos, o como máximo, 50, 100, 200, 250, 300, 400, 500, 600, 700, 800, 900, 1000, si los códigos de barras estocásticos comprenden aproximadamente 6561 etiquetas moleculares. con secuencias distintas. En algunas formas de realización, el estado de secuenciación insuficiente puede no depender del número de etiquetas moleculares de los códigos de barras estocásticos con secuencias distintas.

[0326] En el bloque de decisión 1104, si el objetivo tiene el estado de sobresecuenciación, el método 1100 continúa con el bloque 1108. En el bloque 1108, la cobertura de EM del objetivo se puede reducir, por ejemplo, submuestreando la cobertura de EM para todos objetivos. Por ejemplo, el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación se puede submuestrear hasta aproximadamente el umbral de

sobresecuenciación predeterminado para todas las dianas (por ejemplo, 10). El método 1100 procede al bloque 928 desde el bloque 1108.

[0327] En el bloque de decisión 1104, si el objetivo no tiene el estado de sobresecuenciación, el método 1100 procede al bloque 928 para filtrar los recuentos de etiquetas moleculares. El filtrado de recuentos de etiquetas moleculares puede incluir, en el bloque de decisión 932, que puede determinarse el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación que son menores que un umbral de pseudopuntos.

[0328] En el bloque de decisión 932, si el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación es menor que el umbral de pseudopuntos, el método 900 puede opcionalmente proceder al bloque 936, donde se pueden agregar pseudopuntos al número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación antes de determinar el número de etiquetas moleculares de ruido con secuencias distintas asociadas con la diana en los datos de secuenciación. En algunas formas de realización, si el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación es menor que el umbral de pseudopuntos, el método 900 puede proceder al bloqueo si la cantidad de etiquetas moleculares con secuencias distintas asociadas con la diana en la Los datos de secuenciación son inferiores al umbral de pseudopuntos, el método 944.

[0329] En el bloque de decisión 932, si el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación no es menor que el umbral de pseudopuntos, las etiquetas moleculares no únicas se pueden eliminar en el bloque 940. Las etiquetas se pueden eliminar para determinar el número de etiquetas moleculares de ruido con secuencias distintas asociadas con la diana en los datos de secuenciación en el bloque 944. Las etiquetas moleculares no únicas pueden incluir etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación que son mayores que un umbral de etiqueta molecular reciclado predeterminado.

[0330] En el bloque 944, los recuentos de etiquetas moleculares se pueden ajustar usando un método de corrección de errores basado en distribución. El método de corrección de errores basado en distribución puede incluir la determinación del número de etiquetas moleculares de ruido con secuencias distintas asociadas con el objetivo en los datos de secuenciación. La determinación del número de etiquetas moleculares de ruido puede comprender: ajustar dos distribuciones binomiales negativas, una distribución binomial negativa de señal y una distribución binomial negativa de ruido, al número de etiquetas moleculares con secuencias distintas asociadas con el objetivo en los datos de secuenciación. La distribución binomial de señal negativa corresponde a una serie de etiquetas moleculares con secuencias distintas asociadas con el objetivo en los datos de secuenciación contados como etiquetas moleculares de señal. La distribución binomial de ruido negativo corresponde a una serie de etiquetas moleculares con secuencias distintas asociadas con el objetivo en los datos de secuenciación que se cuentan como etiquetas moleculares de ruido. La determinación del número de etiquetas moleculares de ruido puede comprender determinar el número de etiquetas moleculares de ruido utilizando la distribución binomial negativa de señal ajustada y la distribución binomial negativa de ruido ajustada.

[0331] Después de ajustar los recuentos de etiquetas moleculares usando un método de corrección de errores basado en distribución en el bloque 944, el método 1100 opcionalmente continúa con el bloque 1112. En el bloque 1112, los recuentos de etiquetas moleculares ajustados del bloque 944 se pueden combinar con recuentos de etiquetas moleculares ajustados basado en la adyacencia direccional determinada en el bloque 916. Por ejemplo, las etiquetas moleculares no únicas se eliminan en el bloque 940 y no se usan para el ajuste de distribución en el bloque 944. Sin embargo, estas etiquetas moleculares todavía están presentes en los recuentos de etiquetas moleculares ajustados en base a la adyacencia direccional determinada en el bloque 916. adyacencia determinada en el bloque 916. En consecuencia, los recuentos de etiquetas moleculares ajustados del bloque 944 y los recuentos de etiquetas moleculares ajustados en el bloque 944 se pueden combinar para generar una salida en el bloque 948.

[0332] En el bloque de decisión 912, si el estado de secuenciación del objetivo en los datos de secuenciación es el estado de secuenciación saturado, el método 1100 puede proceder al bloque 948 para generar resultados sin ajustar las etiquetas moleculares basadas en la adyacencia direccional y la corrección de errores basada en la distribución. En el bloque de decisión 924, si el estado de secuenciación del objetivo en los datos de secuenciación es el estado de secuenciación insuficiente, el método 1100 puede proceder al bloque 948 para generar resultados sin ajustar las etiquetas moleculares basándose en la corrección de errores basada en la distribución. Por ejemplo, el número de marcas moleculares de ruido determinadas puede ser cero. El método 1100 puede, por ejemplo, terminar en el bloque 952.

[0333] FIG. 12 es un diagrama de flujo que muestra una forma de realización ejemplar 1200 no limitante de corrección de errores usando dos distribuciones binomiales negativas. Los bloques del método 1200 (tales como los bloques 904-952 y el bloque 1104) se han descrito con referencia a las FIGS. 9 y 11. Brevemente, el método 1200 comienza en el bloque 904 después de recibir datos de secuenciación de una pluralidad de objetivos con códigos de barras estocásticos. En algunas formas de realización, el método 1200 comprende además codificar de forma estocástica una pluralidad de objetivos usando una pluralidad de códigos de barras estocásticos para crear la pluralidad de objetivos con códigos de barras estocásticos, en donde cada uno de la pluralidad de códigos de barras estocásticos comprende una etiqueta molecular. En algunas formas de realización, el método 1200 comprende además secuenciar la pluralidad de objetivos con códigos de barras estocásticos para obtener los datos de secuenciación.

[0334] En el bloque 908, para uno o más de la pluralidad de objetivos: se puede contar el número de etiquetas moleculares con secuencias distintas asociadas con el objetivo en los datos de secuenciación. En el bloque de decisión 912, se puede determinar si los datos de secuenciación tienen un estado de secuenciación saturado. Si los datos de secuenciación no tienen un estado de secuenciación saturado en el bloque de decisión 912, el método 1200 puede pasar al bloque 916, donde los recuentos de etiquetas moleculares se pueden ajustar en función de la adyacencia direccional. En algunas formas de realización, el ajuste de los recuentos moleculares en función de la adyacencia direccional puede ser como se describe con referencia a la FIG 7.

[0335] En el bloque 920, se puede determinar el estado de secuenciación del objetivo en los datos de secuenciación. El estado de secuenciación del objetivo en los datos de secuenciación puede incluir, o estar, bajo secuenciación. En el bloque de decisión 924, se puede determinar si el estado de secuenciación del objetivo en los datos de secuenciación es el estado de secuenciación insuficiente.

[0336] En el bloque de decisión 924, si el estado de secuenciación del objetivo en los datos de secuenciación no es el estado de secuenciación inferior, el método 1200 puede opcionalmente proceder al bloque de decisión 1104. En el bloque de decisión 1104, si el estado de secuenciación del objetivo en los datos de secuenciación son los que pueden determinarse el estado de sobresecuenciación.

[0337] En el bloque de decisión 1104, si el objetivo tiene el estado de sobresecuenciación o si el objetivo es un objetivo de alta expresión, el método 1200 opcionalmente continúa con el bloque 1208. En el bloque 1208, la cobertura de EM del objetivo se puede reducir en, por ejemplo, submuestrear la cobertura de EM objetivo por objetivo. Por ejemplo, el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación se puede submuestrear hasta aproximadamente el umbral de secuenciación predeterminado diana por diana. El método 1200 procede al bloque 928 desde el bloque 1208.

[0338] En el bloque de decisión 1104, si el objetivo no tiene el estado de sobresecuenciación, el método 1200 procede al bloque 928 para filtrar los recuentos de etiquetas moleculares. El filtrado de recuentos de etiquetas moleculares puede incluir, en el bloque de decisión 932, que puede determinarse el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación que son menores que un umbral de pseudopuntos.

[0339] En el bloque de decisión 932, si el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación es menor que el umbral de pseudopuntos, el método 900 puede opcionalmente proceder al bloque 936, donde opcionalmente se pueden agregar pseudopuntos al número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación antes de determinar el número de etiquetas moleculares de ruido con secuencias distintas asociadas con la diana en los datos de secuenciación. En algunas formas de realización, si el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación es menor que el umbral de pseudopuntos, el método 900 puede proceder al bloqueo si la cantidad de etiquetas moleculares con secuencias distintas asociadas con la diana en la Los datos de secuenciación son inferiores al umbral de pseudopuntos, el método 944.

[0340] En el bloque de decisión 932, si el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación no es menor que el umbral de pseudopuntos, las etiquetas moleculares no únicas se pueden eliminar en el bloque 940. Las etiquetas se pueden eliminar para determinar el número de etiquetas moleculares de ruido con secuencias distintas asociadas con la diana en los datos de secuenciación en el bloque 944. Las etiquetas moleculares no únicas pueden incluir etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación que son mayores que un umbral de etiqueta molecular reciclado predeterminado.

[0341] En el bloque 944, los recuentos de etiquetas moleculares se pueden ajustar usando un método de corrección de errores basado en distribución. El método de corrección de errores basado en distribución puede incluir la determinación del número de etiquetas moleculares de ruido con secuencias distintas asociadas con el objetivo en los datos de secuenciación. La determinación del número de etiquetas moleculares de ruido puede comprender: ajustar dos distribuciones binomiales negativas, una distribución binomial negativa de señal y una distribución binomial negativa de ruido, al número de etiquetas moleculares con secuencias distintas asociadas con el objetivo en los datos de secuenciación. La distribución binomial de señal negativa corresponde a una serie de etiquetas moleculares con secuencias distintas asociadas con el objetivo en los datos de secuenciación contados como etiquetas moleculares de señal. La distribución binomial de ruido negativo corresponde a una serie de etiquetas moleculares con secuencias distintas asociadas con el objetivo en los datos de secuenciación se cuentan como etiquetas moleculares de ruido. La determinación del número de etiquetas moleculares de ruido puede comprender determinar el número de etiquetas moleculares de ruido utilizando la distribución binomial negativa de señal ajustada y la distribución binomial negativa de ruido ajustada.

[0342] Después de ajustar los recuentos de etiquetas moleculares usando un método de corrección de errores basado en distribución en el bloque 944, el método 1200 opcionalmente continúa con el bloque 1112. En el bloque 1112, los recuentos de etiquetas moleculares ajustados del bloque 944 se pueden combinar con recuentos de etiquetas moleculares ajustados basado en la adyacencia direccional determinada en el bloque 916. Por ejemplo, las etiquetas moleculares no

únicas se eliminan en el bloque 940 y no se usan para el ajuste de distribución en el bloque 944. Sin embargo, estas etiquetas moleculares todavía están presentes en los recuentos de etiquetas moleculares ajustados en base a la adyacencia direccional determinada en el bloque 916. la adyacencia se determina en el bloque 916. En consecuencia, los recuentos de etiquetas moleculares ajustados del bloque 944 y los recuentos de etiquetas moleculares ajustados en el bloque 944 se pueden combinar para generar una salida en el bloque 948.

[0343] En el bloque de decisión 912, si el estado de secuenciación del objetivo en los datos de secuenciación es el estado de secuenciación saturado, el método 1200 puede proceder al bloque 948 para generar resultados sin ajustar las etiquetas moleculares basadas en la adyacencia direccional y la corrección de errores basada en la distribución. En el bloque de decisión 924, si el estado de secuenciación del objetivo en los datos de secuenciación es el estado de secuenciación insuficiente, el método 1200 puede proceder al bloque 948 para generar resultados sin ajustar las etiquetas moleculares basándose en la corrección de errores basada en la distribución. Por ejemplo, el número de marcas moleculares de ruido determinadas puede ser cero. El método 1200 termina en el bloque 952.

Corrección de errores de secuenciación y PCR basándose en la adyacencia direccional y la corrección de errores basada en la distribución con estimaciones de parámetros iniciales para el ajuste de la distribución

[0344] FIG. 13 es un diagrama de flujo que muestra una forma de realización ejemplar 13 no limitante de corrección de errores de secuenciación y PCR basándose en la corrección de errores de sustitución recursiva y la corrección de errores basada en distribución por recursividad. Los bloques del método 1300 (tales como los bloques 904-952) se han descrito con referencia a la FIG. 9. Brevemente, el método 1300 comienza en el bloque 904 después de recibir datos de secuenciación de una pluralidad de objetivos con códigos de barras estocásticos. En algunas formas de realización, el método 1300 comprende además codificar de barras estocásticamente una pluralidad de objetivos usando una pluralidad de códigos de barras estocásticos para crear la pluralidad de objetivos con códigos de barras estocásticos, en donde cada uno de la pluralidad de códigos de barras estocásticos comprende una etiqueta molecular. En algunas formas de realización, el método 1300 comprende además secuenciar la pluralidad de objetivos con códigos de barras estocásticos para obtener los datos de secuenciación.

[0345] En el bloque 908, para uno o más de la pluralidad de objetivos: se puede contar el número de etiquetas moleculares con secuencias distintas asociadas con el objetivo en los datos de secuenciación. En el bloque de decisión 912, se puede determinar si los datos de secuenciación tienen un estado de secuenciación saturado. Si los datos de secuenciación no tienen un estado de secuenciación saturado en el bloque de decisión 912, el método 1300 puede pasar al bloque 916, donde los recuentos de etiquetas moleculares se pueden ajustar en función de la adyacencia direccional. En algunas formas de realización, el ajuste de los recuentos moleculares en función de la adyacencia direccional puede ser como se describe con referencia a la FIG 7.

[0346] En el bloque 920, se puede determinar el estado de secuenciación del objetivo en los datos de secuenciación. El estado de secuenciación del objetivo en los datos de secuenciación puede incluir, o estar, bajo secuenciación. En el bloque de decisión 924, se puede determinar si el estado de secuenciación del objetivo en los datos de secuenciación es el estado de secuenciación insuficiente.

[0347] En el bloque de decisión 924, si el estado de secuenciación de la diana en los datos de secuenciación no es el estado de secuenciación insuficiente, el método 1300 puede proceder al bloque 928 para filtrar los recuentos de etiquetas moleculares. El filtrado de recuentos de etiquetas moleculares puede incluir, en el bloque de decisión 932, que puede determinarse el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación que son menores que un umbral de pseudopuntos.

[0348] En el bloque de decisión 932, si el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación es menor que el umbral de pseudopuntos, el método 900 puede opcionalmente proceder al bloque 936, donde se pueden agregar pseudopuntos al número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación antes de determinar el número de etiquetas moleculares de ruido con secuencias distintas asociadas con la diana en los datos de secuenciación. En algunas formas de realización, si el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación es menor que el umbral de pseudopuntos, el método 900 puede proceder al bloqueo si la cantidad de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación son inferiores al umbral de pseudopuntos, el método 944.

[0349] En el bloque de decisión 932, si el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación no es menor que el umbral de pseudopuntos, las etiquetas moleculares no únicas se pueden eliminar en el bloque 940.

[0350] Antes de ajustar los recuentos de etiquetas moleculares en el bloque 944, los parámetros iniciales de las dos distribuciones binomiales negativas se pueden estimar opcionalmente en el bloque 1304. Los parámetros iniciales de las dos distribuciones binomiales negativas pueden ser diferentes en diferentes implementaciones. En algunas formas de realización, la media y la dispersión de cada una de las dos distribuciones binomiales negativas pueden ser una. En algunas formas de realización, la media y la dispersión de las dos distribuciones binomiales negativas se pueden estimar

como la media y la dispersión de un subconjunto no vacío de los recuentos de etiquetas moleculares filtradas del bloque 928. Por ejemplo, el subconjunto puede ser 25 % - 75 % de los cuantiles del marcador molecular filtrado cuenta desde el bloque 928. El rango superior o inferior de los cuantiles puede ser diferente en diferentes implementaciones. En algunas formas de realización, el rango superior o inferior puede ser, o ser aproximadamente, 1 %, 2 %, 3 %, 4 %, 5 %, 6 %, 7 %, 8 %, 9 %, 10 %, 20 %, 30 %, 40 %, 50 %, 70 %, 80 %, 90 %, 99 %, o un número o rango entre dos de estos valores. En algunas formas de realización, el rango superior o inferior puede ser al menos, o como máximo, 1 %, 2 %, 3 %, 4 %, 5 %, 6 %, 7 %, 8 %, 9 %, 10 %, 20 %, 30 %, 40 %, 50 %, 70 %, 80 %, 90 %, 99 % o 100 %.

[0351] En el bloque 944, los recuentos de etiquetas moleculares se pueden ajustar usando un método de corrección de errores basado en distribución. El método de corrección de errores basado en distribución puede incluir la determinación del número de etiquetas moleculares de ruido con secuencias distintas asociadas con el objetivo en los datos de secuenciación. La determinación del número de etiquetas moleculares de ruido puede comprender: ajustar dos distribuciones binomiales negativas, una distribución binomial negativa de señal y una distribución binomial negativa de ruido, al número de etiquetas moleculares con secuencias distintas asociadas con el objetivo en los datos de secuenciación. La distribución binomial de señal negativa corresponde a una serie de etiquetas moleculares con secuencias distintas asociadas con el objetivo en los datos de secuenciación contados como etiquetas moleculares de señal. La distribución binomial de ruido negativo corresponde a una serie de etiquetas moleculares con secuencias distintas asociadas con el objetivo en los datos de secuenciación que se cuentan como etiquetas moleculares de ruido. La determinación del número de etiquetas moleculares de ruido puede comprender determinar el número de etiquetas moleculares de ruido utilizando la distribución binomial negativa de señal ajustada y la distribución binomial negativa de ruido ajustada.

[0352] En el bloque 948, se puede estimar el número del objetivo para generar salida después de correcciones de errores basadas en adyacencia y distribución. En el bloque de decisión 912, si el estado de secuenciación del objetivo en los datos de secuenciación es el estado de secuenciación saturado, el método 1300 puede proceder al bloque 948 para generar resultados sin ajustar etiquetas moleculares basándose en la adyacencia direccional y la corrección de errores basada en la distribución.

[0353] En el bloque de decisión 924, si el estado de secuenciación del objetivo en los datos de secuenciación es el estado de secuenciación insuficiente, el método 1300 puede proceder al bloque 948 para generar resultados sin ajustar las etiquetas moleculares basándose en la corrección de errores basada en la distribución. Por ejemplo, el número de marcas moleculares de ruido determinadas puede ser cero. El método 1300 puede, por ejemplo, terminar en el bloque 952.

[0354] FIG. 14 es un diagrama de flujo que muestra una forma de realización ejemplar no limitante de la corrección de errores de secuenciación y PCR basándose en la corrección de errores de sustitución recursiva y la corrección de errores basada en la distribución utilizando la segunda etiqueta molecular más alta para las estimaciones de parámetros iniciales. Los bloques del método 1400 (tales como los bloques 904-952) se han descrito con referencia a la FIG. 9. Brevemente, el método 1400 comienza en el bloque 904 después de recibir datos de secuenciación de una pluralidad de objetivos con códigos de barras estocásticos. En algunas formas de realización, el método 1400 comprende además codificar de barras estocásticamente una pluralidad de objetivos usando una pluralidad de códigos de barras estocásticos para crear la pluralidad de objetivos con códigos de barras estocásticos, en donde cada uno de la pluralidad de códigos de barras estocásticos comprende una etiqueta molecular. En algunas formas de realización, el método 1400 comprende además secuenciar la pluralidad de objetivos con códigos de barras estocásticos para obtener los datos de secuenciación.

[0355] En el bloque 908, para uno o más de la pluralidad de objetivos: se puede contar el número de etiquetas moleculares con secuencias distintas asociadas con el objetivo en los datos de secuenciación. En el bloque de decisión 912, se puede determinar si los datos de secuenciación tienen un estado de secuenciación saturado. Si los datos de secuenciación no tienen un estado de secuenciación saturado en el bloque de decisión 912, el método 1400 puede pasar al bloque 916, donde los recuentos de etiquetas moleculares se pueden ajustar en función de la adyacencia direccional. En algunas formas de realización, el ajuste de los recuentos moleculares en función de la adyacencia direccional puede ser como se describe con referencia a la FIG 7.

[0356] En el bloque 920, se puede determinar el estado de secuenciación del objetivo en los datos de secuenciación. El estado de secuenciación del objetivo en los datos de secuenciación puede incluir, o estar, bajo secuenciación. En el bloque de decisión 924, se puede determinar si el estado de secuenciación del objetivo en los datos de secuenciación es el estado de secuenciación insuficiente.

[0357] En el bloque de decisión 924, si el estado de secuenciación de la diana en los datos de secuenciación no es el estado de secuenciación insuficiente, el método 1400 puede proceder al bloque 928 para filtrar los recuentos de etiquetas moleculares. El filtrado de recuentos de etiquetas moleculares puede incluir, en el bloque de decisión 932, que puede determinarse el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación que son menores que un umbral de pseudopuntos.

[0358] En el bloque de decisión 932, si el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación es menor que el umbral de pseudopuntos, el método 900 puede opcionalmente

proceder al bloque 936, donde se pueden agregar pseudopuntos al número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación antes de determinar el número de etiquetas moleculares de ruido con secuencias distintas asociadas con la diana en los datos de secuenciación.

[0359] En el bloque de decisión 932, si el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación no es menor que el umbral de pseudopuntos, las etiquetas moleculares no únicas se pueden eliminar en el bloque 940.

[0360] En el bloque 944, los recuentos de etiquetas moleculares se pueden ajustar usando un método de corrección de errores basado en distribución. Los parámetros iniciales para el método de corrección de errores basado en la distribución pueden basarse en el recuento de una etiqueta molecular. Por ejemplo, los parámetros iniciales (tales como la media y la dispersión) para una de las distribuciones binomiales negativas (por ejemplo, la distribución binomial negativa de señal o la distribución binomial negativa de ruido) pueden basarse en el recuento de una etiqueta molecular o la media o promedio o los recuentos de una serie de etiquetas moleculares. Esta etiqueta molecular puede ser la etiqueta molecular con el segundo recuento más alto o una etiqueta molecular con cualquier clasificación (por ejemplo, con el décimo recuento más alto). La clasificación de la etiqueta molecular puede ser diferente en diferentes implementaciones. En algunas formas de realización, la clasificación puede ser, o ser aproximadamente, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, o un número o rango entre dos de estos valores. En algunas formas de realización, la clasificación puede ser al menos, o como máximo, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50, 60, 70, 80, 90., o 100. El número de etiquetas moleculares puede ser diferente en diferentes implementaciones. En algunas formas de realización, el número de etiquetas moleculares puede ser, o ser aproximadamente, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50, 60, 70, 80. 90, 100 o un número o rango entre dos de estos valores. En algunas formas de realización, el número de etiquetas moleculares puede ser al menos, o como máximo, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50, 60, 70. 80, 90 o 100.

[0361] En el bloque 948, se puede estimar el número del objetivo para generar salida después de correcciones de errores basadas en adyacencia y distribución. En el bloque de decisión 912, si el estado de secuenciación del objetivo en los datos de secuenciación es el estado de secuenciación saturado, el método 1400 puede proceder al bloque 948 para generar resultados sin ajustar las etiquetas moleculares basándose en la adyacencia direccional y la corrección de errores basada en la distribución.

[0362] En el bloque de decisión 924, si el estado de secuenciación del objetivo en los datos de secuenciación es el estado de secuenciación insuficiente, el método 1400 puede proceder al bloque 948 para generar resultados sin ajustar las etiquetas moleculares basándose en la corrección de errores basada en la distribución. Por ejemplo, el número de marcas moleculares de ruido determinadas puede ser cero. El método 1400 termina en el bloque 952.

Secuenciación

[0363] En algunas formas de realización, estimar el número de diferentes objetivos con códigos de barras estocásticos puede comprender determinar las secuencias de los objetivos marcados, el marcador espacial, el marcador molecular, el marcador de muestra, el marcador celular o cualquier producto de los mismos (por ejemplo, amplicones marcados), o moléculas de ADNc marcadas). Una diana amplificada puede someterse a secuenciación. La determinación de la secuencia de la diana con código de barras estocástico o cualquier producto del mismo puede comprender realizar una reacción de secuenciación para determinar la secuencia de al menos una porción de una etiqueta de muestra, una etiqueta espacial, una etiqueta celular, una etiqueta molecular, al menos una porción de la diana etiquetada estocásticamente, un complemento del mismo, un complemento inverso del mismo, o cualquier combinación de los mismos.

[0364] Se puede realizar la determinación de la secuencia de una diana con código de barras estocástico (por ejemplo, ácido nucleico amplificado, ácido nucleico marcado, copia de ADNc de un ácido nucleico marcado, etc.) usando una variedad de métodos de secuenciación que incluyen, entre otros, secuenciación por hibridación (SBH), secuenciación por ligación (SBL), secuenciación incremental cuantitativa por adición de nucleótidos fluorescentes (QIFNAS), ligadura y escisión por pasos, transferencia de energía por resonancia de fluorescencia (FRET), balizas moleculares, digestión con sonda indicadora TaqMan, pirosecuenciación, secuenciación fluorescente in situ (FISSEQ), perlas FISSEQ, secuenciación oscilante, secuenciación múltiple, secuenciación de colonias polimerizadas (POLONY); secuenciación de círculo rodante con nanogrid (ROLONY), ensayos de oligoligación específicos de alelo (p. ej., ensayo de oligoligadura (OLA), molécula de plantilla única OLA usando una sonda lineal ligada y una lectura de amplificación de círculo rodante (RCA), sondas de candado ligadas o plantilla única molécula OLA usando una sonda de candado circular ligada y una lectura de amplificación de círculo rodante (RCA), y similares.

[0365] En algunas formas de realización, la determinación de la secuencia del objetivo con código de barras estocástico o cualquier producto del mismo comprende secuenciación de extremos pares, secuenciación de nanoporos, secuenciación de alto rendimiento, secuenciación de escopeta, secuenciación con terminador de colorante, secuenciación de ADN con cebadores múltiples, recorrido con cebador, Secuenciación didesoxi de Sanger, secuenciación Maxim-Gilbert, pirosecuenciación, secuenciación verdadera de molécula única o cualquier combinación de las mismas. Alternativamente, la secuencia del objetivo con código de barras estocástico o cualquier producto del mismo se puede determinar mediante microscopía electrónica o una matriz de transistores de efecto de campo sensibles a productos químicos (chemFET).

[0366] También se pueden utilizar métodos de secuenciación de alto rendimiento, tales como secuenciación de matriz cíclica utilizando plataformas como Roche 454, Illumina Solexa, ABI-SOLiD, ION Torrent, Complete Genomics, Pacific Bioscience, Helicos o la plataforma Polonator. En alguna forma de realización, la secuenciación puede comprender secuenciación MiSeq. En alguna forma de realización, la secuenciación puede comprender secuenciación HiSeq.

[0367] Las dianas marcadas estocásticamente pueden comprender ácidos nucleicos que representan desde aproximadamente el 0,01 % de los genes del genoma de un organismo hasta aproximadamente el 100 % de los genes del genoma de un organismo. Por ejemplo, aproximadamente del 0,01 % de los genes del genoma de un organismo a aproximadamente el 100 % de los genes del genoma de un organismo se pueden secuenciar usando una región complementaria diana que comprende una pluralidad de multímeros capturando los genes que contienen una secuencia complementaria de la muestra. En algunas formas de realización, las dianas con códigos de barras estocásticos comprenden ácidos nucleicos que representan desde aproximadamente el 0,01 % de las transcripciones del transcriptoma de un organismo hasta aproximadamente el 100 % de las transcripciones del transcriptoma de un organismo. Por ejemplo, aproximadamente del 0,501 % de las transcripciones del transcriptoma de un organismo a aproximadamente el 100 % de las transcripciones del transcriptoma de un organismo se pueden secuenciar usando una región complementaria diana que comprende una cola poli(T) capturando los ARNm de la muestra.

[0368] La determinación de las secuencias de las etiquetas espaciales y las etiquetas moleculares de la pluralidad de códigos de barras estocásticos puede incluir la secuenciación 0,00001 %, 0,0001 %, 0,001 %, 0,01 %, 0,1 %, 1 %, 2 %, 3 %, 4 %, 5 %, 6 %, 7 %, 8 %, 9 %, 10 %, 20 %, 30 %, 40 %, 50 %, 60 %, 70 %, 80 %, 90 %, 99 %, 100 %, o un número o un rango entre dos cualesquiera de estos valores, de la pluralidad de códigos de barras estocásticos. Determinación de las secuencias de las etiquetas de la pluralidad de códigos de barras estocásticos, por ejemplo las etiquetas de muestra, las etiquetas espaciales y las etiquetas moleculares, pueden incluir la secuenciación 1, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 10^3, 10^4, 10^5, 10^6, 10^7, 10^8, 10^9, 10^{10}, 10^{11}, 10^{12}, 10^{13}, 10^{14}, 10^{15}, 10^{16}, 10^{17}, 10^{18}, 10^{19}, 10^{20}, o un número o un rango entre dos cualesquiera de estos valores, de la pluralidad de códigos de barras estocásticos. Secuenciar parte o la totalidad de la pluralidad de códigos de barras estocásticos puede incluir generar secuencias con longitudes de lectura de, aproximadamente, al menos o como máximo, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10000, o un número o rango entre dos de estos valores, de nucleótidos o bases.

[0369] La secuenciación puede comprender secuenciar al menos o al menos aproximadamente 10, 20, 30, 40, 50, 60, 70, 80, 90, 100 o más nucleótidos o pares de bases de las dianas con código de barras estocástico. Por ejemplo, la secuenciación puede comprender generar datos de secuenciación con secuencias con longitudes de lectura de 50, 75 o 100 o más nucleótidos realizando una amplificación por reacción en hebra de la polimerasa (PCR) en la pluralidad de objetivos con códigos de barras estocásticos. La secuenciación puede comprender secuenciar al menos o al menos aproximadamente 200, 300, 400, 500, 600, 700, 800, 900, 1.000 o más nucleótidos o pares de bases de las dianas con código de barras estocástico. La secuenciación puede comprender secuenciar al menos o al menos aproximadamente 1500, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000 o 10000 o más nucleótidos o pares de bases de las dianas con códigos de barras estocásticos.

[0370] La secuenciación puede comprender al menos aproximadamente 200, 300, 400, 500, 600, 700, 800, 900, 1000 o más lecturas de secuenciación por ejecución. En algunas formas de realización, la secuenciación comprende secuenciar al menos o al menos aproximadamente 1500, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000 o 10000 o más lecturas de secuenciación por ejecución. La secuenciación puede comprender menos o igual a aproximadamente 1.600.000.000 lecturas de secuenciación por ejecución. La secuenciación puede comprender menos o igual a aproximadamente 200.000.000 de lecturas por ejecución.

Muestras

[0371] La pluralidad de objetivos está comprendida en una o más muestras. Una muestra comprende una o más células, o ácidos nucleicos de una o más células. Una muestra puede ser una sola célula o ácidos nucleicos de una sola célula. La una o más células pueden ser de uno o más tipos de células. Al menos uno de uno o más tipos de células puede ser una célula cerebral, una célula cardíaca, una célula cancerosa, una célula tumoral circulante, una célula orgánica, una célula epitelial, una célula metastásica, una célula benigna, una célula primaria, una célula circulatoria o cualquier combinación de las mismas.

[0372] Una muestra para usar en el método de la divulgación comprende una o más células. Una muestra puede hacer referencia a una o más células. En algunas formas de realización, la pluralidad de células puede incluir uno o más tipos de células. Al menos uno de uno o más tipos de células puede ser una célula cerebral, una célula cardíaca, una célula cancerosa, una célula tumoral circulante, una célula orgánica, una célula epitelial, una célula metastásica, una célula benigna, una célula primaria, una célula circulatoria o cualquier combinación de las mismas. En algunas formas de realización, las células son células cancerosas extirpadas de un tejido canceroso, por ejemplo, cáncer de mama, cáncer de pulmón, cáncer de colon, cáncer de próstata, cáncer de ovario, cáncer de páncreas, cáncer de cerebro, melanoma y cánceres de piel no melanoma, y similares. En algunas formas de realización, las células se derivan de un cáncer pero se recolectan de un fluido corporal (por ejemplo, células tumorales circulantes). Los ejemplos no limitantes de cánceres pueden incluir adenoma, adenocarcinoma, carcinoma de células escamosas, carcinoma de células basales, carcinoma

de células pequeñas, carcinoma indiferenciado de células grandes, condrosarcoma y fibrosarcoma. La muestra puede incluir un tejido, una monocapa celular, células fijadas, una sección de tejido o cualquier combinación de los mismos. La muestra puede incluir una muestra biológica, una muestra clínica, una muestra ambiental, un fluido biológico, un tejido o una célula de un sujeto. La muestra se puede obtener de un ser humano, un mamífero, un perro, una rata, un ratón, un pez, una mosca, un gusano, una planta, un hongo, una bacteria, un virus, un vertebrado o un invertebrado.

[0373] En algunas formas de realización, las células son células que han sido infectadas con virus y contienen oligonucleótidos virales. En algunas formas de realización, la infección viral puede ser causada por un virus tal como virus de ADN monocatenario (de hebra + o "sentido") (por ejemplo, parvovirus) o virus de ARN bicatenario (por ejemplo, reovirus). En algunas formas de realización, las células son bacterias. Estos pueden incluir bacterias grampositivas o gramnegativas. En algunas formas de realización, las células son hongos. En algunas formas de realización, las células son protozoos u otros parásitos.

[0374] Como se usa en el presente documento, el término "célula" puede referirse a una o más células. En algunas formas de realización, las células son células normales, por ejemplo, células humanas en diferentes etapas de desarrollo, o células humanas de diferentes órganos o tipos de tejido. En algunas formas de realización, las células son células no humanas, por ejemplo, otros tipos de células de mamíferos (por ejemplo, ratón, rata, cerdo, perro, vaca o caballo). En algunas formas de realización, las células son otros tipos de células animales o vegetales. En otras formas de realización, las células pueden ser cualquier célula procariótica o eucariota.

[0375] En algunas formas de realización, las células se clasifican antes de asociar una célula con una perla. Por ejemplo, las células se pueden clasificar mediante clasificación de células activada por fluorescencia o clasificación de células activada magnéticamente, o más generalmente mediante citometría de flujo. Las células se pueden filtrar por tamaño. En algunas formas de realización, un retenido contiene las células que se asociarán con la perla. En algunas formas de realización, el flujo contiene las células que se asociarán con la perla.

[0376] Una muestra puede hacer referencia a una pluralidad de células. La muestra puede referirse a una monocapa de células. La muestra puede referirse a una sección delgada (por ejemplo, sección delgada de tejido). La muestra puede referirse a una colección de células sólidas o semisólidas que se pueden colocar en una dimensión en una matriz.

Software de análisis y visualización de datos

Análisis de datos y visualización de resolución espacial de objetivos.

[0377] La divulgación proporciona métodos para estimar el número y la posición de objetivos con códigos de barras estocásticos y recuento digital utilizando etiquetas espaciales. Los datos obtenidos de los métodos de la divulgación se pueden visualizar en un mapa. Se puede construir un mapa del número y ubicación de objetivos de una muestra utilizando información generada utilizando los métodos descritos en este documento. El mapa se puede utilizar para localizar la ubicación física de un objetivo. El mapa se puede utilizar para identificar la ubicación de múltiples objetivos. Los objetivos múltiples pueden ser la misma especie de objetivo, o los objetivos múltiples pueden ser múltiples objetivos diferentes. Por ejemplo, se puede construir un mapa de un cerebro para mostrar el recuento digital y la ubicación de múltiples objetivos.

[0378] El mapa se puede generar a partir de datos de una única muestra. El mapa se puede construir utilizando datos de múltiples muestras, generando así un mapa combinado. El mapa se puede construir con datos de decenas, cientos y/o miles de muestras. Un mapa construido a partir de múltiples muestras puede mostrar una distribución de recuentos digitales de objetivos asociados con regiones comunes a las múltiples muestras. Por ejemplo, los ensayos replicados se pueden mostrar en el mismo mapa. Se pueden mostrar (por ejemplo, superpuestas) al menos 1, 2, 3, 4, 5, 6, 7, 8, 9 o 10 o más réplicas en el mismo mapa. Como máximo se pueden mostrar (por ejemplo, superpuestas) 1, 2, 3, 4, 5, 6, 7, 8, 9 o 10 o más réplicas en el mismo mapa. La distribución espacial y el número de objetivos se pueden representar mediante una variedad de estadísticas.

[0379] La combinación de datos de múltiples muestras puede aumentar la resolución de ubicación del mapa combinado. La orientación de múltiples muestras puede registrarse mediante puntos de referencia comunes, donde las mediciones de ubicación individuales entre muestras son al menos en parte no contiguas. Un ejemplo particular es seccionar una muestra usando un microtomo en un eje y luego seccionar una segunda muestra a lo largo de un acceso diferente. El conjunto de datos combinado proporcionará ubicaciones espaciales tridimensionales asociadas con recuentos digitales de objetivos. La multiplexación del enfoque anterior permitirá obtener mapas tridimensionales de alta resolución de estadísticas de conteo digital.

[0380] En algunas formas de realización del sistema de instrumentos, el sistema comprenderá medios legibles por computadora que incluyen código para proporcionar análisis de datos para los conjuntos de datos de secuencia generados al realizar ensayos de códigos de barras estocásticos de una sola célula. Ejemplos de funcionalidad de análisis de datos que puede proporcionar el software de análisis de datos incluyen, entre otros, (i) algoritmos para decodificar/demultiplexar la etiqueta de muestra, la etiqueta celular, la etiqueta espacial y la etiqueta molecular, y los datos de secuencia diana proporcionados secuenciando la biblioteca de códigos de barras estocásticos creada al ejecutar el ensayo, (ii) algoritmos para determinar el número de lecturas por gen por célula y el número de moléculas de transcripción únicas por gen por

célula, en función de los datos, y creando tablas de resumen, (iii) análisis estadístico de los datos de secuencia, por ejemplo para agrupamiento de células mediante datos de expresión génica, o para predecir intervalos de confianza para determinaciones del número de moléculas de transcripción por gen por célula, etc., (iv) algoritmos para identificar subpoblaciones de células raras, por ejemplo, utilizando análisis de componentes principales, agrupamiento jerárquico, agrupamiento k-media, mapas autoorganizados, redes neuronales, etc., (v) capacidades de alineación de secuencias para la alineación de datos de secuencias genéticas con secuencias de referencia conocidas y detección de mutaciones, marcadores polimórficos y variantes de empalme, y (vi) agrupación automatizada de etiquetas moleculares para compensar errores de amplificación o secuenciación. En algunas formas de realización, se puede usar software disponible comercialmente para realizar todo o una parte del análisis de datos, por ejemplo, se puede usar el software Seven Bridges (https://www.sbgenomics.com/) para compilar tablas del número de copias de uno o más genes que ocurren en cada célula para toda la colección de células. En algunas formas de realización, el software de análisis de datos puede incluir opciones para generar los resultados de la secuenciación en formatos gráficos útiles, por ejemplo, mapas de calor que indican el número de copias de uno o más genes que ocurren en cada célula de una colección de células. En algunas formas de realización, el software de análisis de datos puede comprender además algoritmos para extraer significado biológico de los resultados de la secuenciación, por ejemplo, correlacionando el número de copias de uno o más genes que ocurren en cada célula de una colección de células con un tipo de célula rara, o una célula derivada de un sujeto que tiene una enfermedad o condición específica. En alguna forma de realización, el software de análisis de datos puede comprender además algoritmos para comparar poblaciones de células en diferentes muestras biológicas.

[0381] En algunas formas de realización, toda la funcionalidad de análisis de datos se puede empaquetar dentro de un único paquete de software. En algunas formas de realización, el conjunto completo de capacidades de análisis de datos puede comprender un conjunto de paquetes de software. En algunas formas de realización, el software de análisis de datos puede ser un paquete independiente que se pone a disposición de los usuarios independientemente del sistema de instrumentos de ensayo. En algunas formas de realización, el software puede estar basado en web y puede permitir a los usuarios compartir datos.

[0382] En algunas formas de realización, toda la funcionalidad de análisis de datos se puede empaquetar dentro de un único paquete de software. En algunas formas de realización, el conjunto completo de capacidades de análisis de datos puede comprender un conjunto de paquetes de software. En algunas formas de realización, el software de análisis de datos puede ser un paquete independiente que se pone a disposición de los usuarios independientemente del sistema de instrumentos de ensayo. En algunas formas de realización, el software puede estar basado en web y puede permitir a los usuarios compartir datos.

Procesadores y redes del sistema

[0383] En general, la computadora o procesador adecuado para su uso en los métodos de los sistemas de instrumentos actualmente divulgados, como se ilustra en la FIG. 15, puede entenderse además como un aparato lógico que puede leer instrucciones de los medios 1511 o un puerto de red 1505, que puede conectarse opcionalmente al servidor 1509 que tiene medios fijos 1512. El sistema 1500, como se muestra en la FIG. 15 puede incluir una CPU 1501, unidades de disco 1503, dispositivos de entrada opcionales tales como teclado 1515 o mouse 1516 y monitor opcional 1507. La comunicación de datos se puede lograr a través del medio de comunicación indicado a un servidor en una ubicación local o remota. El medio de comunicación puede incluir cualquier medio de transmisión o recepción de datos. Por ejemplo, el medio de comunicación puede ser una conexión de red, una conexión inalámbrica o una conexión a Internet. Una conexión de este tipo puede permitir la comunicación a través de la World Wide Web. Se prevé que los datos relacionados con la presente divulgación puedan transmitirse a través de tales redes o conexiones para su recepción o revisión por parte de una parte 1522 como se ilustra en la FIG. 15.

[0384] FIG. 16 ilustra una forma de realización ejemplar de una primera arquitectura de ejemplo de un sistema informático 1600 que se puede usar en conexión con formas de realización ejemplares de la presente divulgación. Como se representa en la FIG. 16, el sistema informático de ejemplo puede incluir un procesador 1602 para procesar instrucciones. Ejemplos no limitativos de procesadores incluyen: procesador Intel XeonTM, procesador AMD Opteron™, procesador Samsung RISC ARM 1176JZ(F)-S v1.0™ de 32 bits, procesador ARM Cortex-A8 Samsung S5PC100TM, procesador ARM Cortex-A8 Apple A4TM, procesador Marvell PXA 930TM o un procesador funcionalmente equivalente. Se pueden utilizar varios subprocesos de ejecución para el procesamiento paralelo. En algunas formas de realización, también se pueden usar múltiples procesadores o procesadores con múltiples núcleos, ya sea en un único sistema informático, en un clúster o distribuidos entre sistemas a través de una red que comprende una pluralidad de computadoras, teléfonos celulares o dispositivos de asistente de datos personales.

[0385] Como se ilustra en la FIG. 16, se puede conectar o incorporar una memoria caché de alta velocidad 1604 al procesador 1602 para proporcionar una memoria de alta velocidad para instrucciones o datos que han sido utilizados recientemente, o que son utilizados con frecuencia, por el procesador 1602. El procesador 1602 está conectado a un puente norte 1606 mediante un bus de procesador 1608. El puente norte 1606 está conectado a la memoria de acceso aleatorio (RAM) 1610 mediante un bus de memoria 1612 y gestiona el acceso a la RAM 1610 mediante el procesador 1602. El puente norte 1606 también está conectado a un bus de procesador 1608. puente 1614 mediante un bus de chipset 1616. El puente sur 1614 está, a su vez, conectado a un bus periférico 1618. El bus periférico puede ser, por ejemplo, PCI, PCI-X, PCI Express u otro bus periférico. El puente norte y el puente sur a menudo se denominan conjunto

de chips de procesador y gestionan la transferencia de datos entre el procesador, la RAM y los componentes periféricos en el bus periférico 1618. En algunas arquitecturas alternativas, la funcionalidad del puente norte se puede incorporar al procesador. en lugar de utilizar un chip de puente norte separado.

[0386] En algunas formas de realización, el sistema 1600 puede incluir una tarjeta aceleradora 1622 unida al bus periférico 1618. El acelerador puede incluir matrices de puertas programables en campo (FPGA) u otro hardware para acelerar cierto procesamiento. Por ejemplo, se puede utilizar un acelerador para la reestructuración adaptativa de datos o para evaluar expresiones algebraicas utilizadas en el procesamiento de conjuntos extendidos.

[0387] El software y los datos se almacenan en el almacenamiento externo 1624 y se pueden cargar en la RAM 1610 o en la memoria caché 1604 para su uso por parte del procesador. El sistema 1600 incluye un sistema operativo para gestionar los recursos del sistema; los ejemplos no limitantes de sistemas operativos incluyen: Linux, Windows™, MACOS™, BlackBerry OS™, iOS™ y otros sistemas operativos funcionalmente equivalentes, así como software de aplicación que se ejecuta sobre el sistema operativo para gestionar el almacenamiento y la optimización de datos de acuerdo con formas de realización de ejemplo de la presente invención.

[0388] En este ejemplo, el sistema 1600 también incluye tarjetas de interfaz de red (NIC) 1620 y 1621 conectadas al bus periférico para proporcionar interfaces de red para almacenamiento externo, tal como almacenamiento conectado a la red (NAS) y otros sistemas informáticos que pueden usarse para procesamiento paralelo distribuido.

[0389] FIG. 17 ilustra un diagrama ejemplar que muestra una red 1700 con una pluralidad de sistemas informáticos 1702a y 1702b, una pluralidad de teléfonos móviles y asistentes de datos personales 1702c y un almacenamiento conectado a la red (NAS) 1704a y 1704b adecuado para su uso en los métodos de la divulgación. En formas de realización de ejemplo, los sistemas 1712a, 1712b y 1712c pueden gestionar el almacenamiento de datos y optimizar el acceso a los datos almacenados en el almacenamiento conectado a la red (NAS) 1714a y 1714b. Se puede usar un modelo matemático para los datos y evaluarlo usando procesamiento paralelo distribuido entre los sistemas informáticos 1712a y 1712b, y los sistemas de asistente de datos personales y de teléfonos móviles 1712c. Los sistemas informáticos 1712a y 1712b, y los sistemas de asistente de datos personales y de teléfonos móviles 1712c también pueden proporcionar procesamiento paralelo para la reestructuración adaptativa de los datos almacenados en el almacenamiento conectado a la red (NAS) 1714a y 1714b. FIG. 17 ilustra sólo un ejemplo, y se puede utilizar una amplia variedad de otras arquitecturas y sistemas informáticos junto con las diversas formas de realización de la presente invención. Por ejemplo, se puede utilizar un servidor Blade para proporcionar procesamiento paralelo. Las hojas del procesador se pueden conectar a través de un plano posterior para proporcionar procesamiento en paralelo. El almacenamiento también se puede conectar al plano posterior o como almacenamiento conectado a la red (NAS) a través de una interfaz de red separada.

[0390] En algunas formas de realización de ejemplo, los procesadores pueden mantener espacios de memoria separados y transmitir datos a través de interfaces de red, plano posterior u otros conectores para el procesamiento en paralelo por parte de otros procesadores. En otras formas de realización, algunos o todos los procesadores pueden usar un espacio de memoria de direcciones virtuales compartido.

[0391] FIG. 18 ilustra un diagrama de bloques ejemplar de un sistema informático multiprocesador 1800 que utiliza un espacio de memoria de direcciones virtuales compartido de acuerdo con una forma de realización ejemplar. El sistema incluye una pluralidad de procesadores 1802a-f que pueden acceder a un subsistema de memoria compartida 1804. El sistema incorpora una pluralidad de procesadores de algoritmos de memoria de hardware (MAP) programables 1806a-f en el subsistema de memoria 1804. Cada MAP 1806a-f puede comprender un memoria 1808a-f y una o más matrices de puertas programables en campo (FPGA) 1810a-f. El MAP proporciona una unidad funcional configurable y se pueden proporcionar algoritmos particulares o partes de algoritmos a los FPGA 1810a-f para su procesamiento en estrecha coordinación con un procesador respectivo. Por ejemplo, los MAP se pueden usar para evaluar expresiones algebraicas con respecto al modelo de datos y para realizar una reestructuración adaptativa de datos en formas de realización de ejemplo. En este ejemplo, todos los procesadores pueden acceder globalmente a cada MAP para estos fines. En una configuración, cada MAP puede usar el acceso directo a la memoria (DMA) para acceder a una memoria asociada 1808a-f, lo que le permite ejecutar tareas independientemente, y de forma asincrónica, del respectivo microprocesador 1802a-f. En esta configuración, un MAP puede enviar resultados directamente a otro MAP para la canalización y ejecución paralela de algoritmos.

[0392] Las arquitecturas y sistemas informáticos anteriores son sólo ejemplos, y se puede utilizar una amplia variedad de otras arquitecturas y sistemas de ordenadores, teléfonos móviles y asistentes de datos personales en conexión con formas de realización de ejemplo, incluidos sistemas que utilizan cualquier combinación de procesadores generales, coprocesadores, FPGA y otros dispositivos lógicos programables, sistemas en chips (SOC), circuitos integrados de aplicaciones específicas (ASIC) y otros elementos lógicos y de procesamiento. En algunas formas de realización, todo o parte del sistema informático se puede implementar en software o hardware. Se puede utilizar cualquier variedad de medios de almacenamiento de datos en relación con formas de realización de ejemplo, incluyendo memoria de acceso aleatorio, discos duros, memoria flash, unidades de cinta, matrices de discos, almacenamiento conectado a la red (NAS) y otros dispositivos y sistemas de almacenamiento de datos locales o distribuidos.

[0393] En formas de realización de ejemplo, el subsistema informático de la presente divulgación se puede implementar utilizando módulos de software que se ejecutan en cualquiera de los sistemas y arquitecturas informáticas anteriores u otras. En otras formas de realización, las funciones del sistema se pueden implementar parcial o completamente en firmware, dispositivos lógicos programables tales como conjuntos de puertas programables en campo (FPGA), sistemas en chips (SOL), circuitos integrados de aplicación específica (ASIC) u otros procesos y elementos lógicos. Por ejemplo, el conjunto de procesador y optimizador se puede implementar con aceleración de hardware mediante el uso de una tarjeta aceleradora de hardware, tal como una tarjeta aceleradora.

Procesadores y redes del sistema

[0394] En general, la computadora o procesador incluido en los sistemas de instrumentos actualmente divulgados, como se ilustra en la FIG. , puede entenderse además como un aparato lógico que puede leer instrucciones de los medios 11 o un puerto de red 05, que puede conectarse opcionalmente al servidor 09 que tiene medios fijos 12. El sistema 00, como se muestra en la FIG. puede incluir una CPU 01, unidades de disco 03, dispositivos de entrada opcionales tales como teclado 15 o ratón 16 y monitor opcional 07. La comunicación de datos se puede lograr a través del medio de comunicación indicado a un servidor en una ubicación local o remota. El medio de comunicación puede incluir cualquier medio de transmisión o recepción de datos. Por ejemplo, el medio de comunicación puede ser una conexión de red, una conexión inalámbrica o una conexión a Internet. Una conexión de este tipo puede permitir la comunicación a través de la World Wide Web. Se prevé que los datos relacionados con la presente divulgación puedan transmitirse a través de tales redes o conexiones para su recepción o revisión por parte de una parte 22 como se ilustra en la FIG..

[0395] FIG. ilustra una forma de realización ejemplar de una primera arquitectura de ejemplo de un sistema informático 00 que se puede utilizar en conexión con formas de realización ejemplares de la presente divulgación. Como se representa en la FIG. , el sistema informático de ejemplo puede incluir un procesador 02 para procesar instrucciones. Ejemplos no limitativos de procesadores incluyen: procesador Intel XeonTM, procesador AMD Opteron™, procesador Samsung RISC ARM 1176JZ(F)-S v1.0™ de 32 bits, procesador ARM Cortex-A8 Samsung S5PC100TM, procesador ARM Cortex-A8 Apple A4TM, procesador Marvell PXA 930TM o un procesador funcionalmente equivalente. Se pueden utilizar varios subprocesos de ejecución para el procesamiento paralelo. En algunas formas de realización, también se pueden usar múltiples procesadores o procesadores con múltiples núcleos, ya sea en un único sistema informático, en un clúster o distribuidos entre sistemas a través de una red que comprende una pluralidad de computadoras, teléfonos celulares o dispositivos de asistente de datos personales.

[0396] Como se ilustra en la FIG., se puede conectar o incorporar una memoria caché de alta velocidad 04 al procesador 02 para proporcionar una memoria de alta velocidad para instrucciones o datos que han sido utilizados recientemente, o que son utilizados con frecuencia, por el procesador 02. El procesador 02 está conectado a un norte puente 06 mediante un bus de procesador 08. El puente norte 06 está conectado a la memoria de acceso aleatorio (RAM) 10 mediante un bus de memoria 12 y gestiona el acceso a la RAM 10 por parte del procesador 02. El puente norte 06 también está conectado a un puente sur 14 mediante un bus de chipset 16. El puente sur 14 está, a su vez, conectado a un bus periférico 18. El bus periférico puede ser, por ejemplo, PCI, PCI-X, PCI Express u otro bus periférico. El puente norte y el puente sur a menudo se denominan conjunto de chips de procesador y gestionan la transferencia de datos entre el procesador, la RAM y los componentes periféricos en el bus periférico 18. En algunas arquitecturas alternativas, la funcionalidad del puente norte se puede incorporar al procesador. en lugar de utilizar un chip de puente norte separado.

[0397] En algunas formas de realización, el sistema 00 puede incluir una tarjeta aceleradora 22 conectada al bus periférico 18. El acelerador puede incluir matrices de puertas programables en campo (FPGA) u otro hardware para acelerar cierto procesamiento. Por ejemplo, se puede utilizar un acelerador para la reestructuración adaptativa de datos o para evaluar expresiones algebraicas utilizadas en el procesamiento de conjuntos extendidos.

[0398] El software y los datos se almacenan en el almacenamiento externo 24 y se pueden cargar en la RAM 10 o en la memoria caché 04 para su uso por parte del procesador. El sistema 00 incluye un sistema operativo para gestionar los recursos del sistema; los ejemplos no limitantes de sistemas operativos incluyen: Linux, Windows™, MACOS™, BlackBerry OS™, iOS™ y otros sistemas operativos funcionalmente equivalentes, así como software de aplicación que se ejecuta sobre el sistema operativo para gestionar el almacenamiento y la optimización de datos de acuerdo con formas de realización de ejemplo de la presente invención.

[0399] En este ejemplo, el sistema 00 también incluye tarjetas de interfaz de red (NIC) 20 y 21 conectadas al bus periférico para proporcionar interfaces de red para almacenamiento externo, como almacenamiento conectado a la red (NAS) y otros sistemas informáticos que se pueden usar para procesamiento paralelo distribuido.

[0400] FIG. ilustra un diagrama ejemplar que muestra una red 00 con una pluralidad de sistemas informáticos 02a y 02b, una pluralidad de teléfonos móviles y asistentes de datos personales 02c y un almacenamiento conectado a la red (NAS) 04a y 04b. En formas de realización de ejemplo, los sistemas 12a, 12b y 12c pueden gestionar el almacenamiento de datos y optimizar el acceso a los datos almacenados en el almacenamiento conectado a la red (NAS) 14a y 14b. Se puede utilizar un modelo matemático para los datos y evaluarlo utilizando procesamiento paralelo distribuido entre los sistemas informáticos 12a y 12b, y los sistemas de asistente de datos personales y de teléfonos móviles 12c. Los sistemas informáticos 12a y 12b, y los sistemas de asistente de datos personales y de teléfonos móviles 12c también pueden

proporcionar procesamiento paralelo para la reestructuración adaptativa de los datos almacenados en el almacenamiento conectado a la red (NAS) 14a y 14b. FIG. ilustra sólo un ejemplo, y se puede utilizar una amplia variedad de otras arquitecturas y sistemas informáticos junto con las diversas formas de realización de la presente invención. Por ejemplo, se puede utilizar un servidor Blade para proporcionar procesamiento paralelo. Las hojas del procesador se pueden conectar a través de un plano posterior para proporcionar procesamiento en paralelo. El almacenamiento también se puede conectar al plano posterior o como almacenamiento conectado a la red (NAS) a través de una interfaz de red separada.

[0401] En algunas formas de realización de ejemplo, los procesadores pueden mantener espacios de memoria separados y transmitir datos a través de interfaces de red, plano posterior u otros conectores para el procesamiento en paralelo por parte de otros procesadores. En otras formas de realización, algunos o todos los procesadores pueden usar un espacio de memoria de direcciones virtuales compartido.

[0402] FIG. ilustra un diagrama de bloques ejemplar de un sistema informático multiprocesador 00 que utiliza un espacio de memoria de direcciones virtuales compartido de acuerdo con una forma de realización ejemplar. El sistema incluye una pluralidad de procesadores 02af que pueden acceder a un subsistema de memoria compartida 04. El sistema incorpora una pluralidad de procesadores de algoritmos de memoria de hardware (MAP) programables 06a-f en el subsistema de memoria 04. Cada MAP 06a-f puede comprender una memoria 08a -f y una o más matrices de puertas programables en campo (FPGA) 10a-f. El MAP proporciona una unidad funcional configurable y se pueden proporcionar algoritmos particulares o partes de algoritmos a los FPGA 10a-f para su procesamiento en estrecha coordinación con un procesador respectivo. Por ejemplo, los MAP se pueden usar para evaluar expresiones algebraicas con respecto al modelo de datos y para realizar una reestructuración adaptativa de datos en formas de realización de ejemplo. En este ejemplo, todos los procesadores pueden acceder globalmente a cada MAP para estos fines. En una configuración, cada MAP puede usar el acceso directo a la memoria (DMA) para acceder a una memoria asociada 08a-f, lo que le permite ejecutar tareas de forma independiente y asíncrona desde el microprocesador respectivo 02a-f. En esta configuración, un MAP puede enviar resultados directamente a otro MAP para la canalización y ejecución paralela de algoritmos.

[0403] Las arquitecturas y sistemas de computadora anteriores son solo ejemplos, y se puede usar una amplia variedad de otras arquitecturas y sistemas de computadora, teléfono celular y asistente de datos personales en conexión con formas de realización de ejemplo, incluidos sistemas que usan cualquier combinación de procesadores generales, coprocesadores, FPGA y otros dispositivos lógicos programables, sistemas en chips (SOC), circuitos integrados de aplicaciones específicas (ASIC) y otros elementos lógicos y de procesamiento. En algunas formas de realización, todo o parte del sistema informático se puede implementar en software o hardware. Se puede utilizar cualquier variedad de medios de almacenamiento de datos en relación con formas de realización de ejemplo, incluyendo memoria de acceso aleatorio, discos duros, memoria flash, unidades de cinta, conjuntos de discos, almacenamiento conectado a la red (NAS) y otros dispositivos y sistemas de almacenamiento de datos locales o distribuidos.

[0404] En formas de realización de ejemplo, el subsistema informático de la presente divulgación se puede implementar utilizando módulos de software que se ejecutan en cualquiera de los sistemas y arquitecturas informáticas anteriores u otras. En otras formas de realización, las funciones del sistema se pueden implementar parcial o completamente en firmware, dispositivos lógicos programables tales como conjuntos de puertas programables en campo (FPGA), sistemas en chips (SOL), circuitos integrados de aplicación específica (ASIC) u otros procesos y elementos lógicos. Por ejemplo, el conjunto de procesador y optimizador se puede implementar con aceleración de hardware mediante el uso de una tarjeta aceleradora de hardware, tal como una tarjeta aceleradora.

EJEMPLOS

Ejemplo 1

Corrección de errores de sustitución de una base

[0405] Este ejemplo demuestra la corrección de errores de secuenciación o PCR que implican sustituciones de una base. Los errores de PCR o secuenciación que implicaban sustituciones de una base se eliminaron atribuyendo copias de un objetivo con etiquetas moleculares similares y con ocurrencias, es decir, lecturas de secuenciación, ≤ 7 si había 3^8 códigos de barras estocásticos únicos (17 si había 48 códigos de barras estocásticos únicos) que tenían el mismo etiqueta molecular para la pluralidad de objetivos.

[0406] Los códigos de barras estocásticos pueden comprender el uso de un conjunto que no se agota de 3^8 (6561) códigos de barras estocásticos únicos con oligo(dT) como sus regiones de unión objetivo para etiquetar ARNm con poli(A) en una muestra antes del paso de TI. El proceso de etiquetado puede ser aleatorio y cada molécula objetivo puede hibridarse con un código de barras estocástico. Para cualquier objetivo, si el número de moléculas objetivo fuera mucho menor que el número de códigos de barras estocásticos, es probable que cada molécula objetivo se hibride con un código de barras estocástico diferente. Entonces, si solo estuvieran presentes unas pocas moléculas objetivo, es poco probable que las pocas moléculas objetivo se hibriden con códigos de barras estocásticos con etiquetas moleculares (EM) similares durante la hibridación.

[0407] Se calculó la probabilidad de muestrear al menos un par de códigos de barras estocásticos con etiquetas moleculares similares de 3^8 códigos de barras estocásticos únicos que no se agotan. Das etiquetas moleculares pueden tener secuencias similares si difieren en una base. Este evento de muestreo puede considerarse como muestreo con reemplazo porque los códigos de barras estocásticos pueden prácticamente no agotarse. Esta probabilidad puede ayudar a eliminar códigos de barras estocásticos con etiquetas moleculares similares que es muy poco probable que estén presentes en una muestra determinada que comprende una pluralidad de objetivos. El problema se puede formular como el número de códigos de barras estocásticos necesarios para que se puedan elegir al menos dos códigos de barras estocásticos con etiquetas moleculares similares con una cierta probabilidad. Este problema se puede formular como el tamaño de muestra mínimo requerido para que la probabilidad de que dos códigos de barras estocásticos tengan secuencias similares sea mayor que 0,5 3^8 etiquetas moleculares distintas dadas. Por tanto, este problema puede considerarse como una generalización del problema clásico del cumpleaños. El problema clásico del cumpleaños puede determinar el tamaño de muestra mínimo requerido para que la probabilidad de que dos personas compartan el mismo cumpleaños sea mayor que 0,5 dados 365 cumpleaños distintos.

[0408] Para derivar este tamaño de muestra *r,* se calculó la probabilidad de tener al menos un par de etiquetas moleculares similares dados r códigos de barras estocásticos muestreados de 38 códigos de barras estocásticos únicos usando la probabilidad de su evento complementario. Si solo se eligió aleatoriamente un código de barras estocástico de 3^8 códigos de barras estocásticos únicos, entonces la probabilidad de que su etiqueta molecular no sea similar a las etiquetas moleculares de otros códigos de barras estocásticos, p_1, = 1, porque solo había un código de barras estocástico. Si también se eligió aleatoriamente un segundo código de barras estocástico entre 3^8 códigos de barras estocásticos únicos, entonces la probabilidad de que su etiqueta molecular no sea similar a la etiqueta molecular del primer código de barras estocástico, p_2, = $(3^8 - 16 - 1)/3^8$. Esto se debía a que para una etiqueta molecular determinada, cada posición de base tenía dos posibles alternativas de nucleótidos, lo que daba como resultado un total de 2 * 8 variantes de una base, suponiendo que hubiera tres bases posibles para cada posición en un código de barras estocástico. Si un tercer código de barras estocástico se extrajo continuamente al azar de 3^8 códigos de barras estocásticos con etiquetas moleculares únicas, entonces la probabilidad de que su etiqueta molecular no sea similar a las etiquetas moleculares de los dos anteriores, p_3, = $(3^8 - 1 - 16 - 1 - 16)/3^8$ = $(3^8 - 2 * 17)/3^8$. Los códigos de barras estocásticos se pueden extraer continuamente desde 3^8 códigos de barras estocásticos únicos hasta el código de barras estocástico r. La probabilidad de que este último código de barras estocástico no sea similar a los códigos de barras estocásticos anteriores, p_r, = $(3^8 - (r - 1) * 17)/3^8$. Debido a que todos los r códigos de barras estocásticos se dibujaron de forma independiente, la probabilidad de dibujar todos los códigos de barras estocásticos que no tengan secuencias similares, *P* (todas las etiquetas moleculares no tengan secuencias similares), = $p_1 * p_2 * p_3 * \ldots * p_r$. Por lo tanto, la probabilidad de tener al menos un par de códigos de barras estocásticos similares entre r códigos de barras estocásticos de 3^8 códigos de barras estocásticos con etiquetas moleculares únicas fue (al menos un par de etiquetas moleculares con secuencias similares) = 1 - P (todas las etiquetas moleculares no tienen secuencias similares). A continuación, se calculó el tamaño de muestra r mediante esta ecuación estableciendo un valor deseable para (al menos un par de etiquetas moleculares que tienen secuencias similares) = 0,01, 0,05, 0,1 o un valor deseado.

Tabla 1. Probabilidad de observar códigos de barras estocásticos con etiquetas moleculares similares para un número determinado de etiquetas moleculares únicas

N = 3^8 (6561)		N = 4^8 (65536)	
p	*r*	p	*r*
0,01	4	0,01	8
0,05	7	0,05	17
0.1	10	0,1	25
0,2	14	0,2	35
0,5	24	0,5	61
0,9	43	0,9	111

[0409] La Tabla 1 muestra la probabilidad de tener al menos un par similar entre *r* etiquetas moleculares dadas 3^8 o 4^8 etiquetas moleculares únicas. Si hubiera 3^8 códigos de barras estocásticos únicos y se seleccionaran ≤ 7 códigos de barras estocásticos (17 si hubiera 4^8 códigos de barras estocásticos únicos), la probabilidad de observar un par de códigos de barras estocásticos con etiquetas moleculares similares fue inferior a 0,05, lo cual es insignificante. Por lo tanto, como lo justifica esta pequeña probabilidad, las etiquetas moleculares similares eran más probablemente artefactos que una selección casual real de códigos de barras estocásticos similares y pueden corregirse.

[0410] Sin embargo, si estuvieran presentes más de 7 a 24 códigos de barras estocásticos, entonces la probabilidad de observar más de un par de códigos de barras estocásticos con etiquetas moleculares similares sería mayor (por ejemplo, 0,5). Por lo tanto, no se puede descartar con seguridad la posibilidad de que estos códigos de barras estocásticos fueran verdaderos y no artefactos. Por el contrario, la intuición común puede haber llegado a la conclusión errónea de que si sólo se extrajeran 24 códigos de barras estocásticos de un gran conjunto de 6561 posibilidades únicas, cualquier desviación de una base podría ser el resultado de un error de secuenciación y no de coincidencias.

[0411] Por ejemplo, si se muestrearon aleatoriamente 115 códigos de barras estocásticos, entonces sería 100 % seguro de que habrá al menos un par de códigos de barras estocásticos con etiquetas moleculares similares porque la probabilidad calculada sería uno. Supongamos que había 115 objetivos en la muestra, luego de los procesos de hibridación y transcripción inversa, serían observables dos pares de códigos de barras estocásticos con etiquetas moleculares similares y 111 de códigos de barras estocásticos con etiquetas moleculares no similares (en total 115 códigos de barras estocásticos). Sin embargo, si en los datos de secuenciación se observaron tres pares de códigos de barras estocásticos con etiquetas moleculares similares y 110 de códigos de barras estocásticos con etiquetas moleculares no similares (en total 116 códigos de barras estocásticos), entonces se descarta la posibilidad de que solo dos pares de códigos de barras estocásticos con etiquetas moleculares similares las etiquetas eran verdaderas y el tercer par fue creado por algunos errores. Esta probabilidad del 100 % indicaría que el evento de observar al menos un par de códigos de barras estocásticos con etiquetas moleculares similares puede ocurrir cuando se muestrearon aleatoriamente 115 códigos de barras estocásticos durante la codificación de barras estocásticas; sin embargo, esto puede no significar que todos los pares observados de etiquetas moleculares similares fueran verdaderos. Se pueden generar códigos de barras estocásticos con etiquetas moleculares similares a partir de códigos de barras estocásticos, etiquetas moleculares reales o verdaderas, o a partir de errores de PCR, artefactos o errores de secuenciación, errores o etiquetas moleculares falsas. Por lo tanto, puede ser necesaria una evaluación adicional para determinar si un par particular de etiquetas moleculares sería verdadero si se observaran etiquetas moleculares similares. Además, para cada probabilidad, se pueden requerir más códigos de barras estocásticos para esperar pares similares de etiquetas moleculares al aumentar la variedad total de etiquetas moleculares de 3^8 a 4^8.

[0412] La Tabla 2 y la Tabla 3 muestran que cuando se observaron ≤ 7 códigos de barras estocásticos con etiquetas moleculares únicas, era muy poco probable que ocurrieran etiquetas moleculares similares porque la probabilidad de tal aparición era inferior a 0,05. En consecuencia, es probable que esas etiquetas moleculares similares hayan sido causadas por errores de PCR, artefactos o errores de secuenciación, y estos deben eliminarse de los recuentos de etiquetas moleculares para corregir o ajustar los recuentos de etiquetas moleculares. Por lo tanto, el número total de etiquetas moleculares verdaderas en la Tabla 2 y la Tabla 3 se puede reducir de 5 a 1 y de 7 a 6 respectivamente. Sin embargo, se observaron 23 códigos de barras únicos en la Tabla 4, lo que espera alrededor del 50 % de probabilidad de tener al menos un par de códigos de barras estocásticos con etiquetas moleculares similares. En consecuencia, aunque es probable que 16 pares de códigos de barras estocásticos con etiquetas moleculares similares sean reales, cada par de etiquetas moleculares similares requerirá una evaluación adicional para confirmar si son reales.

Tabla 2. Datos sin procesar de CD69 de 10 pg de ARN de entrada en el pocillo H01 (se observaron 4 pares de etiquetas moleculares similares en el código de barras más abundante)

ID de gen	EM	Número de lecturas por EM
CD69	TGTGCGTG	261
CD69	TGTGCGCG	2
CD69	TGTGCGGG	2
CD69	GGTGCGTG	1
CD69	TGGGCGTG	1

Tabla 3. Datos brutos de CD64 de 500 pg de ARN de entrada en el pocillo A01 (se observó 1 par de etiquetas moleculares similares).

ID de gen	EM	Número de lecturas por EM
CD4	CGTGTCTG	10
CD4	GGGGGCGA	7
CD4	GCTGCTGG	5
CD4	TCGGGCGA	4
CD4	CGCGTTCA	1
CD4	CGCGTTTA	1
CD4	TGGGCTTG	1

Tabla 4. Datos sin procesar de TFRC de 10 pg de ARN de entrada del pocillo E11 (se observaron 16 pares de etiquetas moleculares similares en total a partir de los 4 códigos de barras estocásticos más abundantes

ID de gen	EM	Número de lecturas por EM
TFRC	CCGCTCTG	369
TFRC	CGTGGTTC	363
TFRC	CGGGGGTG	335
TFRC	CCCCCTTG	22
TFRC	CGGGGGGG	6
TFRC	CGGGGCCG	5
TFRC	CCGCGCTG	4
TFRC	CGGCGTTC	3
TFRC	CGTGTCCC	2
TFRC	TGGTGTTC	2
TFRC	CTGCTCTG	2
TFRC	GGCCGTGG	2
TFRC	TCGCTCTG	2
TFRC	CCGCCCTG	2
TFRC	CCCGCTTG	1
TFRC	CCGTTCTG	1
TFRC	GGGTGTTC	1
TFRC	CGTTGTTC	1
TFRC	GGGGGTGG	1
TFRC	GGGTCCGG	1
TFRC	CGGCCTG	1
TFRC	CGGCGTCC	1
TFRC	CCGCTCCG	1

[0413] En conjunto, estos datos demuestran que el número de códigos de barras estocásticos con etiquetas moleculares similares observados se eliminó porque los errores de PCR, artefactos o errores de secuenciación probablemente dieron como resultado estos códigos de barras estocásticos con etiquetas moleculares similares.

Ejemplo 2

Determinación del estado de calidad de un objetivo en datos de secuenciación

[0414] Este ejemplo demuestra que la determinación del estado de calidad de una diana en los datos de secuenciación es el estado de calidad de la secuenciación completa, el estado de calidad de la secuenciación incompleta o el estado de calidad de la secuenciación saturada. El estado de calidad del objetivo dependía de si se observaban todos las etiquetas moleculares verdaderas o reales.

[0415] Como se ilustra en el Ejemplo 1, el recuento completo de códigos de barras estocásticos con etiquetas moleculares únicas presentes en la biblioteca puede depender en gran medida de la profundidad de secuenciación. Cuanto más profunda sea la secuenciación, más probable será que se observaran etiquetas todas las verdaderas moléculas. La secuenciación superficial sería menos costosa, pero puede pasar por alto muchas etiquetas moleculares y posiblemente también comprometa la sensibilidad de la detección de genes. La secuenciación completa puede significar que se observaron todas las etiquetas moleculares verdaderas de los códigos de barras estocásticos utilizados para etiquetar la molécula objetivo, y la secuenciación incompleta puede significar que solo se observaron algunas de las etiquetas moleculares verdaderas. Además, es posible que más de 48568 moléculas diana estuvieran presentes en la muestra inicial (que sería el límite inferior del número de moléculas después de la corrección o ajuste de Poisson basado en ver 6561 - 2 * desviación estándar de distintos códigos de barras estocásticos). Entonces, la secuenciación saturada puede ocurrir cuando el número de moléculas diana sería difícil de determinar debido a una limitación en la diversidad total de etiquetas moleculares. Sin embargo, la secuenciación saturada sería poco probable si se utilizara una cantidad baja de ARN como entrada para códigos de barras estocásticos.

[0416] Para definir matemáticamente la secuenciación completa o incompleta, cada una se comparó con un modelo teórico sin ningún error. En condiciones experimentales perfectas, cada copia de una molécula objetivo en la muestra inicial puede generar $(1 + C)^j$ copias dados j ciclos de PCR y eficiencia de C para cada ciclo. A continuación, para cada molécula con código de barras en la muestra inicial, la secuenciación de Illumina puede considerarse esencialmente como un muestreo de Poisson a partir de copias clonales $(1 + C)^j$ amplificadas a partir de las moléculas con código de barras originales. En teoría, para el mismo gen objetivo, la secuenciación de k moléculas objetivo con códigos de barras estocásticos puede considerarse como un muestreo de Poisson repetido a partir de copias $(1 + C)^j$ porque todas las moléculas con códigos de barras estocásticas pueden ser igualmente representables después de la PCR. Una suposición importante del modelo de Poisson fue que la media es igual a la varianza y las lecturas de secuenciación deben seguir la equidispersión. La dispersión se puede definir como varianza/media.

[0417] En la práctica, la secuenciación completa a menudo puede ir acompañada de errores que normalmente se agrupaban en frecuencias de lectura mucho más bajas porque, a diferencia de las verdaderas etiquetas moleculares, es poco probable que los errores participen en todos los ciclos de PCR y, en consecuencia, darían lugar a menos copias, lo que provocaría una variación en la frecuencia de lectura que es mucho mayor en comparación con Poisson. FIGS. 19A-19B muestran ejemplos de genes secuenciados completa e incompletamente. En la FIG. 19A, la lectura de secuenciación más grande fue más de 350 veces mayor que la lectura de secuenciación más pequeña. Por lo tanto, la secuenciación completa tiende a exhibir un índice de dispersión mayor (> 1) en relación con Poisson.

[0418] Por el contrario, para la secuenciación incompleta, solo se han secuenciado algunos de los códigos de barras estocásticos con etiquetas moleculares verdaderas en la biblioteca y, por lo tanto, la variación en las lecturas de secuenciación sería menor en comparación con Poisson. En la FIG. 19B, la lectura de secuenciación más grande fue solo aproximadamente 3 veces mayor que la lectura de secuenciación más pequeña. Por lo tanto, la secuenciación incompleta tendería a exhibir un índice de dispersión más pequeño (<1) en relación con Poisson.

[0419] Además de calcular el índice de dispersión, la lectura de secuenciación del marcador de molécula más abundante se puede utilizar para decidir si la secuenciación está completa. Por ejemplo, el estado de secuenciación se puede clasificar como completo si la lectura del índice molecular más abundante fue 25 y el índice de dispersión fue 5; de lo contrario, puede clasificarse como incompleto. Se puede utilizar un umbral de 25 lecturas porque la secuenciación puede estar incompleta hasta que comienzan a aparecer errores de secuenciación. Es probable que se generen errores de secuenciación si cualquier etiqueta molecular se observara más de 25 veces.

[0420] En circunstancias en las que los datos de secuenciación para un gen altamente abundante estaban saturados en un código de barras estocástico, por ejemplo excediendo 6557 para 3^8 códigos de barras estocásticos con etiquetas moleculares únicas, la información de secuenciación para otros genes de menor expresión dentro del mismo pocillo se puede usar en su lugar para calcular el índice de dispersión y la lectura máxima de secuenciación de ese gen. Por ejemplo, si el segundo gen más abundante en el mismo pocillo no se ha saturado en códigos de barras estocásticos y se clasifica como secuenciación incompleta, entonces la saturación del primer gen se puede considerar real y no se puede calcular el número de moléculas. Y si el segundo gen más abundante fuera clasificado como de secuenciación completa, entonces la saturación del primer gen podría ser artificial y la aparición de todos los códigos de barras estocásticos podría deberse a errores. Y el algoritmo de umbral basado en el modelo de Poisson se puede utilizar para identificar el número de etiquetas moleculares verdaderas.

[0421] En conjunto, estos datos demuestran que la determinación del estado de secuenciación es secuenciación completa, secuenciación incompleta o secuenciación saturada.

Ejemplo 3

Corrección de errores de secuenciación o PCR con sustituciones de una base para genes completamente secuenciados

[0422] Este ejemplo muestra la corrección de errores de secuenciación o PCR con sustituciones de una base para genes completamente secuenciados, es decir, genes con estatutos de calidad de secuenciación completa en los datos de secuenciación. Este ejemplo también muestra el establecimiento de umbrales de etiquetas moleculares de una diana, por ejemplo un gen, para determinar etiquetas moleculares verdaderas y etiquetas moleculares falsas asociadas con la diana en los datos de secuenciación.

[0423] La tasa de error de secuenciación por nucleótido puede variar del 0,1 al 1 % y normalmente puede verse como lecturas de baja frecuencia. A medida que la secuenciación sea más profunda, es probable que se generen más errores de secuenciación. Por ejemplo, si el verdadero error de secuenciación de nucleótidos fue del 0,5 % y una etiqueta molecular se secuenció 100 veces, entonces el número esperado de errores de secuenciación asociados con esta etiqueta molecular puede ser aproximadamente 4, calculado a partir de $100 * (1 - (1 - 0{,}5\ \%)^8)$ si la etiqueta molecular tenía 8 nucleótidos de longitud. Si la etiqueta molecular se secuenció 300 veces, entonces el número esperado de errores de secuenciación puede ser aproximadamente 12. Estos errores de secuenciación pueden crear secuencias de etiquetas moleculares artificiales que pueden inflar los recuentos. Estas etiquetas moleculares se pueden eliminar antes de realizar más análisis.

[0424] Entre todos los errores de secuenciación, los errores de una base pueden ocurrir con mucha más frecuencia que aquellos que están separados por más de una base. La probabilidad de tener un error de secuenciación de una base se puede derivar de una distribución binomial con un tamaño de muestra de 8 y una probabilidad de éxito igual a la tasa de error de secuenciación de una base. Uno de los objetivos era corregir los errores de secuenciación de una base. Los errores de secuenciación de una base pueden considerarse secundarios de la etiqueta molecular más abundante y cercana (por ejemplo, en términos de distancia de Hamming), la etiqueta molecular principal. Los errores de secuenciación se identificaron encontrando los verdaderos etiquetas secundarias de la etiqueta molecular principal (es decir, etiquetas moleculares secundarias que están a una base de la etiqueta molecular principal).

Tabla 5. Datos de secuenciación TFRC actualizados después de eliminar errores de secuenciación de una base del pocillo E11 de ARN de entrada de 10 pg

ID de gen	EM	Número de lecturas por EM
TFRC	CCGCTCTG	378
TFRC	CGTGGTTC	374
TFRC	CGGGGGTG	335
TFRC	CCCCCTTG	23
TFRC	CGGGGGGG	6
TFRC	CGGGGCCG	5
TFRC	CCGCGCTG	4
TFRC	GGCCGTGG	2
TFRC	GGGGGTGG	1
TFRC	GGGTCCGG	1
TFRC	CGGCGTCC	1

Seleccionar etiquetas moleculares para principales y secundarias

[0425] Se puede exigir que las etiquetas moleculares principales tengan > 25 lecturas de secuenciación y que las etiquetas moleculares secundarias no tengan más de 3 lecturas de secuenciación. Estos requisitos se basaron en el razonamiento siguiente. Supongamos que la tasa de error de secuenciación por nucleótido fuera del 0,5 %. Si una etiqueta molecular se secuenciaba 25 veces y se generaban 200 nucleótidos en total, entonces se esperaba que un nucleótido fuera un error porque 200 * 0,005 = 1. Por lo tanto, para cada etiqueta molecular con una lectura de secuenciación de 25, se esperaba que tener al menos una etiqueta secundaria. Se puede suponer que la etiqueta molecular parental debe tener una lectura de secuenciación de 25. Era poco probable que las etiquetas moleculares infantiles con lecturas de secuenciación de 4 tuvieran errores de secuenciación. Esto se debió a que la probabilidad de introducir el mismo error de secuenciación cuatro veces en una etiqueta molecular es $8 * 0{,}005^4 = 5 * 10^{-9}$. Si hubo 106 lecturas de secuenciación en total, entonces el número esperado de errores de secuenciación que se repitió cuatro veces sería $5 * 10^9 * 10^6 = 0{,}005$, lo cual fue insignificante. Por lo tanto, las etiquetas moleculares infantiles deben tener lecturas ≤ 3.

Dada una etiqueta molecular parental y sus etiquetas moleculares secundarias relacionadas que están separadas por una base, ¿cómo determinar las etiquetas moleculares secundarias que son realmente errores de secuenciación de la etiqueta parental?

[0426] Dada una etiqueta molecular parental y un conjunto de etiquetas moleculares secundarias que eran una base diferentes de la etiqueta molecular parental con lecturas de secuenciación ($R_{secundaria\ 1}$, $R_{secundaria\ 2}$,.., $R_{secundaria\ m}$), se puede realizar una prueba binomial múltiple. Se utiliza para identificar etiquetas moleculares de niños verdaderos. Bajo la hipótesis nula, la abundancia de las etiquetas moleculares de las etiquetas verdaderas debería ser menor o igual que $R_{par} * p$ (matemáticamente, H_0: $p < e/2$); de lo contrario, se puede concluir a favor de la hipótesis alternativa de que la abundancia era mayor que $R_{par} * p$ (H_A: $p > e/2$) y se puede rechazar la hipótesis de que la etiqueta molecular era una verdadera etiqueta molecular secundaria. Entonces, la probabilidad de que una etiqueta molecular secundaria sea una base diferente de su etiqueta molecular parental y se observe una vez sería $p = e/2$. Entonces, matemáticamente, la probabilidad $p_{secundaria}$ de observar esta etiqueta molecular secundaria al menos $R_{secundaria}$ veces de la abundancia total ($R_{secundaria} + R_{par}$) sería la siguiente:

$$p_{secundaria} = p(X \leq R_{secundaria} \mid R_{secundaria} + R_{par}, p = e/2)$$

$$= \sum_{R_{par}}^{j=R_{secundaria}} \binom{R_{par} + R_{secundaria}}{j} p^j (1-p)^{R_{par}+R_{secundaria}-j}, \text{ Ecuación (1)}$$

[0427] Si la etiqueta molecular secundaria era de hecho un error de secuenciación de su etiqueta molecular principal, entonces la probabilidad $p_{secundaria}$ debería ser mayor que el valor crítico al 5 %. Debido a que se prueban múltiples hipótesis simultáneamente, el valor crítico utilizado para rechazar la hipótesis nula puede determinarse mediante las tasas de falso descubrimiento (FDR) controladas a un nivel del 5 %, y la hipótesis puede aceptarse si $p_{secundaria}$ es mayor que FDR a un nivel del 5 %. Con FDR controlado al nivel del 5 %, los valores p no ajustados se pueden ordenar en orden creciente, como $p_1 \leq p_2 \leq \ldots \leq p_m$. A continuación, se encuentra la prueba con su correspondiente rango *j*. Si $p_{secundaria} \leq j/m * 5$ %, entonces se puede aceptar la hipótesis nula de que esta etiqueta molecular secundaria fue un error de secuenciación de una base de la etiqueta molecular principal.

[0428] En conjunto, estos datos demuestran los pasos para corregir errores de secuenciación de una base para un gen completamente secuenciado: Paso (1), seleccione la etiqueta molecular con la lectura de secuenciación más abundante como la primera etiqueta molecular principal si su lectura de secuenciación es mayor que 25. Paso (2), seleccione etiquetas moleculares con lecturas de secuenciación ≤ 3 e identifique aquellas etiquetas moleculares que están a una base de la primera etiqueta molecular principal y llámelas etiquetas moleculares secundarias; Si no se encuentra ninguna etiqueta molecular secundaria o no se encuentran etiquetas moleculares secundarias de una base, vaya al paso (5). Paso (3), realice múltiples pruebas binomiales en todas las etiquetas moleculares secundarias y en las etiquetas moleculares principales, y elimine aquellas etiquetas moleculares secundarias cuya hipótesis nula se acepte y atribuya sus lecturas de secuenciación a sus principales. Si no se aceptara ninguna de las hipótesis nulas, eso implicaría que todas las etiquetas moleculares secundarias no eran errores de secuenciación de una base de la etiqueta molecular principal, y no sería necesario realizar ninguna corrección de lectura. Paso (4), actualice las secuencias de etiquetas moleculares y las lecturas de secuenciación. Paso (5), elija la etiqueta molecular con la siguiente secuencia más grande leída como etiqueta molecular principal y repita los pasos anteriores hasta que no quede ninguna etiqueta molecular parental calificada o ninguna etiqueta molecular secundaria calificada.

[0429] La Tabla 5 muestra datos de secuenciación TFRC actualizados después de eliminar errores de secuenciación de una base utilizando el análisis anterior. El número único de etiquetas moleculares se redujo de 23 (que se muestra en la Tabla 4) a 11.

Utilizar modelos de Poisson para establecer umbrales

[0430] Es más probable que aparezcan errores de secuenciación durante la secuenciación completa. Algunos tipos de errores, como los errores de secuenciación de una base, pueden corregirse, pero otros errores, como la incorporación aleatoria de etiquetas moleculares artificiales, no pueden corregirse basándose en la similitud de secuencia. En cambio, este tipo de errores se pueden identificar mediante modelos. Como se analizó anteriormente, la secuenciación completa tendería a estar demasiado dispersa en relación con Poisson. Por lo tanto, se crearon dos modelos de Poisson distintivos que presentan la sobredispersión: uno puede usarse para modelar las lecturas de secuenciación de las etiquetas moleculares verdaderas (es decir, las secuencias de etiquetas moleculares utilizadas para etiquetar las moléculas objetivo durante la codificación de barras estocástica) y el segundo modelo puede usarse para las etiquetas moleculares de error (es decir, secuencias de etiquetas moleculares no utilizadas durante la codificación de barras estocástica pero que aparecieron después de la secuenciación debido a errores). La tasa de error de secuenciación puede ser de aproximadamente 0,1-1 % y la tasa de error de PCR puede ser de aproximadamente 0,001 %. Los errores de PCR pueden ocurrir más en los últimos ciclos de PCR, lo que genera errores en las etiquetas moleculares con lecturas de secuenciación bajas, pero contribuyen a una gran fracción de todas las secuencias de etiquetas moleculares observadas. En consecuencia, los errores generados mediante PCR y secuenciación a menudo pueden tener lecturas de secuenciación más bajas que las verdaderas etiquetas moleculares. Por lo tanto, se supuso que la media de Poisson para las lecturas de secuenciación de las etiquetas moleculares verdaderas sería mayor que la media de Poisson para las etiquetas moleculares de error.

[0431] Supongamos que había *k* etiquetas moleculares distintivas en total, y *t* de ellas eran etiquetas moleculares verdaderas como BC_1, BC_2,.., BC_t y el resto eran etiquetas moleculares de error como BC_{t+1}, BC_{t+2},.., BC_k. Las lecturas de secuenciación que se asignaron a esas etiquetas moleculares verdaderas y de error pueden ser R_1, R_2,.., R_t y R_{t+1}, R_{t+2},.., Rk. Suponiendo también que las medias de Poisson que utilizan etiquetas moleculares verdaderas y de error fueran μ_t y μ_n con $\mu_t > \mu_n$, entonces la probabilidad de todo el proceso sería

$$L_1(X;\ \mu_t, \mu_n) = (X;\ \mu_t) * (X;\ \mu_n)$$
$$= \prod_{i=1}^{t} P(X_i = R_i \mid \mu_t)\ \prod_{j=(t+1)}^{k} P(X_j = R_j \mid \mu_n). \quad \text{Ecuación (3)}$$

donde $P(X_i = R_i \mid \mu_t)$ denota la probabilidad de observar iº molecular con abundancia R_i bajo un proceso de Poisson con media μ_t.

[0432] Para determinar *t* el número de etiquetas moleculares verdaderas, se consideraron varios modelos, comenzando desde el modelo que suponía que todas las etiquetas moleculares eran verdaderas (entonces *l* = *k)*; y el segundo modelo que asumió que la etiqueta molecular menos abundante era un error y que todas las demás etiquetas moleculares eran verdaderas (entonces *l* = *k* - 1); hasta el último modelo que suponía que solo la etiqueta molecular más abundante era verdadera y todas las demás eran etiquetas moleculares de error (por lo que *l* = 1). Por último, el mejor modelo sería el que tenga la mayor probabilidad entre todos los modelos considerados, o su equivalente, el Criterio de Información de Akaike (AIC) más pequeño, que puede usarse en la selección de modelos midiendo la calidad relativa de cada modelo posible para los datos dados. Matemáticamente, AIC se puede definir como AIC = *-2logL + 2p,* donde p es el número de parámetros estimados en el modelo. Entonces, para L_k y L_1, $p = 1$, y para otros casos $p = 2$. El ejemplo de la Tabla 6 muestra que solo tres etiquetas moleculares con las tres lecturas de secuenciación más grandes se consideraron etiquetas moleculares verdaderas entre los 8 modelos posibles comparados. Además, la FIG. 20 muestra que el umbral derivado del modelo seleccionado (los 3 más grandes) separó claramente las lecturas aparentes de etiquetas moleculares verdaderas de aquellas que eran errores más probables.

Tabla 6. Todos los modelos posibles junto con sus probabilidades logarítmicas para TFRC a partir del pocillo E11 de ARN de entrada de 10 pg

ID de gen	EM	Número de lecturas por EM
TFRC	TODOS	-1294
TFRC	más grande 8	-952
TFRC	más grande 7	-813
TFRC	más grande 6	-660
TFRC	más grande 5	-481
TFRC	más grande 4	-263
TFRC	más grande 3	-51
TFRC	más grande 2	-656
TFRC	más grande 1	-1036

[0433] Estos datos demuestran las lecturas de secuenciación de genes completamente secuenciados corregidos eliminando errores de secuenciación de una base y umbrales utilizando modelos de Poisson.

Ejemplo 4

Ajuste de genes secuenciados de forma incompleta

[0434] Este ejemplo muestra el ajuste de genes secuenciados de forma incompleta mediante la eliminación de genes ruidosos y el uso de un modelo de Poisson truncado en cero para estimar el número total de etiquetas moleculares que se espera que estén presentes en la biblioteca.

Eliminar genes ruidosos

[0435] Además de considerar las estadísticas de las etiquetas moleculares y sus lecturas de secuenciación, el análisis a nivel genético también puede ser informativo. Un gen puede considerarse ruidoso si se detectaron muy pocas etiquetas moleculares y cada etiqueta molecular tiene lecturas inusualmente bajas en relación con genes completamente secuenciados. Esta suposición se basó en el razonamiento de que las moléculas con códigos de barras estocásticos dentro de la misma biblioteca deberían amplificarse y secuenciarse aproximadamente a la misma frecuencia. Esta expectativa puede verse afectada por la PCR o el sesgo de secuenciación causado por diferencias en la secuenciación de cada molécula, pero se esperaba que fuera pequeña en relación con el "ruido" creado por eventos como la contaminación de la muestra o la recombinación molecular no deseada durante la PCR. Un gen puede ser ruidoso si su

tasa de amplificación (lecturas promedio por etiqueta molecular) fue similar a la tasa de amplificación de errores que se derivaron de genes completamente secuenciados en la misma biblioteca.

[0436] Específicamente, supongamos que un gen g_1 completamente secuenciado constaba de etiquetas moleculares verdaderas t_1 y etiquetas moleculares de error e_1 en total, de modo que $R_{g1,1}$, $R_{g1,2}$,.., $R_{g1,t1}$ fueran lecturas de secuenciación asignadas a las etiquetas moleculares verdaderas, y $R^*_{g1,1}$, $R^*_{g1,2}$,.., $R^*_{g1,e1}$ fueron lecturas de secuenciación asignadas a las etiquetas moleculares de error. Entonces, la tasa de amplificación de etiquetas moleculares de error (EAMP) para g_1 fue

$$EAMP_{g1} = \sum_{i=1}^{e_1} R^*_{g1,i} / e_1$$

De manera similar, EAMP se puede calcular para g_2, g_3,..., g_x para todos los demás genes completamente secuenciados. Se puede aplicar un límite para un gen potencialmente ruidoso g'$_1$ que tiene menos de 5 etiquetas moleculares observadas en total y lecturas de secuenciación $R_{g'1,1}$, $R_{g'1,2}$,.., $R_{g'1,k}$ asignadas a cada etiqueta molecular y determinar su tasa de amplificación como

$$amp_{g'1} = \sum_{i=1}^{k} R_{g'1,i} / k$$

Si $amp_{g'1} < mediana(amp_{g'1}, amp_{g'2}, amp_{g'x})$, el gen g'$_1$ se consideró un gen ruidoso. De lo contrario, puede considerarse un gen incompleto. Otros genes ruidosos pueden probarse y eliminarse de manera similar. La razón para elegir 5 etiquetas moleculares como límite fue porque puede ser deseable tratar genes con tasas de amplificación más bajas en dos casos separados: artefactos (con etiquetas moleculares observadas menos de 5) y secuenciación incompleta (con etiquetas moleculares observadas ≥ 5 debido a fallo del cebador de PCR/secuenciación baja).

Estimaciones utilizando el modelo de Poisson truncado en cero

[0437] Cuando la secuenciación fue incompleta, es posible que todavía haya errores presentes en los datos, pero pueden ser difíciles de identificar debido a lecturas de secuenciación insuficientes en general. Cuando la secuenciación fue superficial y no se observaron todas las etiquetas moleculares presentes en la biblioteca, pueden ser necesarias algunas suposiciones para un análisis significativo. Se puede suponer que todas las etiquetas moleculares observadas eran verdaderas y que las etiquetas moleculares verdaderas no observadas estaban siendo truncadas en cero, es decir, etiquetas moleculares truncadas observadas cero veces. Aunque no se han muestreado en la secuenciación todas las transcripciones con códigos de barras estocásticos para un gen determinado, la frecuencia de lecturas de las etiquetas moleculares detectadas se puede utilizar para estimar la diversidad total de etiquetas moleculares presentes en la biblioteca completa aplicando un modelo de Poisson truncado en cero.

[0438] Supongamos que se observaron k etiquetas moleculares distintivas con lecturas (R_1, R_2,.., R_k) y que las etiquetas moleculares (S - k) no se observaron con lecturas cero. Uno de los objetivos era estimar S, el número total de etiquetas moleculares que se esperaba que estuvieran presentes en la biblioteca. Suponiendo que las frecuencias de lectura de secuenciación sean 1, 2, 3 o más a medida que las variaciones de Poisson se truncan en cero con la media de Poisson μ, y la suma de todas las lecturas de secuenciación fue n, entonces la probabilidad se puede expresar como:

$$L(S, \mu) \propto S!/(S - k)! \mu^n \exp(-S\mu) \quad \text{Ecuación (3)}$$

[0439] Se pueden aplicar procedimientos de inferencia tradicionales para la estimación de μ, S y sus errores estándar. La probabilidad máxima (MLE) de μ fue n/S, y las aproximaciones al MLE de S pueden ser $k/(1 - e^{-n/S})$ o $k/(1 - (1 - 1/S)^n)$. FIG. 21 muestra el modelo de Poisson truncado en cero ajustado basado en el número de etiquetas moleculares y sus correspondientes lecturas de secuenciación. Como se muestra en la FIG. 21, se observaron 33 etiquetas moleculares únicas en un total de 39 lecturas en la biblioteca parcialmente secuenciada. Con base en las frecuencias de las etiquetas moleculares con las lecturas de secuenciación 1, 2, 3 y 4, se aplicó un modelo de Poisson para estimar que un total de 113 etiquetas moleculares en la biblioteca completa habían completado la secuenciación. Se aplicaron procedimientos de inferencia para la estimación de m, S y sus errores estándar. El MLE de μ puede ser n/S, y las aproximaciones al MLE de S pueden ser $k/(1 - e^{-n/S})$ o $k/(1 - (1 - 1/S)^n)$.

[0440] En conjunto, estos datos demuestran las lecturas de secuenciación de genes secuenciados de forma incompleta corregidas eliminando genes ruidosos y usando un modelo de Poisson truncado en cero para estimar el número total de etiquetas moleculares que se espera que estén presentes en la biblioteca.

Ejemplo 5

Ajuste de genes completamente secuenciados y genes incompletamente secuenciados

[0441] Este ejemplo muestra un ejemplo de resultados generados después de ajustar las lecturas de secuenciación de genes completamente secuenciados y genes secuenciados de forma incompleta.

[0442] La Tabla 7 proporciona un ejemplo de resultados generados después de ajustar las lecturas de secuenciación de genes completamente secuenciados y genes secuenciados de forma incompleta. Las descripciones de los encabezados de las columnas fueron las siguientes: "Gene ID" muestra el nombre del gen detectado. El "Estado de secuenciación" muestra tres resultados posibles: completo, incompleto y saturado, lo que determina otros métodos de análisis. La clasificación dependió del índice de dispersión y de la lectura de secuenciación asignada a la etiqueta molecular (EM) más abundante. "EM bruta" muestra el recuento de etiquetas moleculares únicas observadas para ese gen ('0' para genes no detectados). "Lecturas sin procesar" muestra la suma de las lecturas de secuenciación asignadas a Raw EM ('0' para genes no detectados). La EM corregido muestra el recuento de etiquetas moleculares únicas que se consideraron etiquetas moleculares verdaderas después de aplicar el algoritmo (solo para genes secuenciados completos, 'NA' para genes incompletos, '0' para genes ruidosos y no detectados). "Lecturas corregidas" muestra la suma de las lecturas de secuenciación asignadas a EM corregido (solo para genes secuenciados completos, 'NA' para genes incompletos, '0' para genes ruidosos y no detectados). "EM extrapolado" muestra el número estimado de etiquetas moleculares únicas a través del modelo de Poisson truncado cero (solo para datos secuenciados incompletos, 'NA' para datos completos, '0' para genes ruidosos y genes no detectados). "Mol estimado" muestra el número de moléculas estimadas en función de EM corregido (para genes secuenciados completos) o EM extrapolado (para genes secuenciados incompletos), '0' para genes ruidosos y genes no detectados. "Mol estimado LB " muestra el límite inferior para el número estimado de moléculas. "Mol estimado UB" muestra el límite superior del número estimado de moléculas.

[0443] En la Tabla 7, Mol estimado (*n*), número estimado de moléculas de partida, se calculó de la siguiente manera:

$$n = -m \log(1 - k/m), \quad \text{Ecuación (4)}$$

donde *m* fue la variedad total de etiquetas moleculares (3^8) y *k* fue el número total de etiquetas moleculares únicas observadas. La varianza de *n,* var(*n*), se derivó usando la expansión de Taylor: $\text{var}(n) = (m/(m - k))^2 \text{var}(k)$, donde var(*k*) se puede expresar como $m * (1 - (1 - 1/m)^n) (1 - 1/m)^n + m(m - 1) ((1 - 2/m)^n - (1 - 1/m)^{2n})$. Límites inferior y superior del número estimado de moléculas fijas (Mol estimado LB y Mol estimado UB) se calcularon usando

$$n \pm 2\sqrt{var(n)}$$

[0444] En su conjunto, estos datos demuestran el ajuste de genes secuenciados completamente y genes secuenciados de forma incompleta.

Tabla 7. Ejemplo de salida del pocillo A06 de ARN de entrada de 10 pg

ID de gen	Estado de secuenciación	Lecturas brutas	EM bruta	Lecturas corregidas	EM corregido	EM extrapolado	Mol estimado	Estimado Mol LB	Estimado Mol UB
DAP_PRE ETIQUETADO	Incompleto	1	1	N/A	N/A	N/A	N/A	N/A	N/A
PHE_PRE ETIQUETADO	Completo	3129	59	3059	11	N/A	11,01	10,83	11,19
KAN_PRE ETIQUETADO	Completo	1418	71	1259	8	N/A	8	7,87	8,14
CD4	Incompleto	2	2	0	0	0	0	0	0
FOXP3	Incompleto	0	0	0	0	0	0	0	0
ESTAT5A	Completo	96	4	94	1	N/A	1	1	1
FOXO1	Incompleto	1	1	0	0	0	0	0	0
FOXO3	Incompleto	3	3	0	0	0	0	0	0
CD45RA	Incompleto	0	0	0	0	0	0	0	0
CD45RO	Incompleto	0	0	0	0	0	0	0	0
MKI67	Incompleto	1	1	0	0	0	0	0	0
GAPDH	Completo	12059	709	10984	38	N/A	38,11	37,45	38,77
IM C	Completo	2842	63	2765	7	N/A	7	6,89	7,12
LRRC32	Incompleto	0	0	0	0	0	0	0	0
IL1R1	Completo	177	1	177	1	N/A	1	1	1
CCR8	Incompleto	10	1	N/A	N/A	1	1	1	1
IL-26	Incompleto	0	0	0	0	0	0	0	0

Ejemplo 6

Rendimiento de la corrección de genes completamente secuenciados y genes incompletamente secuenciados

[0445] Este ejemplo muestra el rendimiento de la corrección de las lecturas de secuenciación de genes completamente secuenciados. El rendimiento se basó en los errores y el ruido en los datos de recuento de etiquetas moleculares sin procesar eliminados y en las lecturas de secuenciación que quedaron.

[0446] Se seleccionaron varios genes secuenciados completamente para probar el rendimiento de la corrección de las lecturas de secuenciación de genes completamente secuenciados. La Tabla 8 compara algunas medidas para esos genes antes y después de corregir o ajustar las lecturas de secuenciación. EM sin procesar, lecturas sin procesar, EM corregido y lecturas corregidas se importaron directamente desde la tabla de salida. El amplificador sin procesar (tasa de amplificación usando datos sin procesar) y el amplificador filtrado (tasa de amplificación usando datos de etiquetas moleculares reales después de las correcciones) se calcularon usando (lecturas sin procesar/EM sin procesar) y (lecturas corregidas/EM corregida). El porcentaje de EM retenido en comparación con el número de etiquetas moleculares verdaderas después de las correcciones del número total observado de etiquetas moleculares fue 100 * IM corregido/EM sin procesar y el % de lecturas retenidas se definió de manera similar como 100 * Lecturas corregidas/Lecturas sin procesar. La Tabla 8 muestra genes de ejemplo de diferentes niveles de abundancia con GAPDH y ACTB que muestran recuentos de etiquetas moleculares y lecturas totales más altos. El número de etiquetas moleculares verdaderas después de aplicar las correcciones representó menos del 7 % del total de etiquetas moleculares observadas en los datos sin procesar, lo que implica que más del 93 % de las etiquetas moleculares se consideraron etiquetas moleculares erróneas y se descartaron. Aunque el 93 % de las etiquetas moleculares sin procesar se eliminaron como ruido, las etiquetas moleculares verdaderas contribuyen al menos al 72 % de las lecturas, lo que implica que las etiquetas moleculares de error descartadas eran de lecturas mucho más bajas. Además, las tasas de amplificación después de aplicar el algoritmo oscilaron entre 137 y 413, que fueron mucho más altas que las obtenidas con datos sin procesar (6,1 a 29,4). Las tasas de amplificación corregidas fueron mediciones mucho más realistas que se correlacionaban con una eficiencia de la PCR de al menos el 75 %.

Tabla 8. Comparaciones de genes seleccionados del conjunto de datos D704 (entrada de ARN de 10 pg) y Cliente 1 (entrada de ARN unicelular, desconocida) antes y después de usar el algoritmo

	Conjunto de datos	D704			Cliente 1		
	Pocillo	H01	H01	H01	D01	H01	H12
	ID de gen	GAPDH	IM C	TFRC	LEFTY1	SOX17	ACTB
Antes	EM sin procesar	760	76	45	50	28	2070
	Lecturas crudas	11107	1710	1324	434	297	12625
	amp sin procesar	14,61	22,5	29,4	8,68	10,61	6,1
Después	EM corregido	37	4	3	1	1	66
	Lecturas corregidas	9351	1589	1239	387	226	9039
	% EM retenido	5 %	5 %	7 %	2 %	4 %	3 %
	% Lecturas retenidas	84 %	93 %	94 %	89 %	76 %	72 %
	amp filtrado	252,73	397,25	413	387	226	137

[0447] En conjunto, estos datos indican que la corrección de las lecturas de secuenciación de genes completamente secuenciados redujo significativamente los errores y el ruido en los datos de recuento de etiquetas moleculares sin procesar, manteniendo al mismo tiempo la capacidad de utilizar la mayoría de las lecturas de secuenciación.

Ejemplo 7

Herramientas para resumir y visualizar datos de conteo de objetivos con códigos de barras estocásticos

[0448] Este ejemplo muestra herramientas para resumir y visualizar datos de recuento de objetivos con códigos de barras estocásticos ilustrados en los Ejemplos anteriores.

[0449] Para los datos de prueba, se generaron dos placas de células individuales para su procesamiento mediante el ensayo Precise™ (Cellular Research, Inc. (Palo Alto, CA)). El experimento utilizó dos tipos de células diferentes en una proporción de 4:1 y el investigador que realizó el experimento desconoció la identidad de las células depositadas en cada pocillo. El objetivo de este estudio fue identificar tipos de células para cada pocillo utilizando perfiles de expresión genética a partir de recuentos estocásticos de códigos de barras.

[0450] Para evaluar la calidad general de los datos de secuenciación en todos los pocillos, se sumó la suma de las lecturas de secuenciación por pocillo. Y para evaluar el desempeño del método de corrección, se tabularon y compararon algunas medidas estadísticas antes de aplicar y después del método de corrección. Además, los trazados gráficos pueden proporcionar presentaciones visuales de los datos y detectar anomalías o patrones fácilmente.

[0451] Las Tablas 9 y 10 muestran la suma de las lecturas de secuenciación por pocillo para la Placa 1 con lecturas de secuenciación < 5000 en cursiva. Para aquellos pocillos con lecturas mucho más bajas, como lecturas <5000, podría indicar que no se asignó ninguna célula a ese pocillo y un análisis adicional debería excluir esos pocillos.

Tabla 9. Suma de lecturas de secuenciación por pocillo para la placa 1

	1	2	3	4	5	6	7	8	9	10	11	12
A	58346	59687	141814	57269	106511	26894	5908	40547	16783	19896	8993	*3885*
B	*3894*	24164	52131	40458	61725	55568	*3701*	5189	*3353*	*2690*	*3799*	*3848*
C	98195	175593	*3627*	*4508*	22856	17253	*3435*	49083	29189	28969	*4570*	25593
D	59960	29342	27304	22649	29341	55575	*2861*	*2298*	78385	56112	*3357*	15843
E	18148	49333	25546	40571	*4190*	31746	111060	*3956*	57039	12688	13917	31934
F	37105	34979	88619	5457	*3552*	6499	*2951*	*2874*	*3127*	*3118*	166177	*2848*
G	11930	119412	2179	30913	29445	6002	*4260*	41497	16535	48453	15058	17844
H	29988	28987	*3414*	*2548*	74279	*4609*	*4164*	*4043*	35764	*2911*	*3276*	38723

Tabla 10. Suma de lecturas de secuenciación por pocillo para la placa 2

	1	2	3	4	5	6	7	8	9	10	11	12
A	2291	45737	124283	95919	50637	*3822*	*2770*	*4147*	20528	93362	24443	29416
B	26870	26126	3350	42649	63897	*2960*	71319	13673	40682	28120	*2439*	*3387*
C	52029	51582	232	20538	28556	18871	49769	24187	43879	17136	106123	17861
D	41845	124785	26143	*4836*	*2740*	18416	37201	90892	87375	21552	29307	31768
E	22866	29578	8506	76852	15211	18138	4087	19625	17151	22380	15928	2242
F	16868	46605	21848	53195	*3391*	94041	29196	19468	38771	*2801*	5703	*2735*
G	22187	30552	20213	*3010*	*3143*	*4578*	*1901*	*2873*	163669	34898	60463	19940
H	30216	25110	112489	34535	*3976*	32726	17868	9807	58375	22972	38446	18290

[0452] Las Tablas 10 y 11 comparan varias medidas antes y después de usar el método de corrección. A partir de estas tablas, se notaron las enormes variaciones en las 'lecturas sin procesar' (suma de lecturas de secuenciación por pocillo) y el 'EM sin procesar' (número total de recuentos de etiquetas moleculares por pocillo). Las enormes variaciones pueden deberse a que sus desviaciones estándar (DE) fueron mayores que la media, lo que nuevamente indica la presencia de lecturas bajas en los pocillos. Después de utilizar el método, alrededor del 47 % de los genes por pocillo se clasificaron como genes secuenciados completamente entre todos los genes presentes. Si la mayoría de los genes se clasificaran como genes incompletos (como el 0 %), es posible que el método actual no elimine el ruido de los datos. Para cada pocillo, solo se retuvo el 15 % de las etiquetas moleculares después de la corrección de genes completos, pero esas etiquetas moleculares se asignaron al 95 % de las lecturas de secuenciación en promedio. Un valor más alto de % de lecturas retenidas indica que el método de corrección puede capturar señales de manera efectiva (lecturas que fueron aportadas desde etiquetas moleculares verdaderas) mientras elimina el ruido. Además, la tasa de amplificación para cada etiqueta molecular retenida como etiqueta molecular verdadera fue de 163,32, mucho más alta que 22,76 antes de aplicar el método de corrección.

Tabla 11. Resumen de estadísticas usando diferentes medidas antes y después de aplicar el algoritmo usando datos de la placa 1

	Medida/Estadísticas	Media	DE	Mediana	Mín.	Máx.
	Genes	35,31	11,41	35,5	16	60
	EM sin procesar	1184,54	1202,71	837	137	5617
Antes	Lecturas sin procesar	31110,6	35808,4	21272,5	2179	175593
	Amp	22,76	7,42	24,08	4,45	35,53
Después	% genes completos	0,47	0,2	0,54	0,1	0,79
	% IM retenido	0,15	0,03	0,16	0,07	0,22
	% Lecturas retenidas	0,95	0,02	0,96	0,86	0,98
	Amp corregido	163,32	19,36	165,47	116,96	194,86

Tabla 12. Estadísticas resumidas usando diferentes medidas antes y después de aplicar el algoritmo usando datos de la placa 2

	Medida/Estadísticas	Media	DE	Mediana	Mín.	Máx.
	Genes	36,52	9,92	37,5	16	57
	EM sin procesar	1159,13	1050,72	851, 5	79	5303
Antes	Lecturas sin procesar	32602,8	32144,1	2291 9	232	163669
	Amp	25,25	6,44	27,7 2	2,94	34,09
Después	% genes completos	0,53	0,2	0,6	0,05	0,78
	% IM retenido	0,17	0,04	0,17	0,11	0,5
	% Lecturas retenidas	0,95	0,01	0,95	0,89	0,98
	Amp corregido	161,63	24,06	163,88	27	195,74

[0453] FIG. 22 muestran gráficos de barras de lecturas de secuenciación totales por pocillo. FIG. 22 proporciona una visualización directa de la entrada relativa en 96 pocillos. Esta figura muestra que los pocillos C02 y F11 tuvieron lecturas más altas en comparación con otros, lo que podría indicar múltiples células para esos pocillos. Los pocillos A12, B01, B07-B12, C03, C04, C07, C11, D07, D08, D11, E05, E08, F04-F10, F12, G03, G07, H03, H04, H07-H09, H10-H11 tenían mucho. lecturas más bajas en relación con otros pocillos, lo que podría indicar que no se colocaron células en esos pocillos.

[0454] FIG. 23 muestran gráficos de barras de % de genes completamente secuenciados, % de etiquetas moleculares (EM) retenidas como etiquetas moleculares verdaderas y % de lecturas retenidas asignadas a aquellas EM retenidas para cada pocillo. FIG. 23 ilustra el porcentaje de genes por pocillo que se clasificaron como completos y el método de corrección que se puede aplicar para eliminar el ruido (la fila inferior para cada pocillo); el nivel de ruido por pocillo usando etiquetas moleculares (el porcentaje de etiquetas moleculares se consideró como etiquetas moleculares verdaderas después de aplicar el método de corrección en relación con las etiquetas moleculares observadas antes de aplicar el método de corrección, la fila superior para cada pocillo); y el nivel de ruido por pocillo utilizando lecturas de secuenciación (porcentaje de lecturas asignadas a etiquetas moleculares verdaderas en relación con el total de lecturas sin procesar, la fila central para cada pocillo). Como se muestra, el porcentaje de genes completamente secuenciados varió con los pocillos, pero fue mucho menor en los pocillos A12, B01, V07-B12, C03, C04, C07, D07, D08, D11, E05, E08, F04, F07-F10, F12. G03, G07, H03, H06, H07, H10-H11, que correspondieron a aquellos pocillos con lecturas mucho más bajas. El % de EM retenido indicado en la fila superior fue generalmente inferior al 20 % para todos los pocillos, mientras que el % de lecturas retenidas indicado en la fila del medio superó el 90 % en general para todos los pocillos. Este tipo de gráfico puede proporcionar una idea general sobre cuán efectivo fue el método de corrección para eliminar el ruido y al mismo tiempo maximizar su señal para cada pocillo.

[0455] FIG. 24 muestran diagramas de caja del % de lecturas retenidas variadas según los genes para cada pocillo. Los diagramas de caja a nivel de gen revelaron información detallada, como la medida en que cada método de corrección había funcionado para cada gen en un pocillo, lo que puede no reflejarse en el gráfico de barras a nivel de pocillo. Diagrama de caja del % de lecturas retenidas para todos los genes completamente secuenciados por pocillo en la FIG. 24 revelaron que las variaciones entre genes pueden ser sustanciales, como en el caso de los pocillos D11, F4, F8, H3 y H8, que tenían bigotes extendidos más allá de 0,6. Pero esos cinco pocillos correspondieron a lecturas de secuenciación totales mucho más bajas: 3357, 5457, 2874, 3414 y 4043.

[0456] La agrupación se puede utilizar en el análisis de datos de expresión génica. El análisis de componentes principales (PCA) se puede utilizar para la reducción de dimensiones reduciendo la multidimensionalidad y las variables posiblemente correlacionadas a unas pocas variables linealmente no correlacionadas mediante una transformación ortogonal. Los componentes principales de PCA se pueden usar en la búsqueda de grupos en los datos.

[0457] Las FIGS. 25A-25B muestran gráficos de PCA del uso de EM sin procesar frente a IM corregido después de aplicar el algoritmo de dos placas. FIG. 25A muestra el gráfico de PCA del uso de EM sin procesar por gen por pocillo para pocillos con lecturas de secuenciación totales > 5000. Este gráfico de PCA se generó eliminando primero los pocillos que tenían lecturas de secuenciación totales < 5000 (dando como resultado 139 pocillos y 107 genes excluyendo 3 genes controlados); en segundo lugar, eliminar genes con EM sin procesar cero en 139 pocillos (quedaron 85 genes); tercero, tomar el logaritmo de EM sin procesar más uno para incorporar ceros en el conjunto de datos y luego aplicar PCA en los datos de registro después de centrar y escalar. El gráfico PCA muestra dos grupos aparentes, pero para pocillos como D02, D05 y F06, a distancias aproximadamente iguales de ambos grupos, los tipos de células fueron difíciles de determinar. Los resultados de agrupación pueden corromperse debido a la adición de ruido; incluso unas pocas variables de ruido pueden corromper una estructura de grupo clara. Por lo tanto, la agrupación puede beneficiarse de un paso de preprocesamiento de selección de características/variables o de un paso de filtrado o eliminación de ruido. Al aplicar el método de corrección a los datos secuenciados completos, resultó una estructura de grupo clara como se muestra en la FIG. 25B. El gráfico PCA en la FIG. 25B se generó de manera similar al gráfico PCA anterior en la FIG. 25A excepto el uso de EM corregido (recuentos de etiquetas moleculares verdaderas para genes secuenciados completos después de aplicar el método de corrección) en lugar de EM sin procesar (recuentos de etiquetas moleculares para todos los genes detectados antes de aplicar el algoritmo), que utilizó 75 genes en 139 pocillos en total. Se observaron dos grupos

distintivos, que estaban bien separados por el eje y (PC2). En comparación con la FIG. 25A, grupos en la FIG. 25B tenían un tamaño más compacto y las células de cada pocillo se asignaron claramente a un grupo. Además, el grupo más pequeño a la derecha del eje y de la FIG. 25B constan de 31 pocillos, alrededor del 22 % del total de pocillos y bastante cerca del 20 % esperado.

[0458] En conjunto, estos datos demuestran varias herramientas útiles para resumir y visualizar datos de recuento de objetivos con códigos de barras estocásticos.

Ejemplo 8

Cobertura de EM de cada EM en una placa para un gen de alta expresión - ACTB

[0459] Este ejemplo demuestra que las distribuciones distintas de los errores de EM derivados durante la secuenciación o la PCR generalmente tienen distribuciones distintas de las EM verdaderos.

[0460] Además del recuento absoluto de la expresión génica y la corrección del sesgo de la PCR, las EM pueden proporcionar una mejor comprensión de la calidad estadística del procedimiento de preparación de la biblioteca y los datos de secuenciación. Al observar el número de lecturas que presentan el mismo gen EM, denominado cobertura de EM, es posible detectar llamadas de bases erróneas en la secuenciación o errores de PCR generados durante la preparación de la biblioteca. Por ejemplo, un gen EM de un SL determinado que está representado por múltiples lecturas probablemente sea una medida precisa en comparación con un gen EM de un SL determinado que está representado por una sola lectura. Los códigos de barras de baja cobertura de EM en presencia de códigos de barras de alta cobertura de EM en la misma biblioteca suelen ser artefactos o errores generados durante la secuenciación o los pasos de PCR durante la preparación de la biblioteca. Los errores de EM derivados durante la secuenciación o la PCR generalmente tienen distribuciones distintas de los verdaderos EM. FIG. 27 es un gráfico ejemplar que muestra la cobertura de etiquetas moleculares de cada etiqueta molecular a través de una placa de micropocillos para un gen de alta expresión - ATCB, donde se observaron distintas distribuciones entre etiquetas moleculares de error y etiquetas moleculares reales. FIG. 28 es un gráfico ejemplar que muestra el ajuste de dos distribuciones binomiales negativas a la cobertura de etiquetas moleculares de cada etiqueta molecular a través de una placa de micropocillos para un gen de alta expresión: ATCB. El ajuste de dos distribuciones binomiales negativas demuestra que se pueden distinguir estadísticamente los errores de etiqueta molecular con una profundidad de etiqueta molecular más baja y una etiqueta molecular verdadera con una profundidad de etiqueta molecular más alta. El eje x es la profundidad molecular.

[0461] En conjunto, estos datos demuestran que los errores de EM derivados durante la secuenciación o la PCR generalmente tienen distribuciones distintas de las EM verdaderos.

Ejemplo 9

Corrección de etiquetas moleculares debido a errores de sustitución de secuenciación o PCR

[0462] Este ejemplo demuestra un método para corregir etiquetas moleculares debido a PCR y errores de sustitución de secuenciación que se pueden aplicar a ensayos de transcriptoma completo sin asumir una cobertura uniforme y sin requerir una alta cobertura de secuenciación para el estado de secuenciación completo.

[0463] La deduplicación se realizó en la primera coordenada de mapeo y etiquetas moleculares únicas (UMI) de cada lectura, y se supuso que las lecturas eran idénticas dada la misma coordenada de inicio, UML y hebra. Después de la deduplicación, se conservaron los UML con los recuentos más altos por grupo (Tabla 13).

[0464] Las etiquetas moleculares (EM) se corrigieron por gen. Para cada gen, se identificaron grupos de IM con adyacencia direccional. El método de adyacencia direccional agrupaba los IM si las EM estaban dentro de una distancia de Hamming de 1 y un recuento de EM principal ≥ 2* (recuento de IM secundario) - 1. Se consideró que todos las EM dentro del mismo grupo se originaban en el mismo EM principal y los recuentos de EM secundarios se contrajeron al EM principal. FIG. 29 muestra la corrección de la etiqueta molecular, donde la distancia de Hamming por pares de 1 estaba sobrerrepresentada. Después de la corrección de las etiquetas moleculares, las etiquetas moleculares con las distancias de Hamming de una distancia se agruparon y colapsaron en la misma etiqueta molecular principal. FIG. 30 muestra que la curva del número corregido de EM versus el número de lecturas converge. Debido a que se conservaron todas las lecturas, este método también se puede utilizar para eliminar errores de secuenciación o PCR de una base.

Tabla 13. Después de deduplicar etiquetas moleculares, solo se consideró como error un número insuficiente, dado un ensayo de transcriptoma completo, de etiquetas moleculares únicas

		Conteos crudos		Deduplicar UML	
Muestra	Pocillo	Nº de lecturas	Nº de IM únicos	Nº de lecturas	Nº de IM únicos
Placa 2	A02	234646	3079	28925	2335
Placa 3	A02	773050	14126	95023	11410

[0465] En conjunto, estos datos demuestran un método de corrección que se puede aplicar para corregir o ajustar datos de ensayos de transcriptoma completo porque todas las lecturas fueron reentrenadas.

Ejemplo 10

Recuento de etiquetas moleculares para muestras con alta entrada

[0466] Este ejemplo describe etiquetas moleculares únicas utilizadas a medida que las moléculas de entrada aumentan

[0467] El ensayo dirigido BD Precise™ puede ser el más adecuado cuando se utiliza en entradas de muestras pequeñas, como en células individuales, para permitir el etiquetado estocástico y único de los ARNm. A medida que aumenta el número de transcripciones en relación con el conjunto de códigos de barras en experimentos con entrada alta de ARN/células, el porcentaje de EM que se reciclan para marcar el mismo gen aumenta y se calculó teóricamente utilizando una distribución de Poisson (FIG. 26). En estas situaciones, sin corrección estadística, la cuantificación de la expresión génica utilizando IM subestimaría el número de moléculas que están inicialmente presentes sin correcciones de Poisson o correcciones basadas en dos distribuciones binomiales negativas.

[0468] En muestras de entrada extremadamente altas donde el número de ARNm por gen supera la colección completa de 6561 códigos de barras, ya no es posible una corrección de Poisson o una corrección basada en dos distribuciones binomiales negativas. Por ejemplo, independientemente de las 65.000 o 100.000 moléculas de entrada, se espera un máximo de 6.561 códigos de barras saturados en cualquier caso. Por lo tanto, los genes y las muestras que parecen tener un alto aporte de muestra pueden alterarse, por lo que probablemente se subestimarían los recuentos de EM.

[0469] En conjunto, estos datos demuestran la necesidad de ajustar los datos sin procesar al cuantificar la expresión génica utilizando EM.

Ejemplo 11

Corrección de errores de sustitución recursiva (RSEC)

[0470] Este ejemplo demuestra la corrección de errores de sustitución recursiva.

[0471] Se pueden emplear dos métodos colaborativos en el proceso de análisis del ensayo dirigido BD Precise™ para eliminar errores de EM. En resumen, los errores de EM que se derivan de errores de sustitución de células base de secuenciación se identifican y se ajustan al verdadero código de barras de EM mediante la corrección de errores de sustitución recursiva (RSEC). Posteriormente, los errores de EM que se derivan de los pasos de preparación de la biblioteca o de los errores de eliminación de la base de secuenciación se ajustan mediante la corrección de errores basada en distribución (DBEC).

[0472] El algoritmo RSEC puede ajustar los errores de EM que se derivan de la PCR o la sustitución de secuenciación. Estos raros eventos erróneos se han observado al examinar la cobertura de LD. Por ejemplo, la cobertura de EM para EM de error puede ser significativamente menor que la de IM verdaderos en muestras adecuadamente secuenciadas (FIG. 27); en los casos en los que se utilizan dos EM muy similares durante los pasos iniciales de Molecular Indexing™ (transcripción inversa), generalmente tendrían una cobertura de EM similar y no es necesario eliminarlos. A medida que aumenta la profundidad de la secuenciación, aparecen más errores de EM, por lo que RSEC puede ser crucial para ajustar el recuento de EM para bibliotecas de códigos de barras altamente secuenciadas.

[0473] En resumen, RSEC considera dos factores en la corrección de errores: 1) Similitud en la secuencia EM; y 2) y su cobertura de EM. Para cada gen objetivo, las EM están conectados cuando ambas secuencias de EM están dentro de 1 base (distancia de Hamming = 1) entre sí. Para cada conexión entre EM x e y, si:

$$\text{Cobertura}(y) > 2*\text{Cobertura}(x)+1, \qquad \text{Ecuación (5)}$$

donde y denota "EM principal" y x denota "EM secundario".

[0474] Según esta asignación, las EM secundarias se pueden contraer a su EM principal. Este proceso es recursivo hasta que no haya más IM identificables entre principales y secundarios para el gen.

[0475] FIG. 31 muestra una ilustración esquemática de un ejemplo del esquema de corrección de errores de sustitución recursiva anterior. Las EM en los datos sin procesar antes de la corrección RSEC incluyen nueve EM únicas: GTCAAATT, GTCAAAAT, GTCAAAAA, TTCAAAAA, TTCAGAAA, CTCAAAAA, TTCAAACT, TTCAAAAT y TTCAAACA. Al aplicar RSEC, GTCAAATT se puede contraer a GTCAAAAT porque las dos EM difieren en un nucleótido (subrayado) y las EM GTCAAATT tiene un recuento de EM más bajo que el de GTCAAAAT. A su vez, las EM GTCAAAAT se puede colapsar en las EM GTCAAAAA (la diferencia en las secuencias de EM está subrayada), que tiene un recuento de EM mayor que el de GTCAAAAT. De manera similar, las EM TTCAGAAA y CTCAAAAA se pueden contraer en las EM TTCAAAAA. La EM TTCAAACT se puede contraer en las EM TTCAAAAAT, que a su vez se puede contraer en EM TTCAAAA. La EM

TTCAAACA se diferencia de todos los demás EM en más de un nucleótido y, por lo tanto, no se integra en ninguno de los otros ocho EM. Antes de la corrección RSEC, el número de recuento de EM sin procesar era nueve. Después de la corrección RSEC, el número de recuento de EM fue dos: EM TTCAAAAA y TTCAAAAA.

[0476] En conjunto, estos datos demuestran el uso de RSEC para corregir recuentos de EM sin procesar.

Ejemplo 12

Cálculos de cobertura de EM

[0477] Este ejemplo describe el cálculo de cobertura de EM.

[0478] Después de RSEC, se evalúan los recuentos de genes EM por pocillo para determinar su idoneidad para una corrección adicional. Los genes con baja cobertura de EM (< 4 lecturas por EM) omiten los pasos de corrección posteriores y se informan en la tabla de datos de EM final y se registran como "baja profundidad" en el proceso de bioinformática. Para genes con entradas extremadamente altas en las que se observan al menos 6557 de los 6561 códigos de barras posibles, donde resulta difícil determinar la cantidad de moléculas debido a la diversidad de códigos de barras y los genes se marcan como "saturados". Para las EM de genes que no cumplen con ninguno de los 2 puntos de decisión, avance al algoritmo DBEC posterior y se marcan como "Aprobado" en el archivo de registro de salida. Además, los genes con un promedio superior a 650 EM por pocillo se registran como "entrada alta", ya que >5 % de estas EM se reciclan según una distribución de Poisson (FIG. 27).

[0479] En conjunto, este ejemplo describe el cálculo de cobertura de EM.

Ejemplo 13

Corrección de errores basada en distribución (DBEC)

[0480] Este ejemplo describe la corrección de errores basada en distribución.

[0481] A diferencia de RSEC, el algoritmo DBEC es un método para discriminar si una EM es una señal de error o verdadera independientemente de su secuencia de EM. Mientras que RSEC utiliza tanto la secuencia de EM como la información de cobertura de EM para corregir errores, DBEC se basa principalmente en la cobertura de EM solo para corregir errores que no son de sustitución. Como se mencionó anteriormente, los códigos de barras de error generalmente tienen una cobertura de EM baja que se diferencia de la cobertura de EM de los códigos de barras verdaderos; esta diferencia en la cobertura de EM se puede observar en un gráfico de histograma de la cobertura de EM como distribuciones distintas (FIG. 27). Dada esta diferencia, DBEC ajusta dos distribuciones binomiales negativas para distinguir estadísticamente entre errores de EM (con menor cobertura de EM) y una para señales verdaderas con mayor cobertura de EM.

Eliminación de EM recicladas para un ajuste de distribución óptimo

[0482] Para un gen dado, a medida que aumenta las EM detectadas, el porcentaje de EM reciclada (es decir, la misma EM se usa para marcar 2 o más ARNm del mismo gen) aumenta y se puede estimar. Utilizando una distribución de Poisson ($\lambda_{no\ única}$), el número de EM recicladas para el pocillo *i* ($n_{no\ único,\ i}$) se estima a partir de la ecuación de tasa de reciclaje de EM (Ecuación (6)). Si las EM recicladas estimadas es superior al 5 % de la EM total para el gen dado en el pocillo i, este gen en el pocillo i se marca como "Alto aporte". Para estos datos de "alta entrada", las EM de cobertura superior de EM se eliminarían del ajuste de la distribución, pero se conservarían para pasos de conteo posteriores, para obtener una mejor distribución binomial negativa.

$$P(X > 1 | \lambda_{no\ única}), \lambda_{no\ único} = Número\ de\ ML / 6561 \qquad \text{Ecuación (6)}$$

$$n_n \quad N_{no\ único} = \sum_{i=1}^{n\ pocillos} n_{no\ único,i} \qquad \text{Ecuación (7)}$$

Adición de pseudopuntos para genes de baja expresión

[0483] Si el número único de EM es inferior a 10, a menudo resulta más difícil ajustar las distribuciones debido a la escasez de datos. Para aliviar este problema, DBEC agrega pseudopuntos al 1 % de recuentos de señales que se utilizan para ayudar en el ajuste de la distribución, pero no afecta los datos.

Estimación de parámetros

[0484] Para ajustar dos distribuciones binomiales negativas para separar el error de la señal EM, se aproximan dos conjuntos de valores iniciales para la estimación de parámetros. Se supone que la distribución del error es binomial negativa con media y dispersión de 1.

Estimación de probabilidad de error/señal

[0485] Suponga que las distribuciones de señal y error son Binomiales Negativas ($\mu_{señal}$, $tamaño_{señal}$) y Binomiales Negativas (μ_{error}, $tamaño_{error}$), respectivamente. Para determinar el número de señales EM, en orden ascendente, se calculan las probabilidades de que el número de lecturas de una EM determinada provengan de las distribuciones de señal y error hasta que se cumpla la ecuación (8), donde todas las EM anteriores se consideran EM de error.

$$P(X = r \mid \mu = \mu_{error}, tamaño = tamaño_{error}) < P(X = r \mid \mu = \mu_{señal}, tamaño = tamaño_{señal}) \quad \text{Ecuación (8)}$$

[0486] En conjunto, este ejemplo muestra cálculos para realizar la corrección de errores basada en distribución.

Ejemplo 14

Ajuste por errores de SL basado en segundas derivadas

[0487] Este ejemplo demuestra el ajuste de errores de SL basado en segundas derivadas.

[0488] FIG. 32, los paneles (a)-(e) muestran resultados ejemplares de corrección de errores de secuenciación y PCR basados en segundas derivadas del cambio de profundidad del marcador molecular. FIG. 32, panel (a) muestra que los errores de SL y las EM de señal se pueden separar bien. FIG. 32, los paneles (b) y (d) muestran gráficos de suma acumulativa de recuentos de etiquetas moleculares de recuentos de EM mostrados en la FIG. 32, paneles (c) y (e), respectivamente. Las líneas verticales en la FIG. 32, los paneles (b) y (d) muestran las posiciones de los máximos de las segundas derivadas. Las líneas de puntos en la FIG. 32, los paneles (c) y (e) muestran que las posiciones de los máximos de las segundas derivadas pueden separar EM en los gráficos de recuentos de EM frente a la profundidad de lectura de EM.

[0489] En su conjunto, estos datos muestran que se pueden usar máximos de segundas derivadas del cambio de profundidad del marcador molecular para separar errores de SL de señales de EM.

Ejemplo 15

Corrección de errores de secuenciación y PCR basados en DBEC

[0490] Este ejemplo demuestra la corrección de errores de secuenciación y PCR basándose en dos distribuciones binomiales negativas.

[0491] FIG. 33, los paneles (a)-(c) muestran resultados ejemplares de corrección de PCR y errores de secuenciación basados en dos distribuciones binomiales negativas para CD69. FIG. 33, el panel (a) muestra el ajuste de dos distribuciones binomiales negativas (D_n para la distribución binomial negativa de ruido y D_s para la distribución binomial de señal) para CD69 en los datos de recuento de EM mostrados en el histograma de profundidad de EM en la FIG. 33, panel (b). La línea de puntos en 33, panel (b) muestra la separación de señales EM y errores SL determinados por las dos distribuciones binomiales negativas que se muestran en la FIG. 33, panel (a). La línea vertical en la FIG. 33, panel (c) muestra el máximo local de segundas derivadas determinado en función del gráfico de suma acumulada de lecturas. Similarmente a la FIG. 33, FIG. 34, los paneles (a)-(c) muestran resultados ejemplares de corrección de PCR y errores de secuenciación basados en dos distribuciones binomiales negativas para CD3E.

[0492] En conjunto, estos datos muestran que DBEC se puede utilizar para corregir errores de PCR y secuenciación.

Ejemplo 16

Reciclaje EM

[0493] Este ejemplo demuestra el reciclaje de EM para genes de alta expresión y la necesidad de ajustar los datos de entrada para genes de alta expresión antes del ajuste de la distribución.

[0494] FIG. 35, los paneles (a)-(c) muestran resultados ejemplares de corrección de PCR y errores de secuenciación basados en dos distribuciones binomiales negativas para un gen de alta expresión: ACTB. Un gen de alta expresión puede

tener un estado de sobresecuenciación (por ejemplo, con una cobertura de EM de 100 o más). En algunas formas de realización, se puede determinar un gen de alta expresión usando otros criterios. En la FIG. 35, panel (a), las etiquetas moleculares a la derecha de la línea vertical correspondían a EM probablemente reciclados en función de sus altas profundidades. FIG. 35, panel (b) ilustra esquemáticamente que las etiquetas moleculares se pueden dividir en tres categorías (además de los errores de EM): errores de SL, EM de señal y EM probablemente recicladas. FIG. 35, panel (c) demuestra que, sin ajustar por los probables EM recicladas, dos distribuciones binomiales negativas ajustadas no eran ideales.

[0495] FIG. 36 muestra resultados ejemplares del reciclaje de etiquetas moleculares ricas en G para genes de alta expresión. FIG. 36 muestra los 20 EM principales de alta profundidad para genes de alta expresión GAPDH, ACTB y HSP90AB1. Estas EM de alta profundidad tenían G y T altas, que tenían más probabilidades de reciclarse y los códigos de barras no eran estocásticos. Los dobletes de EM se produjeron antes que el cálculo teórico que supone una etiquetado estocástico. Para ACTB, los dobletes reales fueron de alrededor del cuatro por ciento, aunque teóricamente habría un 2,7 % de dobletes si hubiera 350 EM por pocillo.

[0496] FIG. 37, los paneles (a)-(b) muestran resultados ejemplares del ajuste de datos de entrada para genes de alta expresión antes de ajustar dos distribuciones binomiales negativas. FIG. 37, el panel (a) muestra los datos de entrada en la FIG. FIG. 35, panel (a) ajustado para genes de alta expresión. A diferencia de los accesorios de distribución no ideales de la FIG. 35, panel (c), FIG. 37, el panel (b) muestra dos distribuciones binomiales negativas ajustadas.

[0497] En conjunto, estos datos muestran que antes del ajuste de dos distribuciones binomiales negativas, es posible que sea necesario eliminar las EM recicladas de los datos de secuenciación para genes de alta expresión.

Ejemplo 17

IM, corrección de recuentos mediante dos distribuciones binomiales negativas

[0498] Este ejemplo demuestra recuentos de EM de diez objetivos corregidos usando dos distribuciones binomiales negativas.

[0499] FIG. 38, los paneles (a)-(j) muestran una validación ejemplar no limitante del conjunto de datos corregido utilizando dos distribuciones binomiales negativas. Los recuentos de EM a menudo se corrigieron como se muestra en la FIG. 38. La línea vertical en cada panel de la FIG. 38 muestra la separación de señales EM y errores SL para un objetivo determinado usando dos distribuciones binomiales negativas.

[0500] En conjunto, estos datos validan la corrección de los recuentos de EM utilizando dos distribuciones binomiales negativas.

Ejemplo 18

Visualización de incrustación vecina t-estocástica de un ensayo dirigido BD Precise™ a partir de 96 pocillos de células individuales mixtas de Jurkat y cáncer de mama (BrCa)

[0501] Este ejemplo demuestra un método para corregir errores de secuenciación y PCR basado en la corrección de errores de sustitución recursiva y la corrección de errores basada en la distribución para células individuales mixtas de Jurkat y cáncer de mama (BrCa).

[0502] FIG. 39, los paneles (a)-(d) muestran visualizaciones ejemplares de incrustación vecina t-estocástica (t-SNE) del ensayo dirigido Precise™ de 96 pocillos de células individuales mixtas de Jurkat y cáncer de mama (BrCa) (86 genes examinados). FIG. 39, el panel (a) muestra que los grupos de células se identificaron utilizando DBScan con los mismos parámetros antes y después de los ajustes de EM. FIG. 39, los paneles (b)-(d) muestran la expresión de marcador individual escalada tanto por color como por tamaño de punto. FIG. 39, panel (b) muestra PSMB4, un gen de mantenimiento que está presente en ambos tipos de células y después de los ajustes de EM, la falta de señal de PSMB4 se resalta aún más en el grupo "Señal baja". FIG. 39, el panel (c) muestra CD3E, un marcador de linfocitos que resalta los grupos de células Jurkat. FIG. 39, el panel (d) muestra CDH1, un marcador de células epiteliales que resalta el grupo BrCa.

[0503] En conjunto, estos datos demuestran que el ajuste de EM eliminó el ruido de EM que permitió una clara diferenciación de la expresión génica entre grupos de células.

Ejemplo 19

Análisis de expresión diferencial entre grupos de células

[0504] Este ejemplo demuestra un método para corregir errores de secuenciación y PCR basado en la corrección de errores de sustitución recursiva y la corrección de errores basada en la distribución para células de baja señal y células de cáncer de mama (BrCa).

[0505] FIG. 40, los paneles (a)-(b) son gráficos ejemplares no limitantes que muestran análisis de expresión diferencial entre grupos de células para genes con >0 EM en ambos grupos seleccionados calculados por DBScan y determinados por el nivel de marcador genético en cada grupo. FIG. 40, el panel (a) muestra la expresión del gen del grupo de 'baja señal' en comparación con el resto de las células. FIG. 40, la parte superior del panel (a) muestra una comparación de EM sin procesar que muestra que el ruido de EM fue generalmente mayor para los genes con una expresión promedio más alta en otras células. FIG. 40, la parte inferior del panel (a) muestra que después de los ajustes de EM utilizando RSEC y DBEC, el ruido de EM detectado en el grupo de "señal baja" se redujo, lo que permite una distinción más clara de la expresión genética entre los grupos. FIG. 40, el panel (b) muestra la expresión del gen del grupo 'BrCa' en comparación con el resto de las células. FIG. 40 superior del panel (b) muestra EM sin procesar en células sin BrCa que también tenía un recuento de EM significativo de marcadores de BrCa, como KRT1, MUC1. FIG. 40, la parte inferior del panel (b) muestra que las EM ajustados de los marcadores de BrCa estaban altamente enriquecidos en el grupo de BrCa que el resto de las células.

[0506] En conjunto, estos datos demuestran que los errores de secuenciación y PCR se pueden corregir basándose en la corrección de errores de sustitución recursiva y la corrección de errores basada en la distribución para células, tales como células de baja señal y células de cáncer de mama.

Ejemplo 20

Ajuste de los recuentos de etiquetas moleculares para células Jurkat y T47D mixtas

[0507] Este ejemplo demuestra un método para ajustar los recuentos de etiquetas moleculares para células Jurkat y T47D mixtas.

[0508] FIG. 41, los paneles (a)-(d) son gráficos ejemplares no limitantes que muestran la visualización de incrustación vecina t-estocástica (t-SNE) de un ensayo dirigido BD Precise™ de una placa de 96 pocillos de células individuales de Jurkat mixtas y de cáncer de mama (T47D) con 86 genes examinados. FIG. 41, panel (a) muestra que los grupos de células se identificaron utilizando DBScan con los mismos parámetros antes y después de los ajustes de EM. FIG. 41, los paneles (b)-(d) muestran la expresión del marcador individual escalada tanto por color como por tamaño de punto. FIG. 41, panel (b) muestra la escala de PSMB4, un gen de mantenimiento que estaba presente en ambos tipos de células y después de ajustes de EM. La falta de señal PSMB4 se destaca aún más en el grupo de control sin plantilla (NTC). FIG. 41, panel (c) muestra la escala de CD3E, un marcador de linfocitos que resalta los grupos de células Jurkat. FIG. 41, el panel (d) muestra la escala de CDH1, un marcador de células epiteliales que resalta el grupo T47D.

[0509] FIG. 42, los paneles (a)-(b) son mapas de calor ejemplares no limitantes que muestran la expresión génica diferencial mediante recuentos de etiquetas moleculares entre diferentes grupos de células identificados en la FIG. 41 antes de cualquier paso de corrección de errores (EM sin procesar se muestra en la FIG. 42, panel (a)) y después de la corrección RSEC y DBEC (EM ajustado se muestra en la FIG. 42, panel (b)). Los genes con baja expresión están en azul y los genes con alta expresión están en naranja. Los genes que son similares en el patrón de expresión genética entre estos tipos de células se agrupan. Sin corrección de errores, NTC tenía ruido de genes de alta expresión como CD3E y KRT18, que son marcadores Jurkat y T47D, respectivamente. Además, la corrección de errores reveló distintos patrones de expresión genética entre Jurkat y T47D.

[0510] En conjunto, estos datos demuestran que el ajuste de EM puede eliminar el ruido de IM, lo que permite una clara diferenciación de la expresión génica entre grupos de células.

[0511] Con respecto al uso de sustancialmente cualquier término plural y/o singular en el presente documento, aquellos con experiencia en la técnica pueden traducir del plural al singular y/o del singular al plural según sea apropiado para el contexto y/o aplicación. Las diversas permutaciones singulares/plurales pueden establecerse expresamente en el presente documento en aras de la claridad. Tal como se utiliza en esta especificación y en las reivindicaciones adjuntas, las formas singulares "un", "una", "el" y "ella" incluyen referencias en plural a menos que el contexto indique claramente lo contrario. Cualquier referencia a "o" en este documento pretende abarcar "y/o" a menos que se indique lo contrario.

[0512] Los expertos en la técnica entenderán que, en general, los términos utilizados en el presente documento, y especialmente en las reivindicaciones adjuntas (por ejemplo, cuerpos de las reivindicaciones adjuntas) generalmente pretenden ser términos "abiertos" (por ejemplo, el término " incluyendo" debe interpretarse como "incluido pero no limitado a", el término "tener" debe interpretarse como "tener al menos", el término "incluye" debe interpretarse como "incluye pero no se limita a", etc.). Los expertos en la técnica entenderán además que si se pretende un número específico de una recitación de reivindicación introducida, dicha intención se recitará explícitamente en la reivindicación y, en ausencia de dicha recitación, dicha intención no estará presente. Por ejemplo, como ayuda para la comprensión, las siguientes reivindicaciones adjuntas pueden contener el uso de las frases introductorias "al menos uno" y "uno o más" para introducir recitaciones de reivindicaciones. Sin embargo, el uso de tales frases no debe interpretarse en el sentido de que la

introducción de una recitación de reivindicación por los artículos indefinidos "un" o "una" limita cualquier reivindicación particular que contenga dicha recitación de reivindicación introducida a formas de realización que contengan sólo una recitación de este tipo, incluso cuando la misma reivindicación incluye las frases introductorias "uno o más" o "al menos uno" y artículos indefinidos como "un" o "una" (por ejemplo, "un" y/o "una" debe interpretarse en el sentido de "al menos uno" o "uno o más"); lo mismo se aplica al uso de artículos definidos utilizados para introducir recitaciones de afirmaciones. Además, incluso si se recita explícitamente un número específico de una recitación de reivindicación introducida, los expertos en la técnica reconocerán que dicha recitación debe interpretarse en el sentido de al menos el número recitado (por ejemplo, la simple recitación de "dos recitaciones", sin otros modificadores, significa al menos dos recitaciones, o dos o más recitaciones). Además, en aquellos casos en los que se aplique una convención análoga a "al menos uno de A, B y C, etc." se utiliza, en general dicha construcción está pensada en el sentido en que un experto en la técnica entendería la convención (por ejemplo, "un sistema que tiene al menos uno de A, B y C" incluiría, entre otros, sistemas que tienen A solo, B solo, C solo, A y B juntos, A y C juntos, B y C juntos, y/o A, B y C juntos, etc.). En aquellos casos en los que se aplique una convención análoga a "al menos uno de A, B o C, etc." se utiliza, en general dicha construcción está pensada en el sentido en que un experto en la técnica entendería la convención (por ejemplo, "un sistema que tiene al menos uno de A, B o C" incluiría, entre otros, sistemas que tener A solo, B solo, C solo, A y B juntos, A y C juntos, B y C juntos, y/o A, B y C juntos, etc.). Los expertos en la técnica entenderán además que prácticamente cualquier palabra y/o frase disyuntiva que presente dos o más términos alternativos, ya sea en la descripción, las reivindicaciones o los dibujos, debe entenderse que contempla las posibilidades de incluir uno de los términos. cualquiera de los términos, o ambos términos. Por ejemplo, se entenderá que la frase "A o B" incluye las posibilidades de "A" o "B" o "A y B".

[0513] Además, cuando las características o aspectos de la divulgación se describen en términos de grupos Markush, los expertos en la técnica reconocerán que la divulgación también se describe en términos de cualquier miembro individual o subgrupo de miembros del grupo Markush.

[0514] Como entenderá un experto en la técnica, para todos y cada uno de los fines, tales como en términos de proporcionar una descripción escrita, todos los rangos divulgados en el presente documento también abarcan todos y cada uno de los posibles subrangos y combinaciones de subrangos de los mismos. Se puede reconocer fácilmente que cualquier rango enumerado describe suficientemente y permite dividir el mismo rango en al menos mitades, tercios, cuartos, quintos, décimos, etc. iguales. Como ejemplo no limitante, cada rango discutido en el presente documento se puede descomponer fácilmente en un tercio inferior, un tercio medio y un tercio superior, etc. Como también entenderá un experto en la técnica, todos los lenguajes tales como "hasta", "al menos", "mayor que", "menor que" y similares incluyen el número citado y hacen referencia a rangos que posteriormente pueden dividirse en subrangos como se analizó anteriormente. Finalmente, como entenderá un experto en la técnica, un rango incluye cada miembro individual. Así, por ejemplo, un grupo que tiene de 1 a 3 artículos se refiere a grupos que tienen 1, 2 o 3 artículos. De manera similar, un grupo que tiene de 1 a 5 artículos se refiere a grupos que tienen 1, 2, 3, 4 o 5 artículos, y así sucesivamente.

REIVINDICACIONES

1. Un método para determinar el número de objetivos, que comprende:

(a) codificar de forma estocástica una pluralidad de objetivos de una muestra que comprende una o más células, en donde los objetivos comprenden ácidos nucleicos diana, usando una pluralidad de códigos de barras estocásticos para crear una pluralidad de objetivos con código de barras estocástico, en donde cada uno de los códigos de barras estocásticos es un código de barras de ácido nucleico, en el que cada uno de la pluralidad de códigos de barras estocásticos comprende una etiqueta molecular;
(b) obtener datos de secuenciación de los objetivos con códigos de barras estocásticos; y
(c) para uno o más de la pluralidad de objetivos:

(i) contar el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación;
(ii) identificar grupos de etiquetas moleculares de la diana usando adyacencia direccional;
(iii) colapsar los datos de secuenciación obtenidos en (b) usando los grupos de etiquetas moleculares de la diana identificada en (ii); y
(iv) estimar el número de la diana, en donde el número de la diana estimado se correlaciona con el número de etiquetas moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación contados en (i) después de colapsar los datos de secuenciación en (ii), que comprende:

ajustar la distribución de las etiquetas moleculares de la diana y sus apariciones a dos distribuciones binomiales negativas;
determinar el número de etiquetas moleculares verdaderas *n* usando las dos distribuciones binomiales negativas; y eliminar las etiquetas moleculares falsas de los datos de secuenciación obtenidos en (b), en donde las etiquetas moleculares falsas comprenden etiquetas moleculares con apariciones inferiores a la aparición del enésimo marcador molecular más abundante, y en donde las etiquetas moleculares verdaderas comprenden etiquetas moleculares con apariciones mayores que o iguales a la aparición de la enésima etiqueta molecular más abundante.

2. El método de la reivindicación 1, en el que la pluralidad de dianas comprende dianas de todo el transcriptoma de una célula.

3. El método de cualquiera de las reivindicaciones 1-2, en el que las etiquetas moleculares de la diana dentro de un grupo están dentro de un umbral predeterminado de adyacencia direccional entre sí.

4. El método de la reivindicación 3, en el que el umbral de adyacencia direccional es una distancia de Hamming de uno.

5. El método de cualquiera de las reivindicaciones 1-4, en el que las etiquetas moleculares de la diana dentro del grupo comprenden una o más etiquetas moleculares principales y etiquetas moleculares secundarias de una o más etiquetas moleculares principales, y en donde la aparición de la etiqueta molecular original es mayor o igual que un umbral de ocurrencia de adyacencia direccional predeterminada.

6. El método de la reivindicación 5, en el que el umbral de aparición de adyacencia direccional predeterminado es el doble de la aparición de una etiqueta molecular secundaria menos uno.

7. El método de cualquiera de las reivindicaciones 1-6, en el que colapsar los datos de secuenciación obtenidos en (b) usando los grupos de las etiquetas moleculares de la diana identificada en (ii) comprenden:
atribuir la aparición de la etiqueta molecular secundaria a la etiqueta molecular principal.

8. El método de cualquiera de las reivindicaciones 1-7, que comprende además:
determinar una profundidad de secuenciación del objetivo para que esté por encima de un umbral de profundidad de secuenciación predeterminado, opcionalmente en el que el umbral de profundidad de secuenciación predeterminado está entre 15 y 20.

9. El método de cualquiera de las reivindicaciones 1-8, en el que las dos distribuciones binomiales negativas son una primera distribución binomial negativa para las etiquetas moleculares verdaderas y una segunda distribución binomial negativa para las etiquetas moleculares falsas.

10. El método de cualquiera de las reivindicaciones 1-9, en el que la pluralidad de objetivos está comprendida en una muestra que comprende una o más células, en donde la una o más células comprenden una célula cerebral, una célula cardíaca, una célula cancerosa, una célula tumoral circulante, una célula de órgano, una célula epitelial, una célula metastásica, una célula benigna, una célula primaria, una célula circulatoria o cualquier combinación de las mismas, en donde opcionalmente la muestra comprende una sola célula.

11. El método de cualquiera de las reivindicaciones 1-10, en el que la pluralidad de dianas comprende ácido desoxirribonucleico (ADN), ácidos ribonucleicos (ARN), ARN mensajeros (ARNm), microARN, pequeños ARN interferentes (ARNip), productos de degradación de ARN, ARN que comprenden cada uno una cola poli(A), o cualquier combinación de los mismos.

12. Un sistema informático para determinar el número de objetivos que comprende:

un procesador de hardware; y
memoria no transitoria que tiene instrucciones almacenadas en ella, que cuando las ejecuta el procesador de hardware hacen que el procesador realice el método de cualquiera de las reivindicaciones 1 a 11.

13. Un medio legible por computadora que comprende un programa de software que comprende código para realizar el método de cualquiera de las reivindicaciones 1-11.

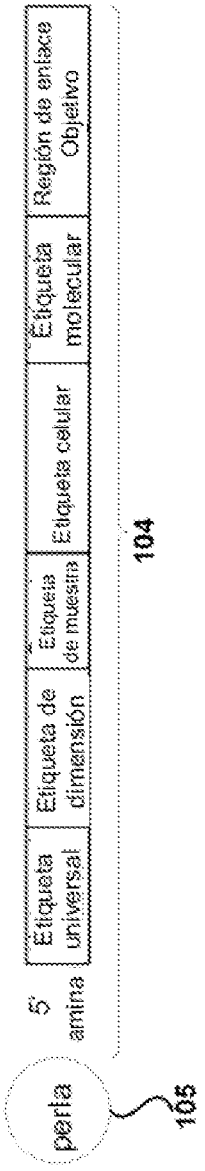

perla
105
5'
amina
Etiqueta universal
Etiqueta de dilucidación
Etiqueta de muestra
Etiqueta celular
Etiqueta molecular
Región de enlace
Objetivo
104

FIG. 1

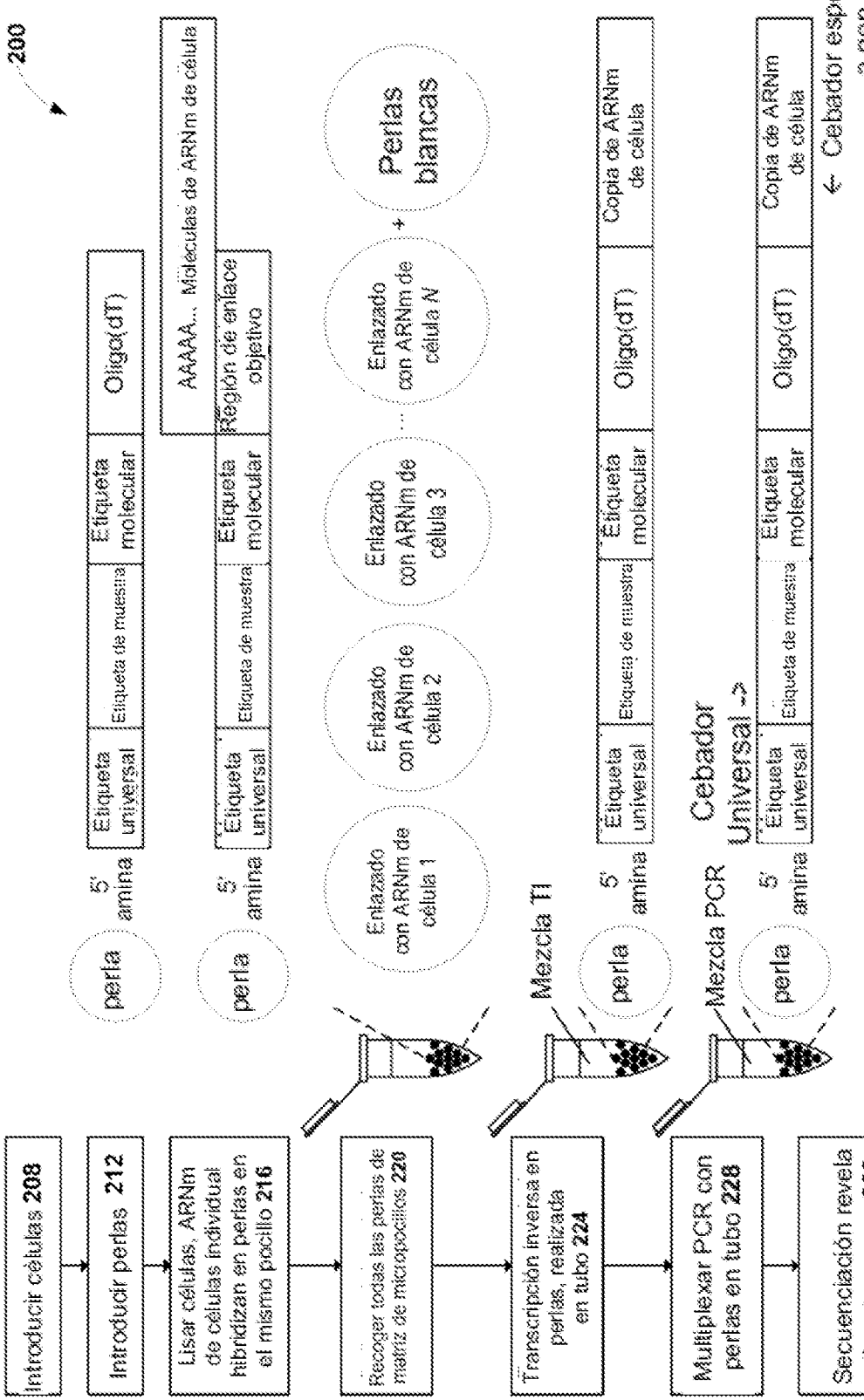

200
Introducir células 208
Introducir perlas 212
Lisar células, ARNm de células individual hibridizan en perlas en el mismo pocillo 216
Recoger todas las perlas de matriz de microposillos 226
Transcripción inversa en perlas, realizada en tubo 224
Multiplexar PCR con perlas en tubo 228
Secuenciación revela etiquetas y gen 232
perla
5' amina
Etiqueta universal
Etiqueta de muestra
Etiqueta molecular
Oligo(dT)
AAAAA... Moléculas de ARNm de célula
Región de enlace objetivo
Enlazado con ARNm de célula 1
Enlazado con ARNm de célula 2
Enlazado con ARNm de célula 3
Enlazado con ARNm de célula N
Perlas blancas
Mezcla T1
Copia de ARNm de célula
Cebador Universal ->
Mezcla PCR
← Cebador específico a gen

FIG. 2

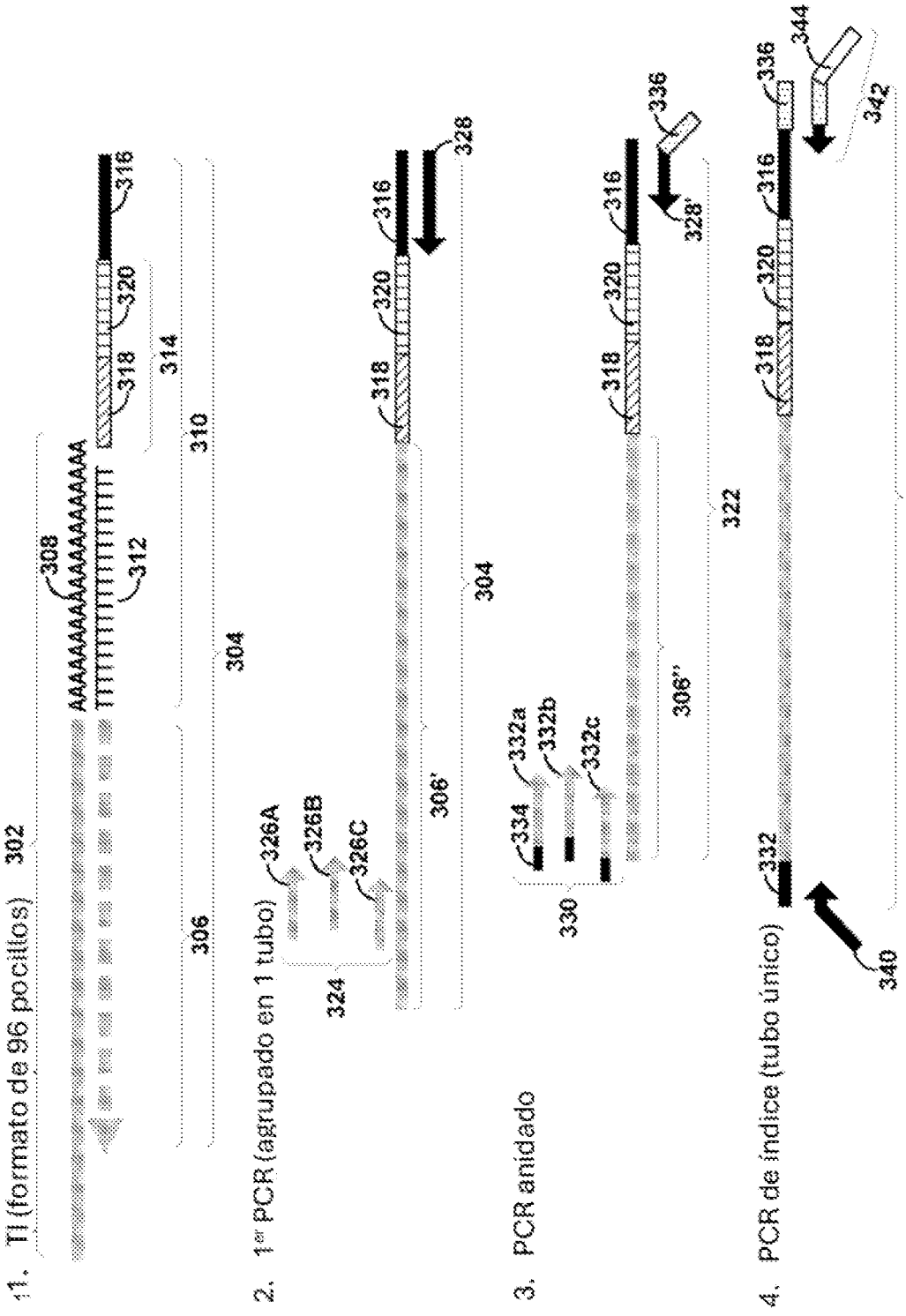

1. TI (formato de 96 pocillos) 302
308
AAAAAAAAAAAAAAAAAA
TTTTTTTTTTTTTTTTTT
312
316
320
318
314
310
306
304
2. 1er PCR (agrupado en 1 tubo)
326A
326B
326C
324
306'
304
318
320
316
328
3. PCR anidado
334
332a
332b
332c
330
306''
322
318
320
316
336
328'
4. PCR de índice (tubo único)
332
340
338
318
320
316
336
344
342

FIG. 3

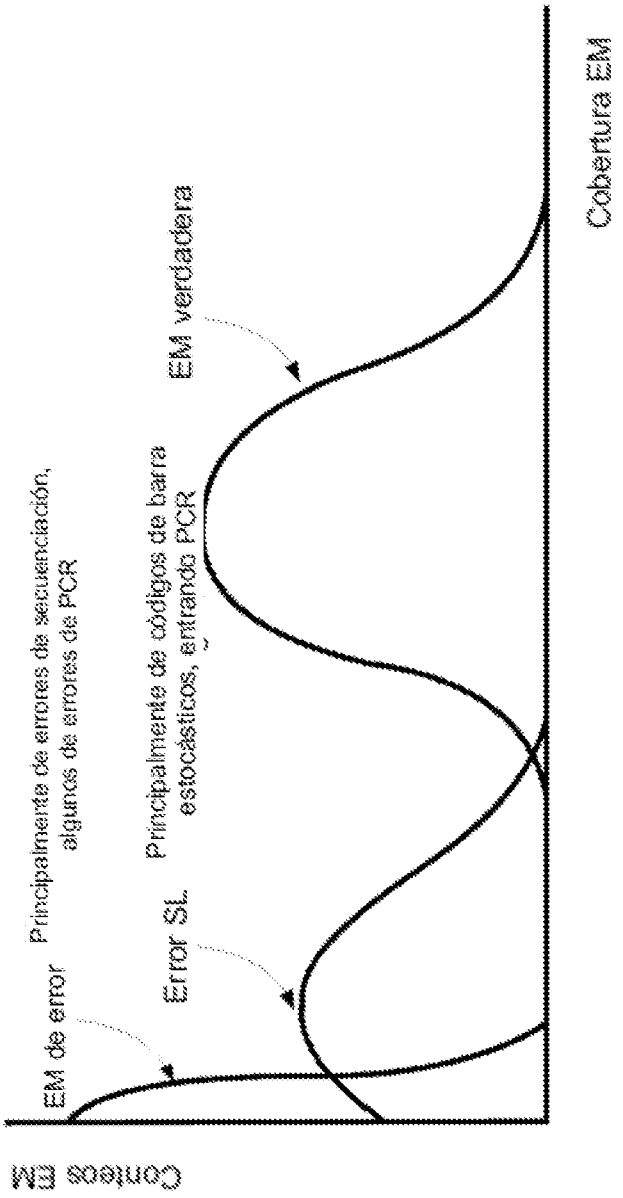

Conteos EM
EM de error
Principalmente de errores de secuenciación, algunos de errores de PCR
Error SL
Principalmente de códigos de barra estocásticos, entrando PCR
EM verdadera
Cobertura EM

FIG. 4

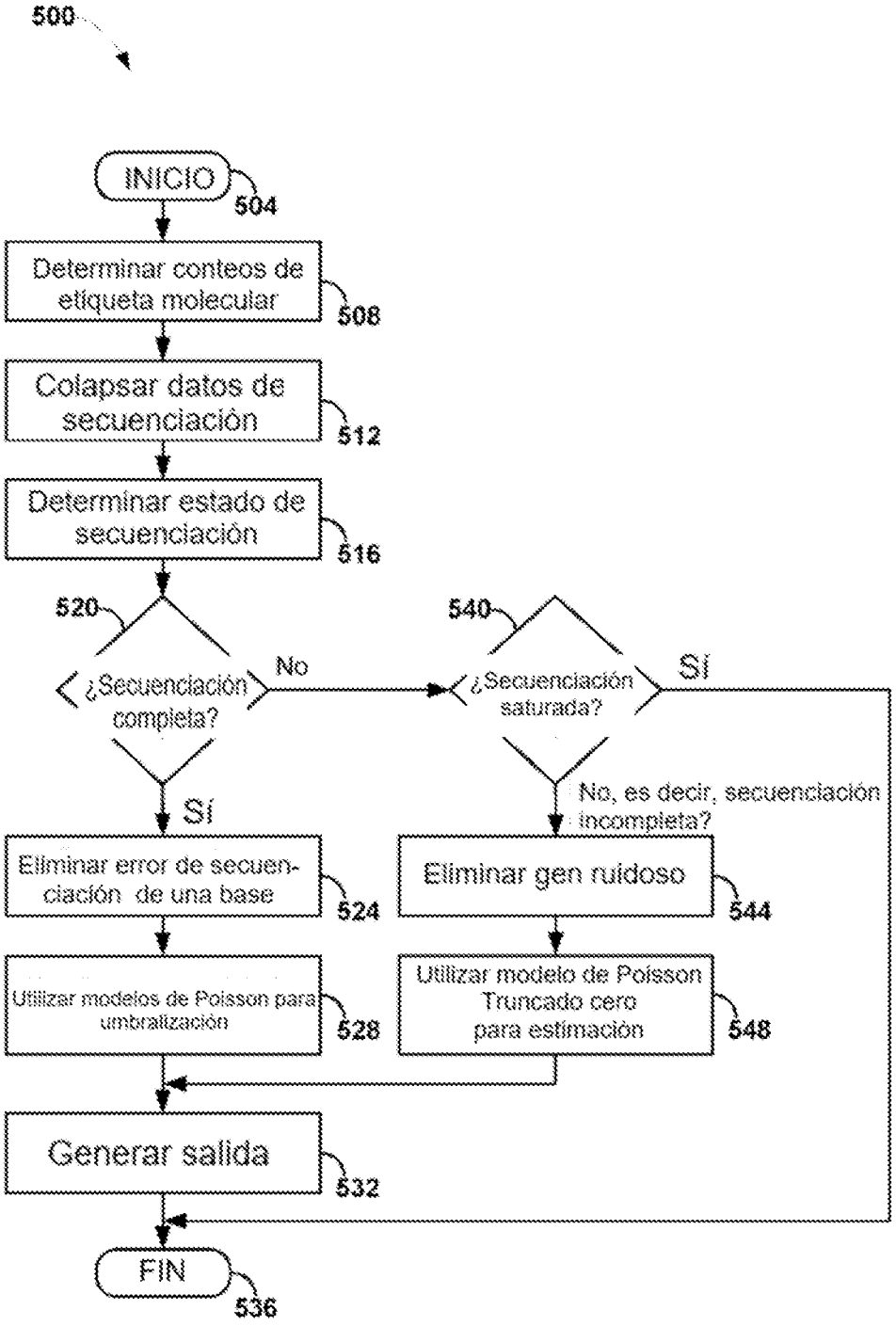

500
INICIO
504
Determinar conteos de etiqueta molecular
508
Colapsar datos de secuenciación
512
Determinar estado de secuenciación
516
520
¿Secuenciación completa?
No
Sí
540
¿Secuenciación saturada?
Sí
No, es decir, secuenciación incompleta?
Eliminar error de secuenciación de una base
524
Eliminar gen ruidoso
544
Utilizar modelos de Poisson para umbralización
528
Utilizar modelo de Poisson Truncado cero para estimación
548
Generar salida
532
FIN
536

FIG. 5

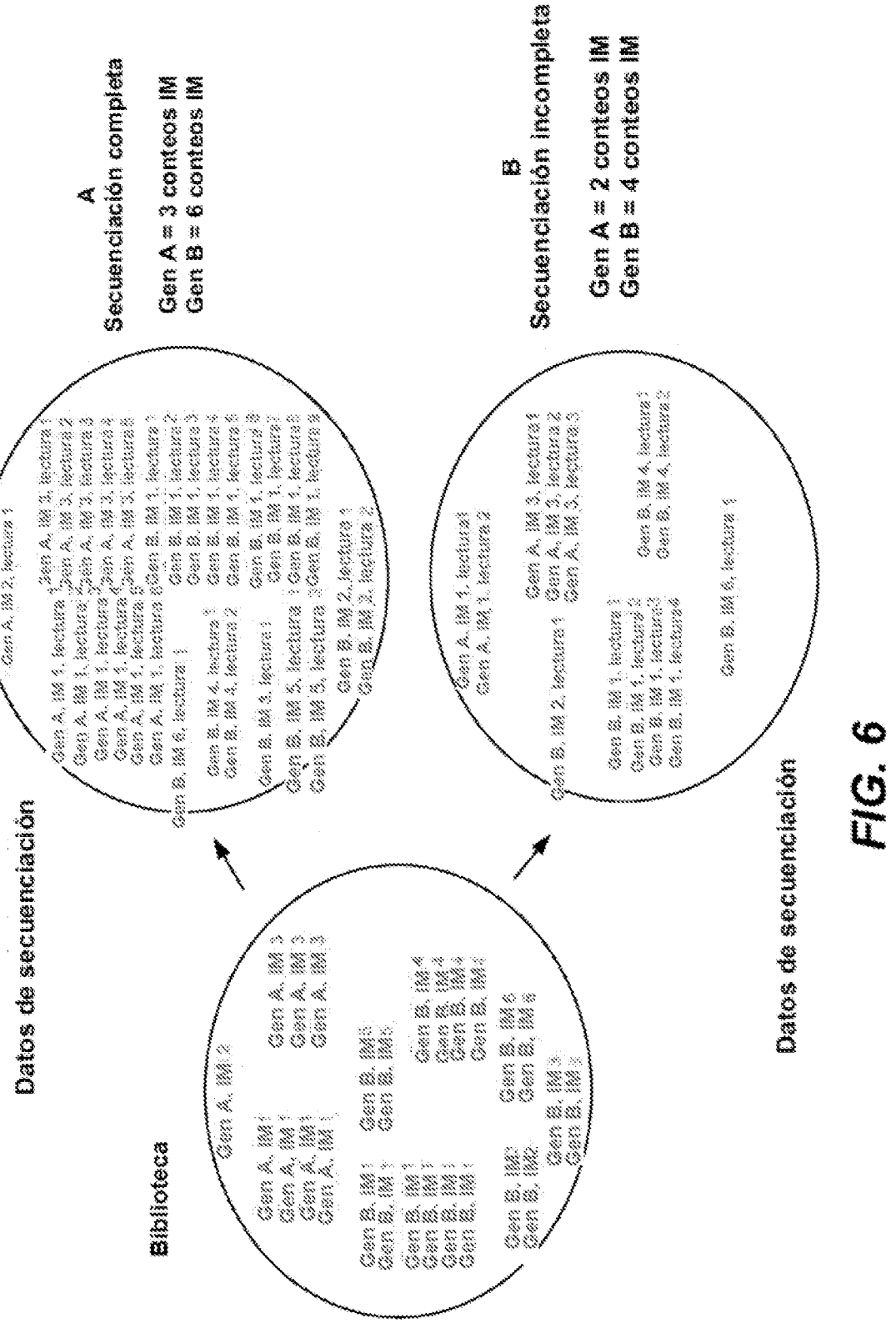

Datos de secuenciación
Biblioteca
A
Secuenciación completa
Gen A = 3 conteos IM
Gen B = 6 conteos IM
B
Secuenciación incompleta
Gen A = 2 conteos IM
Gen B = 4 conteos IM
Datos de secuenciación

FIG. 6

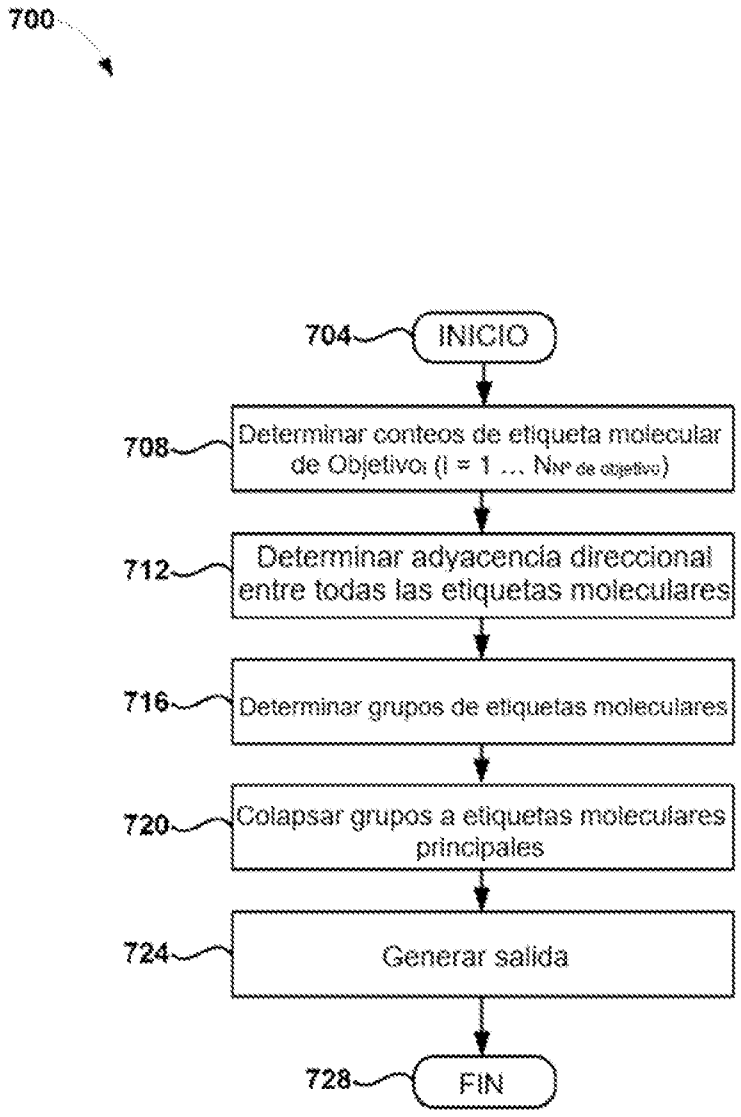

700
704
INICIO
708
Determinar conteos de etiqueta molecular de Objetivo i (i = 1 … N° de objetivo)
712
Determinar adyacencia direccional entre todas las etiquetas moleculares
716
Determinar grupos de etiquetas moleculares
720
Colapsar grupos a etiquetas moleculares principales
724
Generar salida
728
FIN

FIG. 7

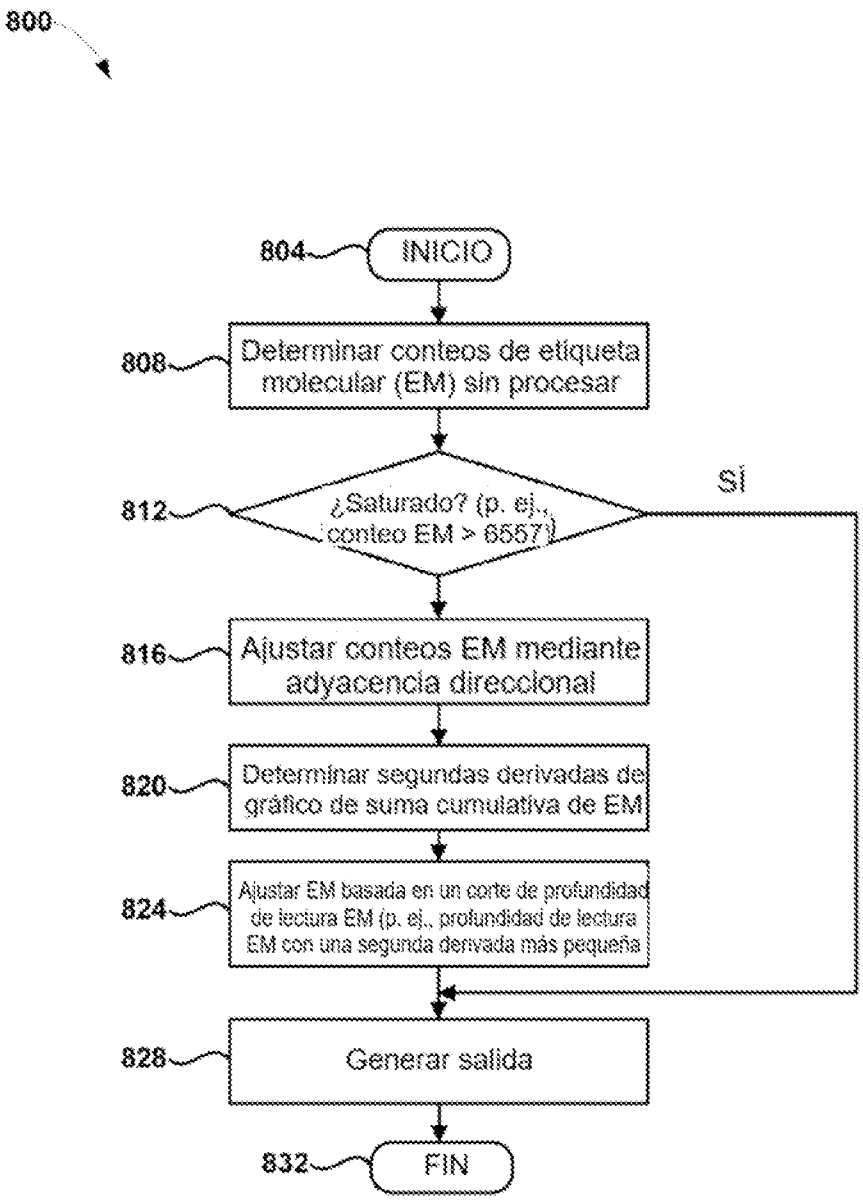

800
804
INICIO
808
Determinar conteos de etiqueta molecular (EM) sin procesar
812
¿Saturado? (p. ej., conteo EM > 6557)
SÍ
816
Ajustar conteos EM mediante adyacencia direccional
820
Determinar segundas derivadas de gráfico de suma cumulativa de EM
824
Ajustar EM basada en un corte de profundidad de lectura EM (p. ej., profundidad de lectura EM con una segunda derivada más pequeña
828
Generar salida
832
FIN

FIG. 8

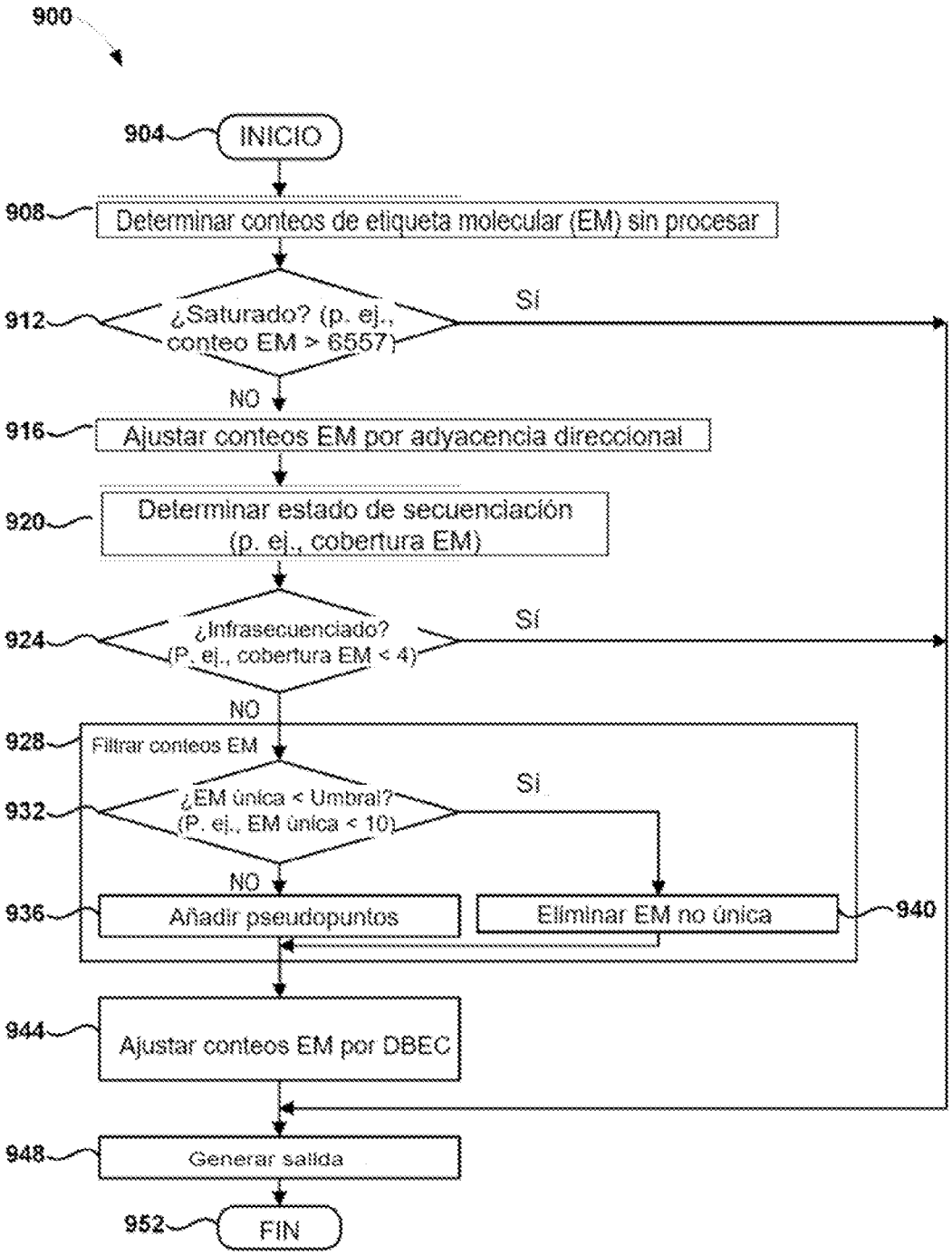

900
904
INICIO
908
Determinar conteos de etiqueta molecular (EM) sin procesar
912
¿Saturado? (p. ej., conteo EM > 6557)
Sí
NO
916
Ajustar conteos EM por adyacencia direccional
920
Determinar estado de secuenciación (p. ej., cobertura EM)
924
¿Infrasecuenciado? (P. ej., cobertura EM < 4)
Sí
NO
928
Filtrar conteos EM
932
¿EM única < Umbral? (P. ej., EM única < 10)
Sí
NO
936
Añadir pseudopuntos
Eliminar EM no única
940
944
Ajustar conteos EM por DBEC
948
Generar salida
952
FIN

FIG. 9

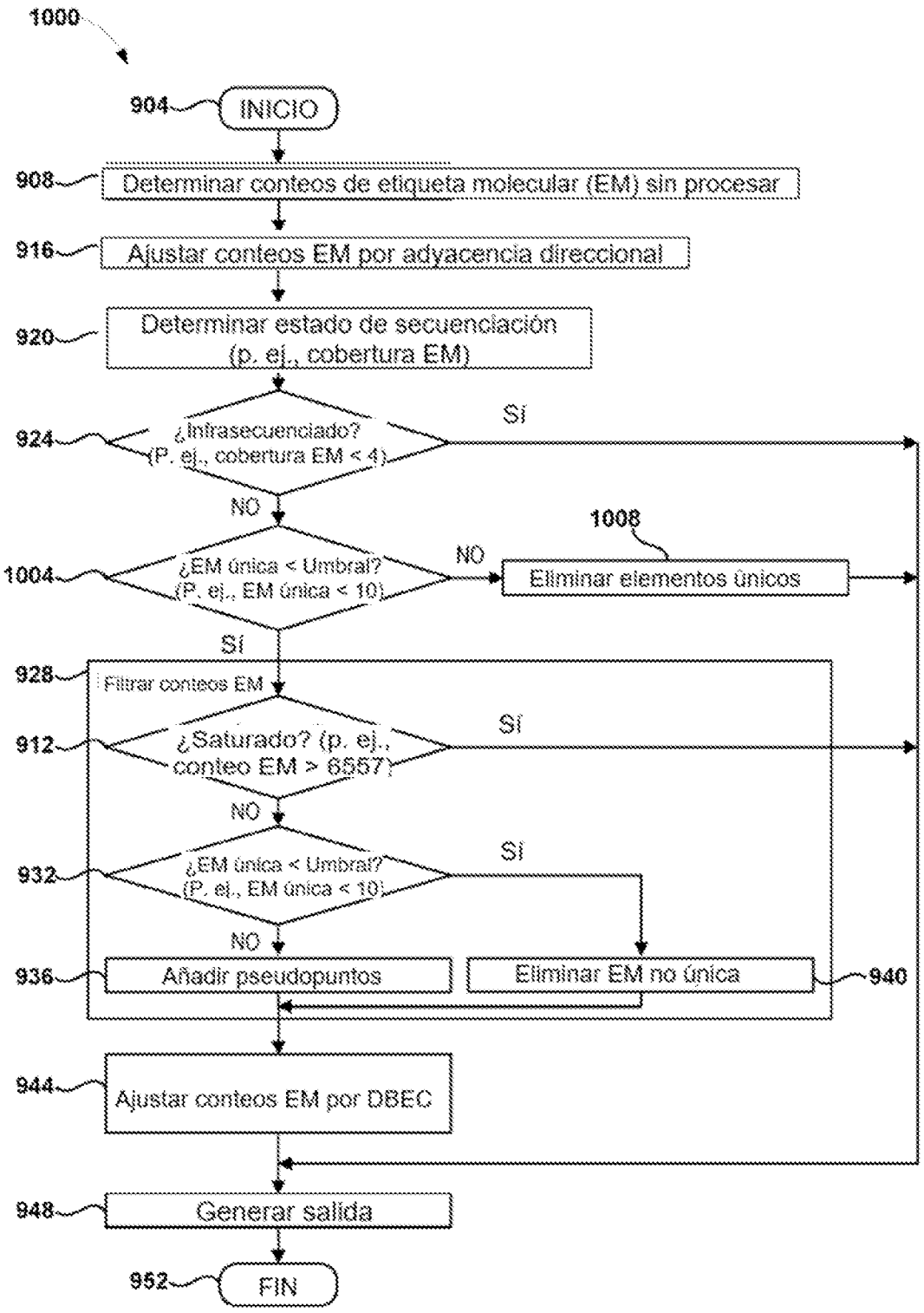

1000
904
INICIO
908
Determinar conteos de etiqueta molecular (EM) sin procesar
916
Ajustar conteos EM por adyacencia direccional
920
Determinar estado de secuenciación (p. ej., cobertura EM)
924
¿Infrasecuenciado? (P. ej., cobertura EM < 4)
Sí
NO
1008
1004
¿EM única < Umbral? (P. ej., EM única < 10)
NO
Eliminar elementos únicos
Sí
928
Filtrar conteos EM
912
¿Saturado? (p. ej., conteo EM > 6557)
Sí
NO
932
¿EM única < Umbral? (P. ej., EM única < 10)
Sí
NO
936
Añadir pseudopuntos
Eliminar EM no única
940
944
Ajustar conteos EM por DBEC
948
Generar salida
952
FIN

FIG. 10

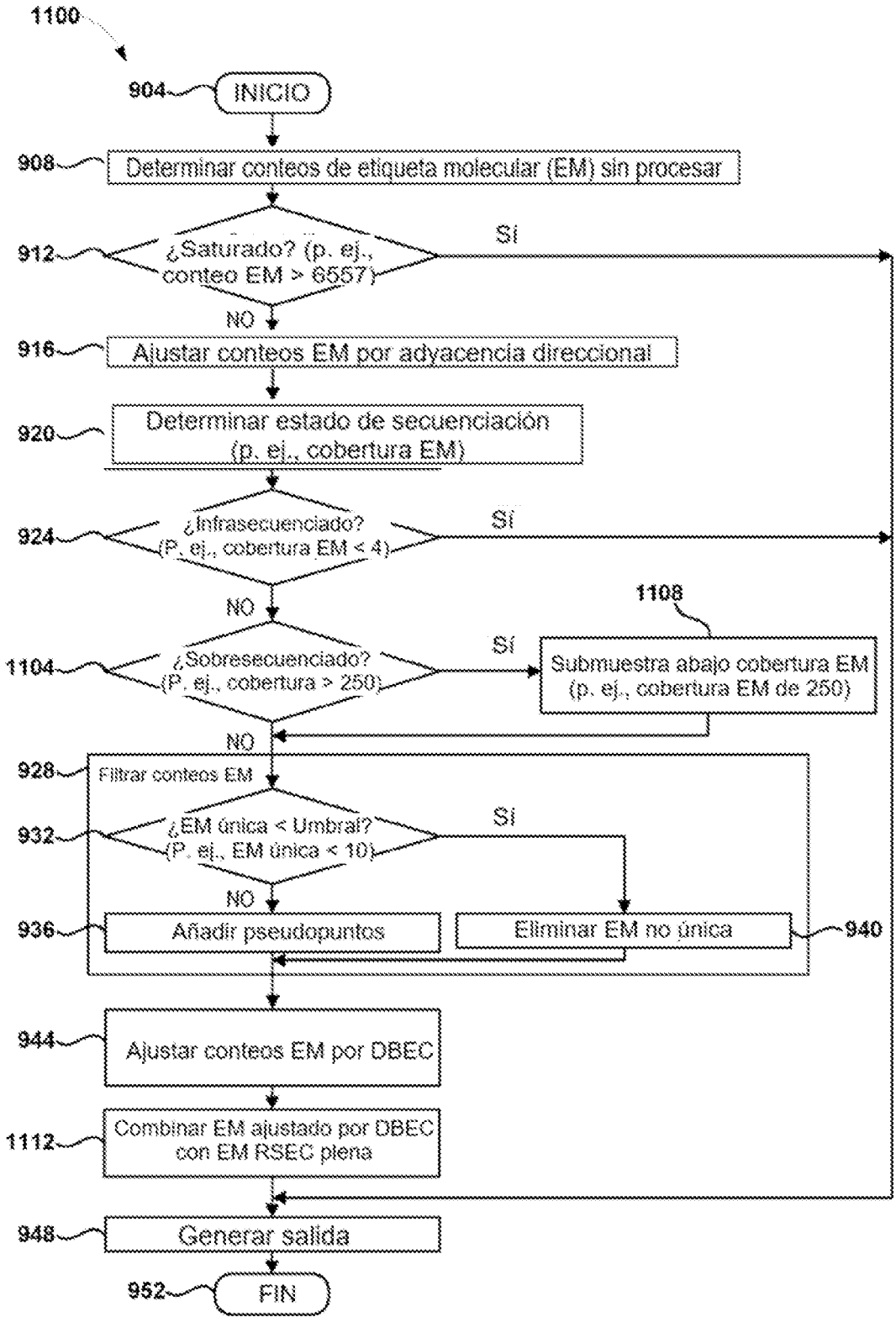

1100
904
INICIO
908
Determinar conteos de etiqueta molecular (EM) sin procesar
912
¿Saturado? (p. ej., conteo EM > 6557)
Sí
NO
916
Ajustar conteos EM por adyacencia direccional
920
Determinar estado de secuenciación (p. ej., cobertura EM)
924
¿Infrasecuenciado? (P. ej., cobertura EM < 4)
Sí
NO
1108
1104
¿Sobresecuenciado? (P. ej., cobertura > 250)
Sí
Submuestra abajo cobertura EM (p. ej., cobertura EM de 250)
NO
928
Filtrar conteos EM
932
¿EM única < Umbral? (P. ej., EM única < 10)
Sí
NO
936
Añadir pseudopuntos
Eliminar EM no única
940
944
Ajustar conteos EM por DBEC
1112
Combinar EM ajustado por DBEC con EM RSEC plena
948
Generar salida
952
FIN

FIG. 11

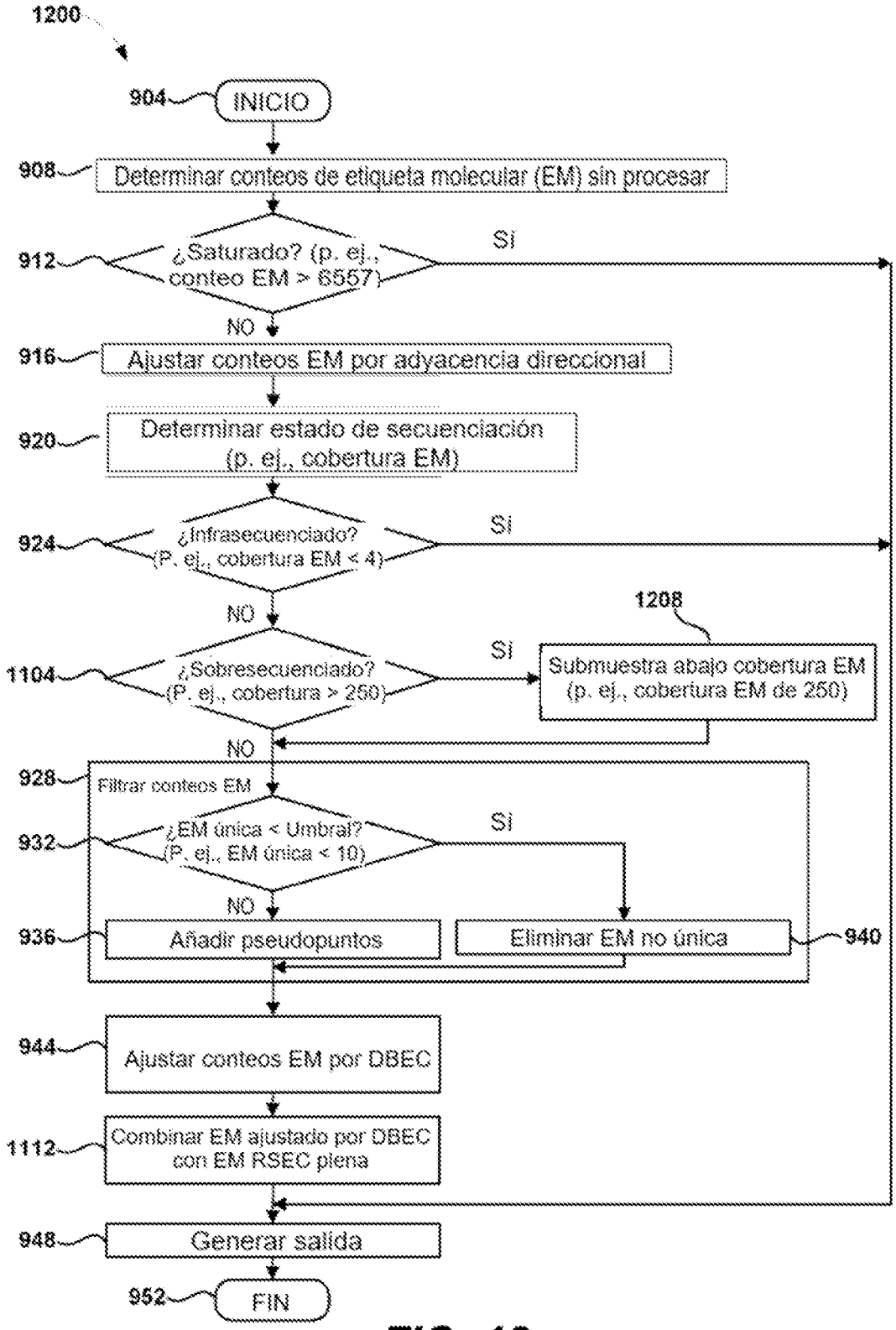

1200
904
INICIO
908
Determinar conteos de etiqueta molecular (EM) sin procesar
912
¿Saturado? (p. ej., conteo EM > 6557)
Sí
NO
916
Ajustar conteos EM por adyacencia direccional
920
Determinar estado de secuenciación (p. ej., cobertura EM)
924
¿Infrasecuenciado? (P. ej., cobertura EM < 4)
Sí
NO
1208
1104
¿Sobresecuenciado? (P. ej., cobertura > 250)
Sí
Submuestra abajo cobertura EM (p. ej., cobertura EM de 250)
NO
928
Filtrar conteos EM
932
¿EM única < Umbral? (P. ej., EM única < 10)
Sí
NO
936
Añadir pseudopuntos
Eliminar EM no única
940
944
Ajustar conteos EM por DBEC
1112
Combinar EM ajustado por DBEC con EM RSEC plena
948
Generar salida
952
FIN

FIG. 12

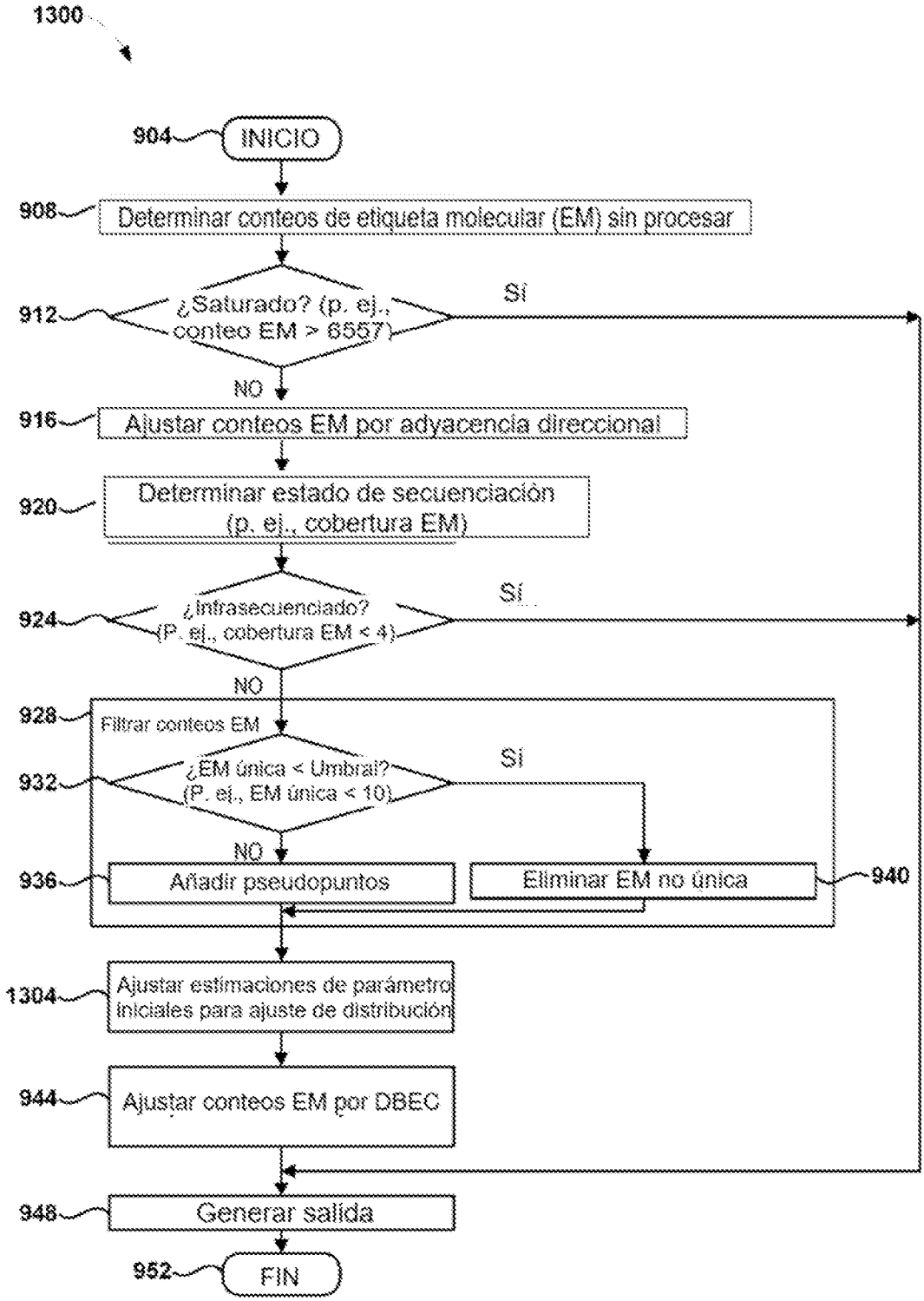

1300
904
INICIO
908
Determinar conteos de etiqueta molecular (EM) sin procesar
912
¿Saturado? (p. ej., conteo EM > 6557)
Sí
NO
916
Ajustar conteos EM por adyacencia direccional
920
Determinar estado de secuenciación (p. ej., cobertura EM)
924
¿Infrasecuenciado? (P. ej., cobertura EM < 4)
Sí
NO
928
Filtrar conteos EM
932
¿EM única < Umbral? (P. ej., EM única < 10)
Sí
NO
936
Añadir pseudopuntos
Eliminar EM no única
940
1304
Ajustar estimaciones de parámetro iniciales para ajuste de distribución
944
Ajustar conteos EM por DBEC
948
Generar salida
952
FIN

FIG. 13

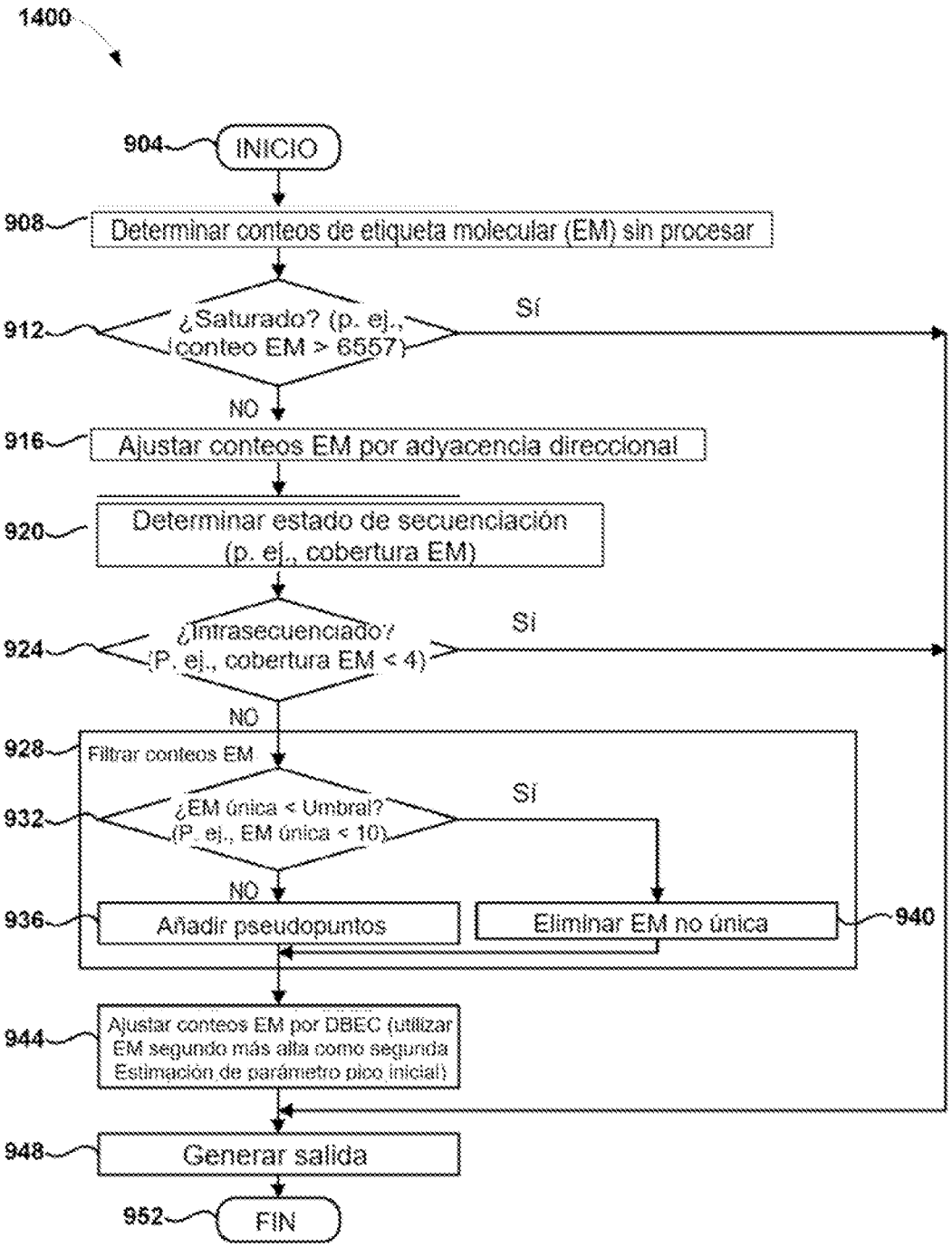

1400
904
INICIO
908
Determinar conteos de etiqueta molecular (EM) sin procesar
912
¿Saturado? (p. ej., conteo EM > 6557)
Sí
NO
916
Ajustar conteos EM por adyacencia direccional
920
Determinar estado de secuenciación (p. ej., cobertura EM)
924
¿Infrasecuenciado? (P. ej., cobertura EM < 4)
Sí
NO
928
Filtrar conteos EM
932
¿EM única < Umbral? (P. ej., EM única < 10)
Sí
NO
936
Añadir pseudopuntos
Eliminar EM no única
940
944
Ajustar conteos EM por DBEC (utilizar EM segundo más alta como segunda Estimación de parámetro pico inicial)
948
Generar salida
952
FIN

FIG. 14

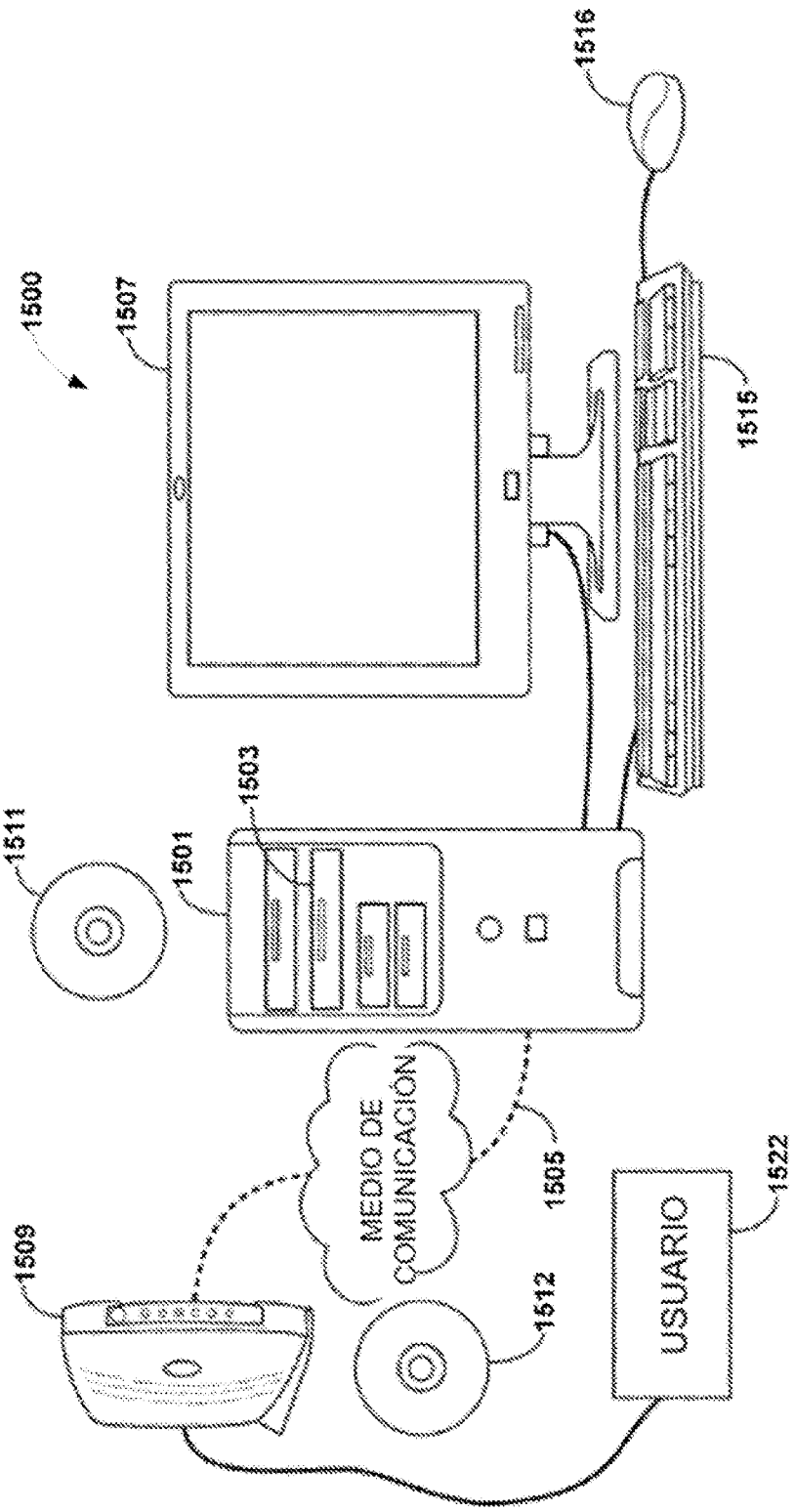

1500
1507
1516
1515
1503
1511
1501
MEDIO DE COMUNICACIÓN
1505
1509
1512
USUARIO
1522

FIG. 15

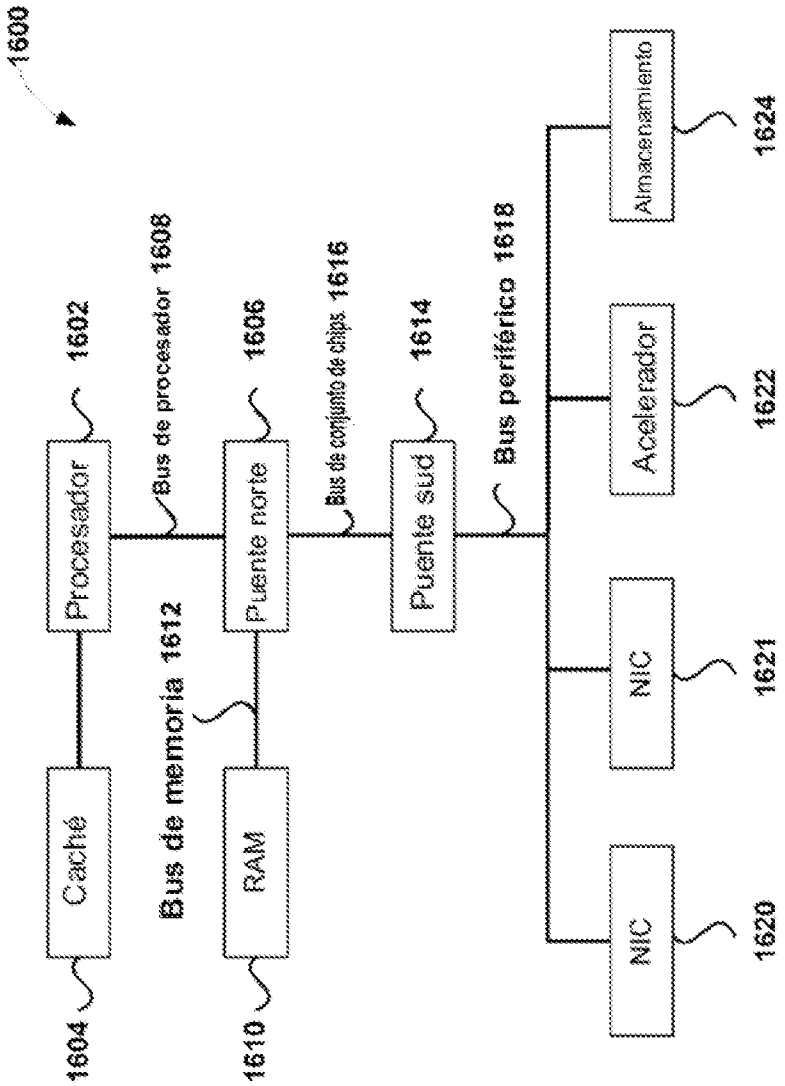

1600
Procesador
1602
Caché
1604
Bus de procesador 1608
Bus de memoria 1612
Puente norte
1606
RAM
1610
Bus de conjunto de chips 1616
Puente sud
1614
Bus periférico 1618
NIC
1620
NIC
1621
Acelerador
1622
Almacenamiento
1624

FIG. 16

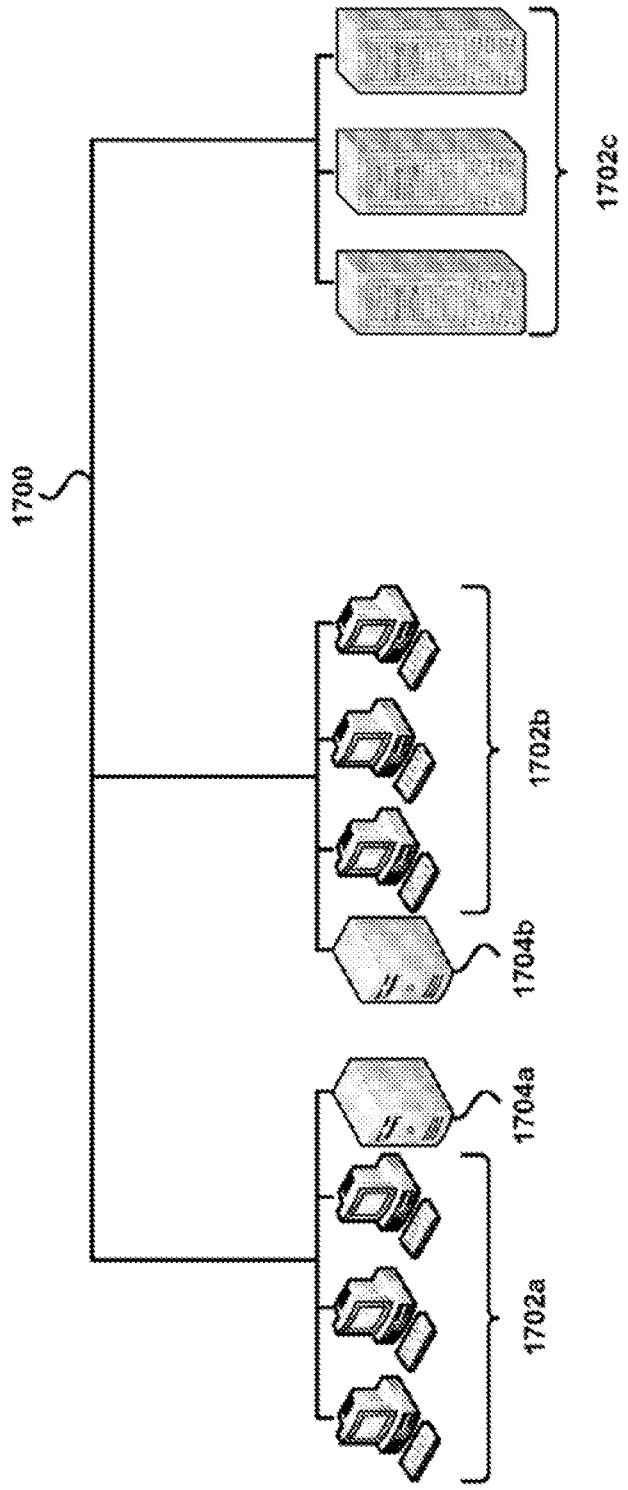


FIG. 17

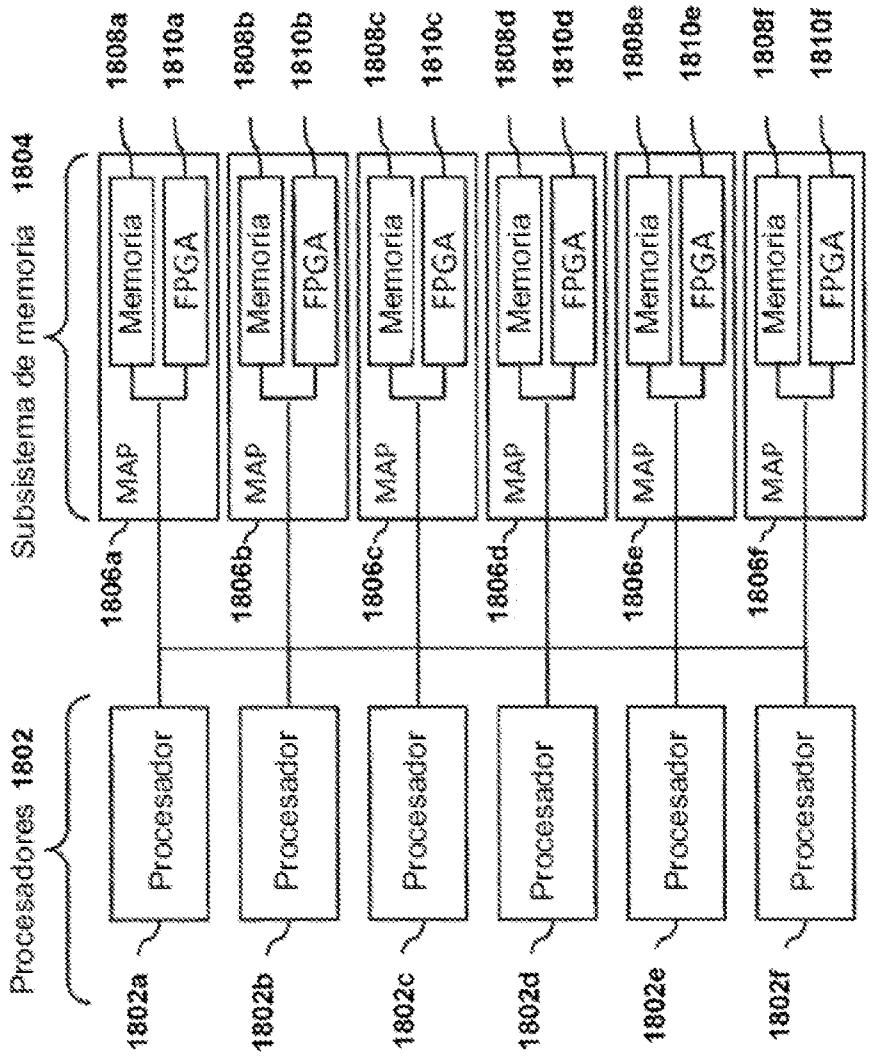

Procesadores 1802
Subsistema de memoria 1804
Procesador
MAP
Memoria
FPGA
1802a
1802b
1802c
1802d
1802e
1802f
1806a
1806b
1806c
1806d
1806e
1806f
1808a
1808b
1808c
1808d
1808e
1808f
1810a
1810b
1810c
1810d
1810e
1810f

FIG. 18

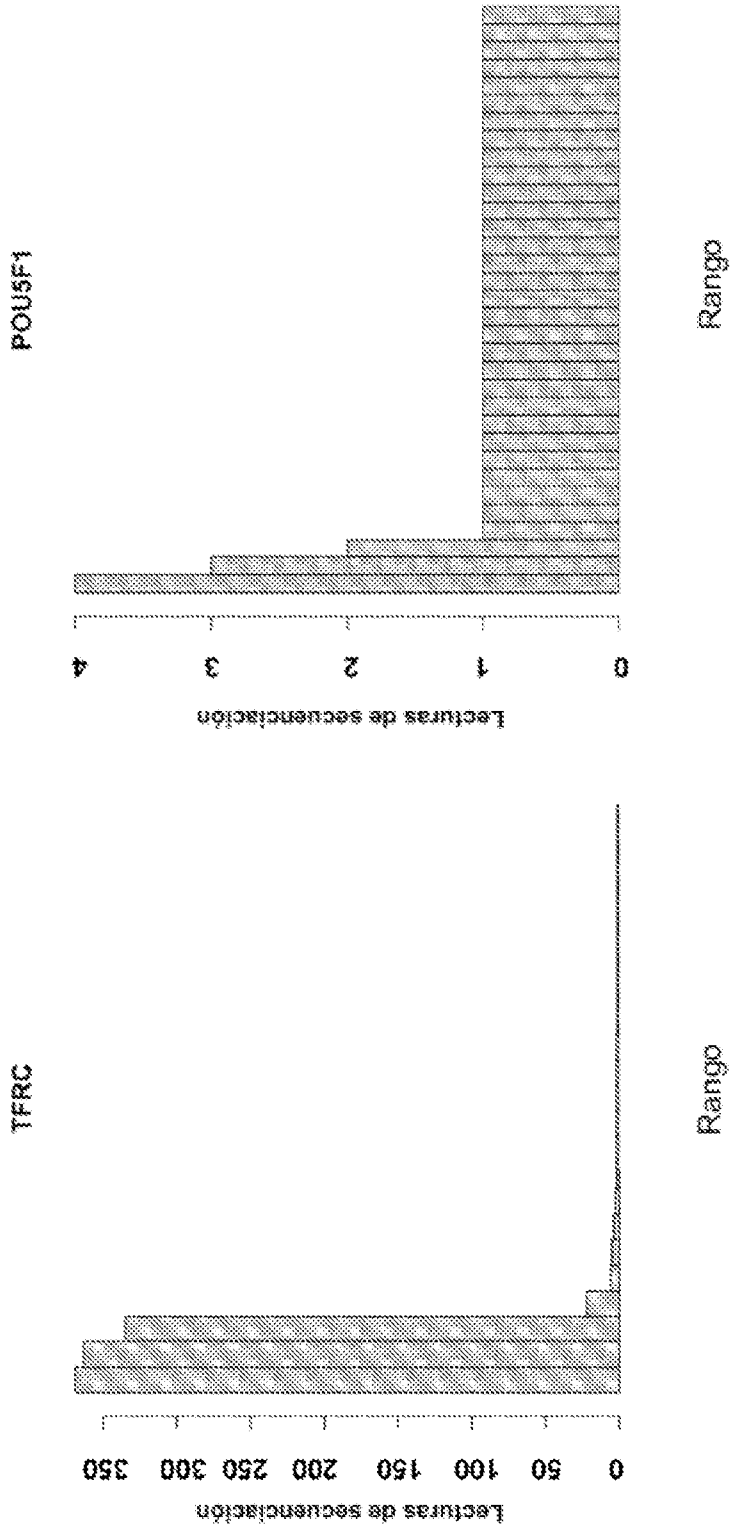


FIG. 19A

FIG. 19B

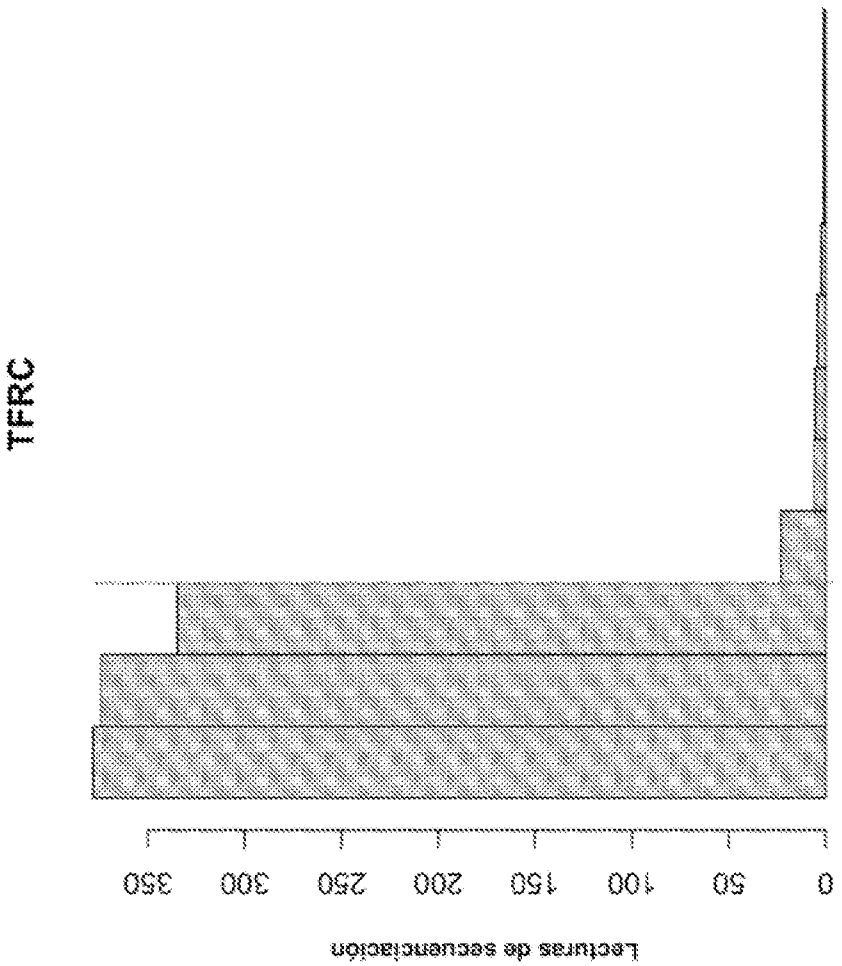

TFRC
Rango
0
50
100
150
200
250
300
350
Lecturas de secuenciación

FIG. 20

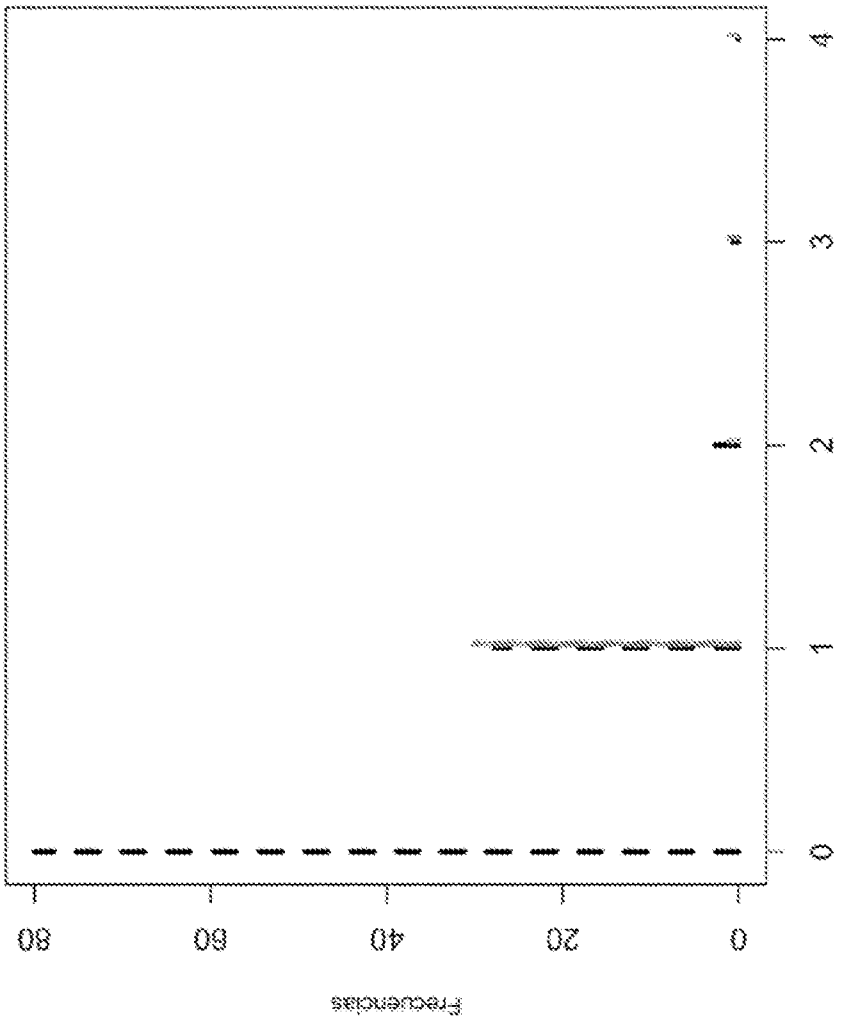
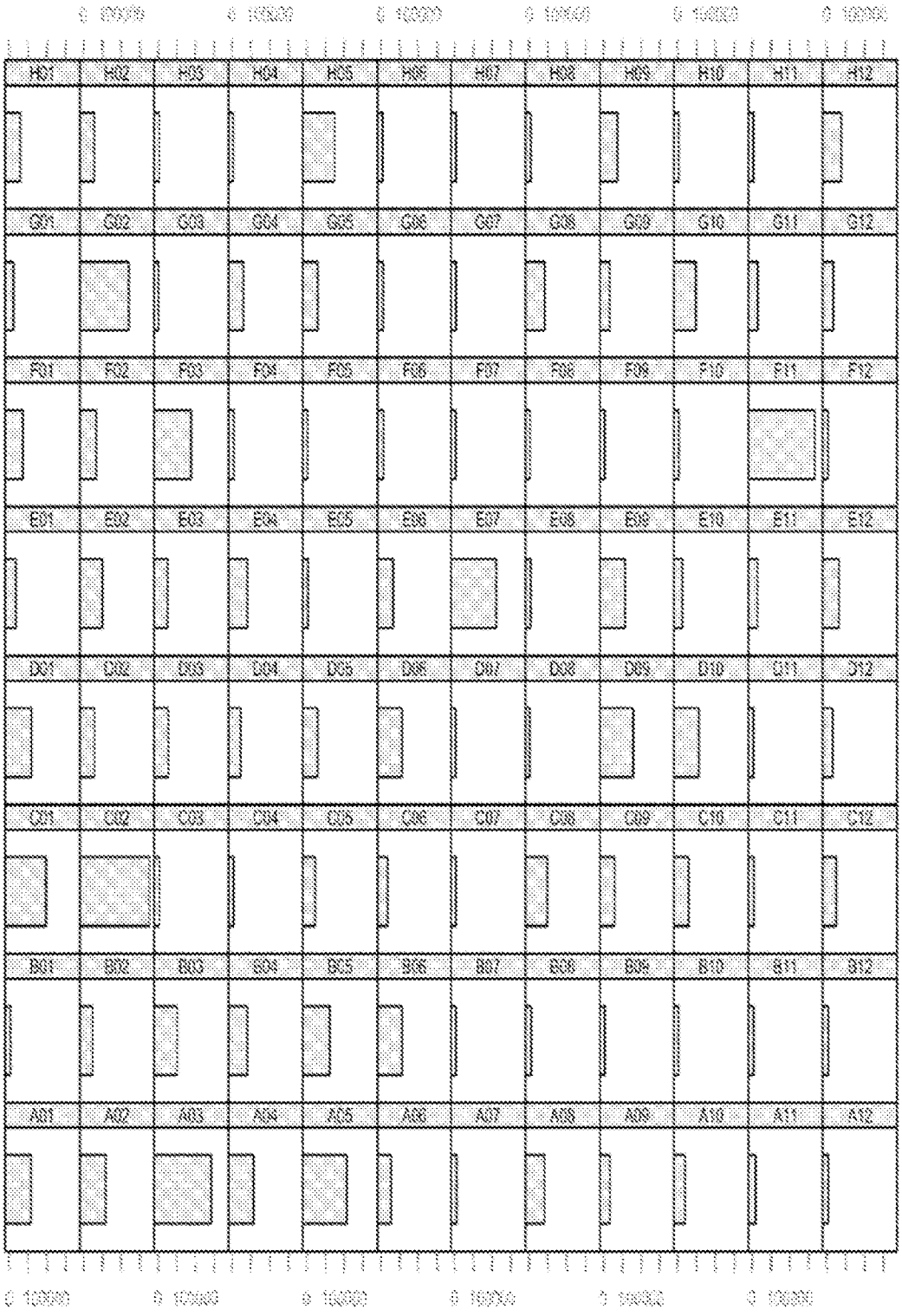


FIG. 21



Lecturas sin procesar

FIG. 22

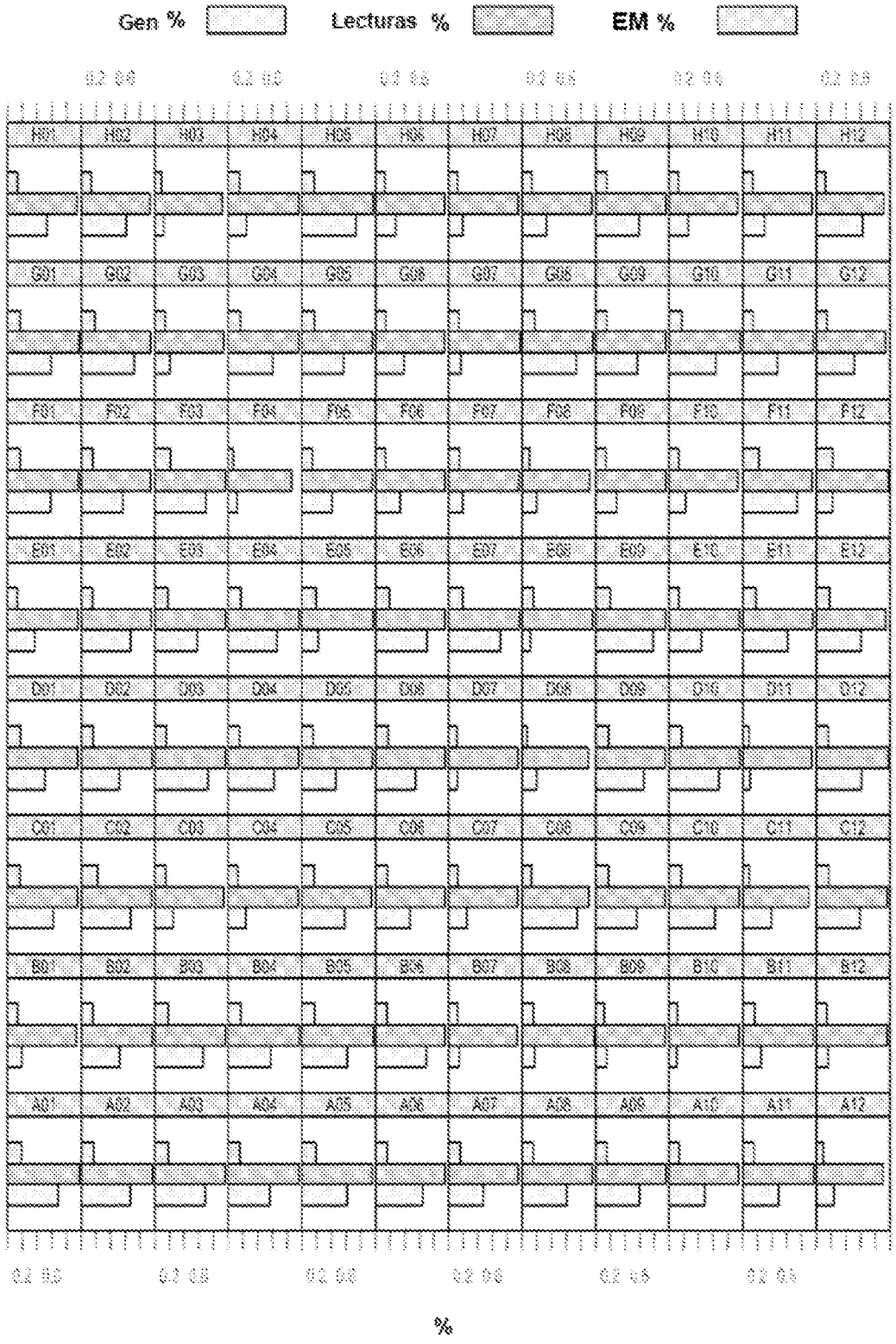


FIG. 23

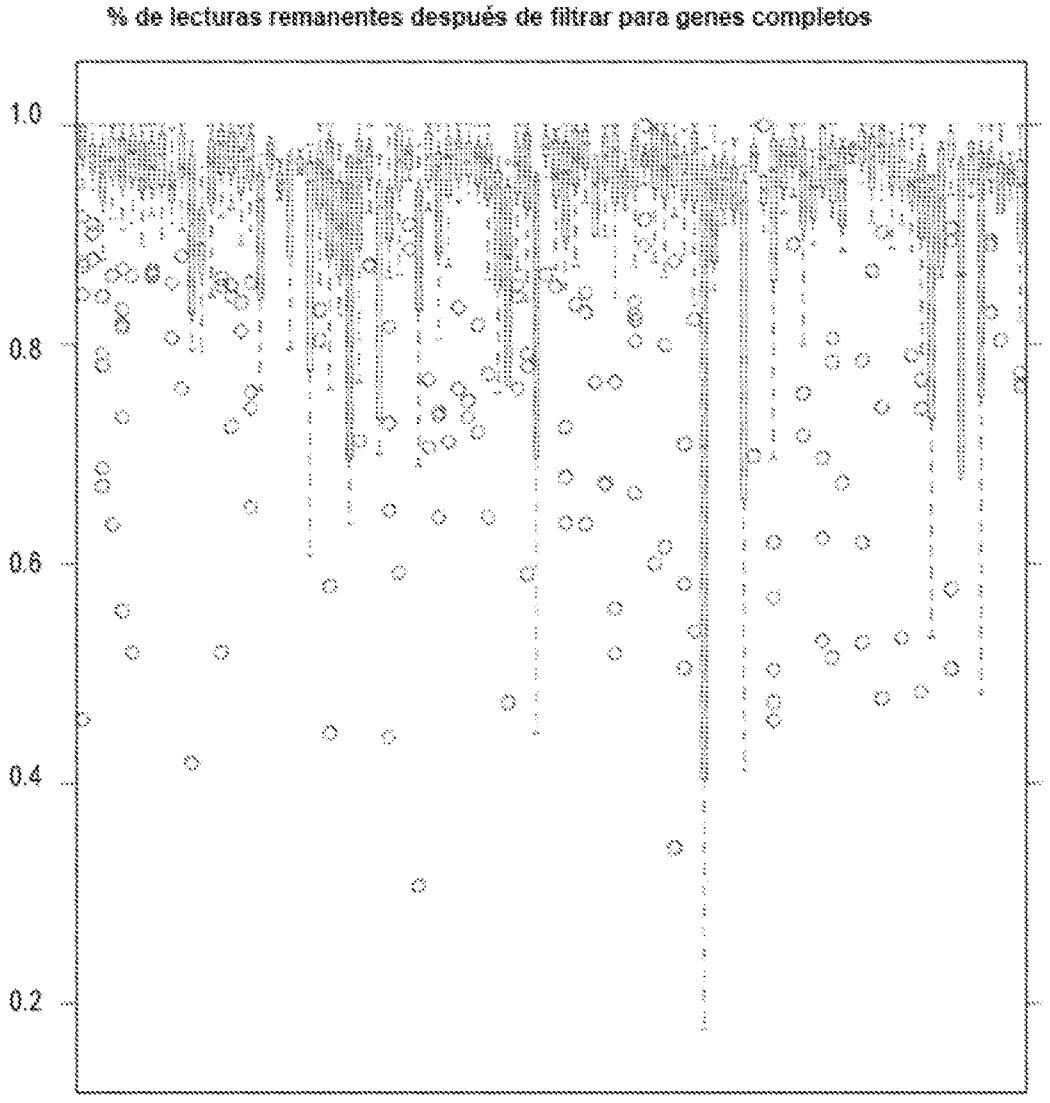

% de lecturas remanentes después de filtrar para genes completos
1.0
0.8
0.6
0.4
0.2

FIG. 24

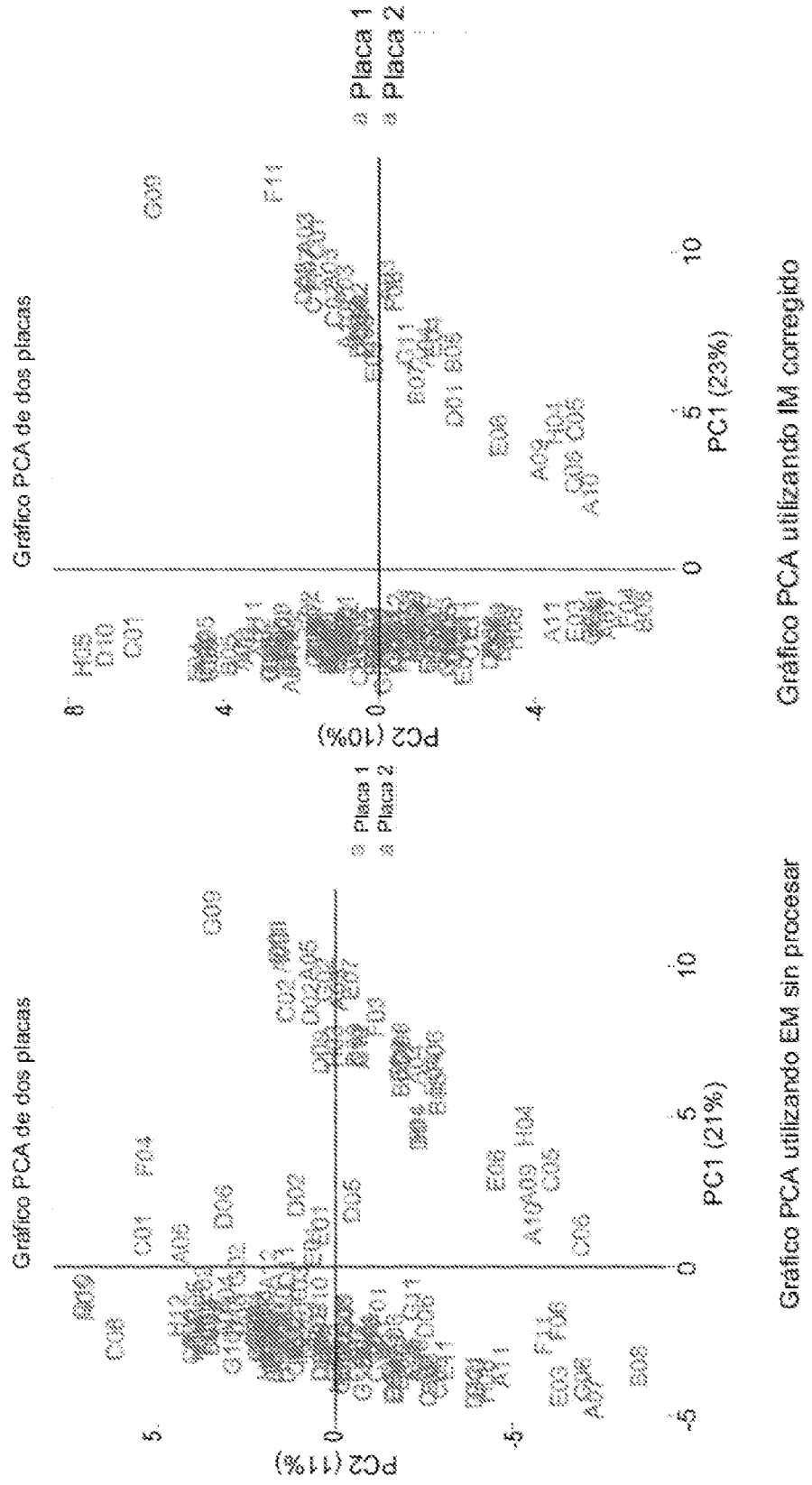


FIG. 25A

FIG. 25B

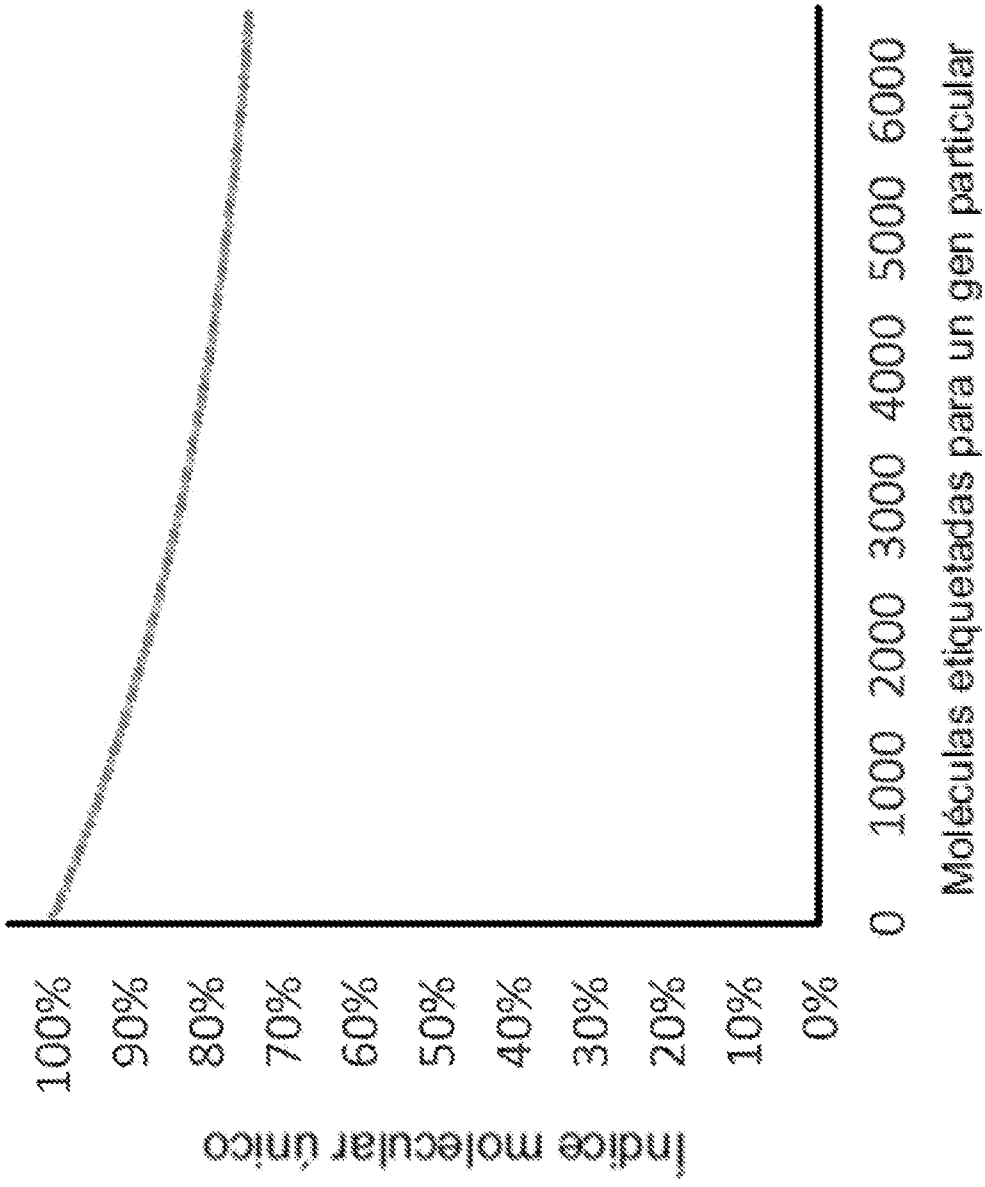

Índice molecular único
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
0
1000
2000
3000
4000
5000
6000
Moléculas etiquetadas para un gen particular

FIG. 26

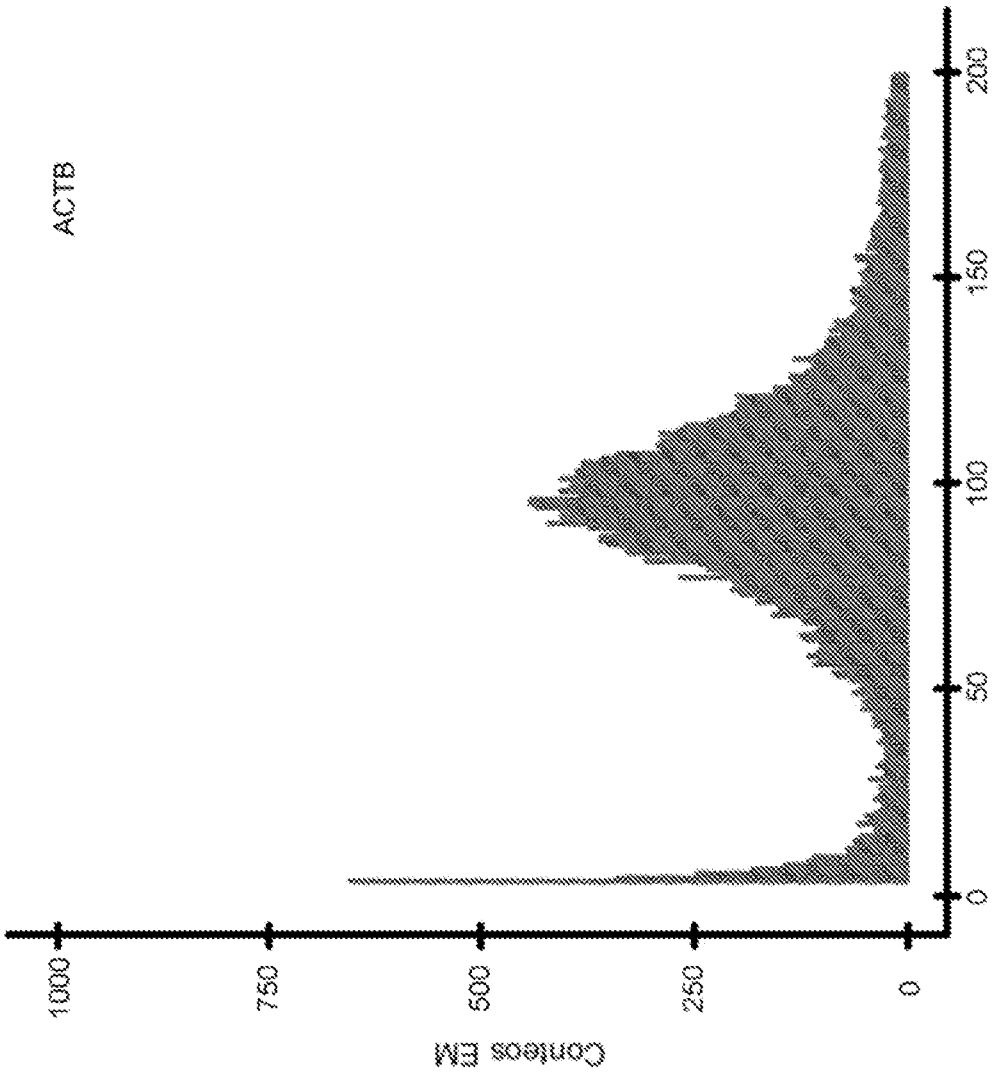

ACTB
Conteos EM
1000
750
500
250
0
0
50
100
150
200
Cobertura EM

FIG. 27

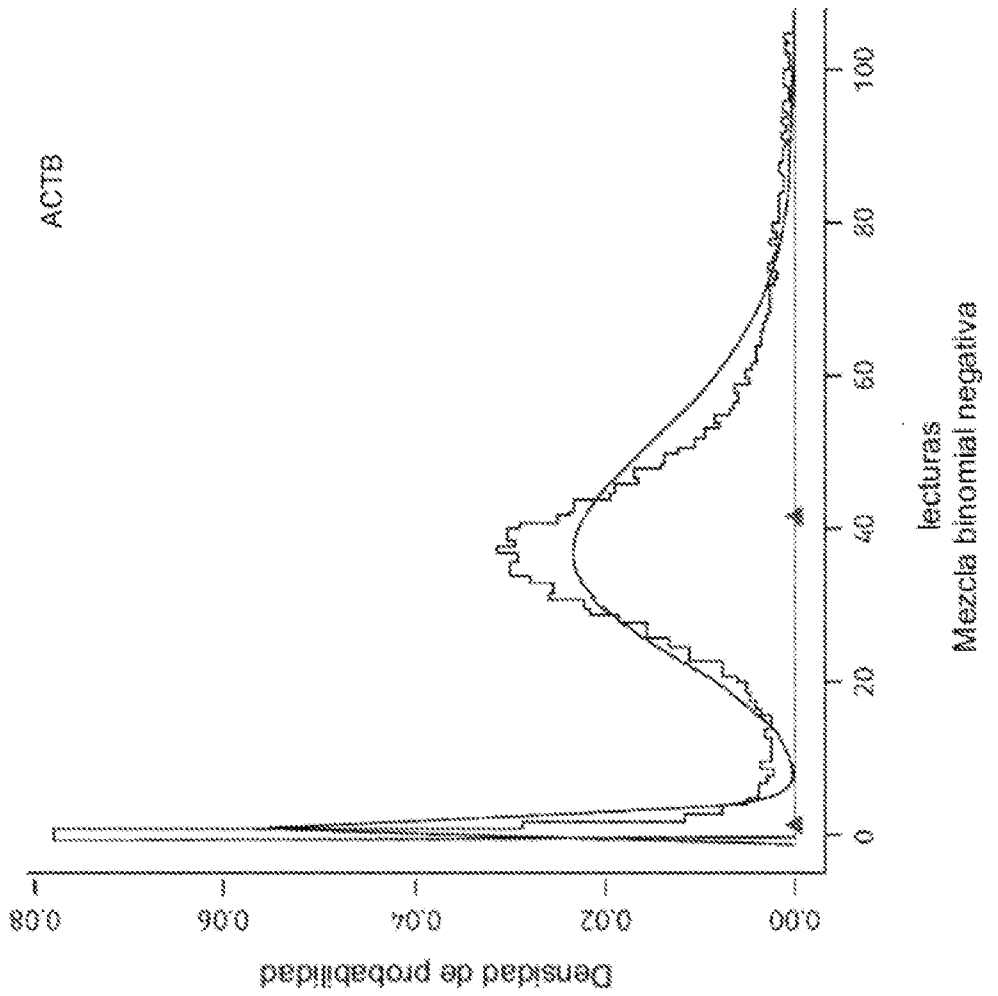

ACTB
Densidad de probabilidad
0,00
0,02
0,04
0,06
0,08
0
20
40
60
80
100
lecturas
Mezcla binomial negativa

FIG. 28

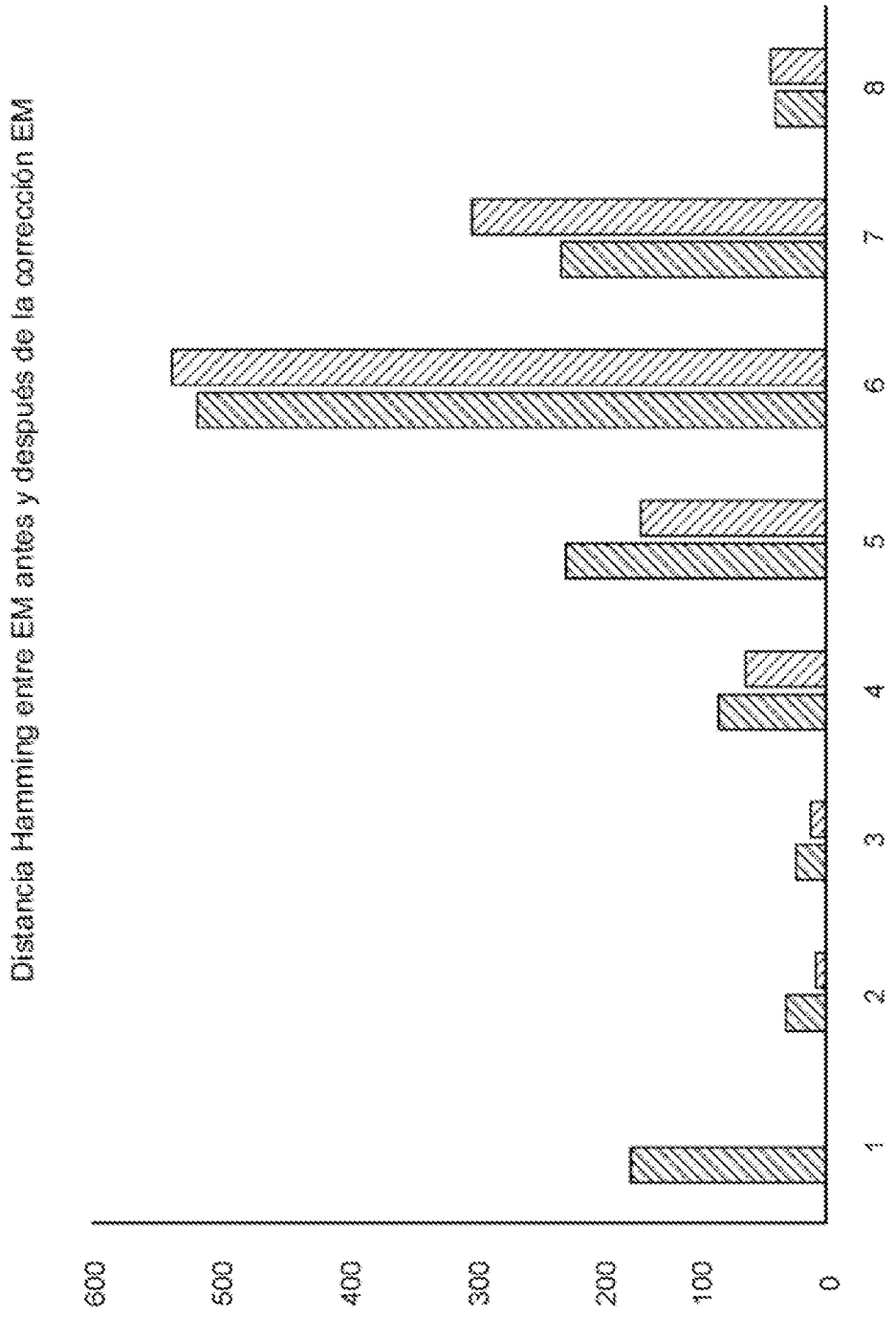

Distancia Hamming entre EM antes y después de la corrección EM
600
500
400
300
200
100
0
1
2
3
4
5
6
7
8
Antes
Después

FIG. 29

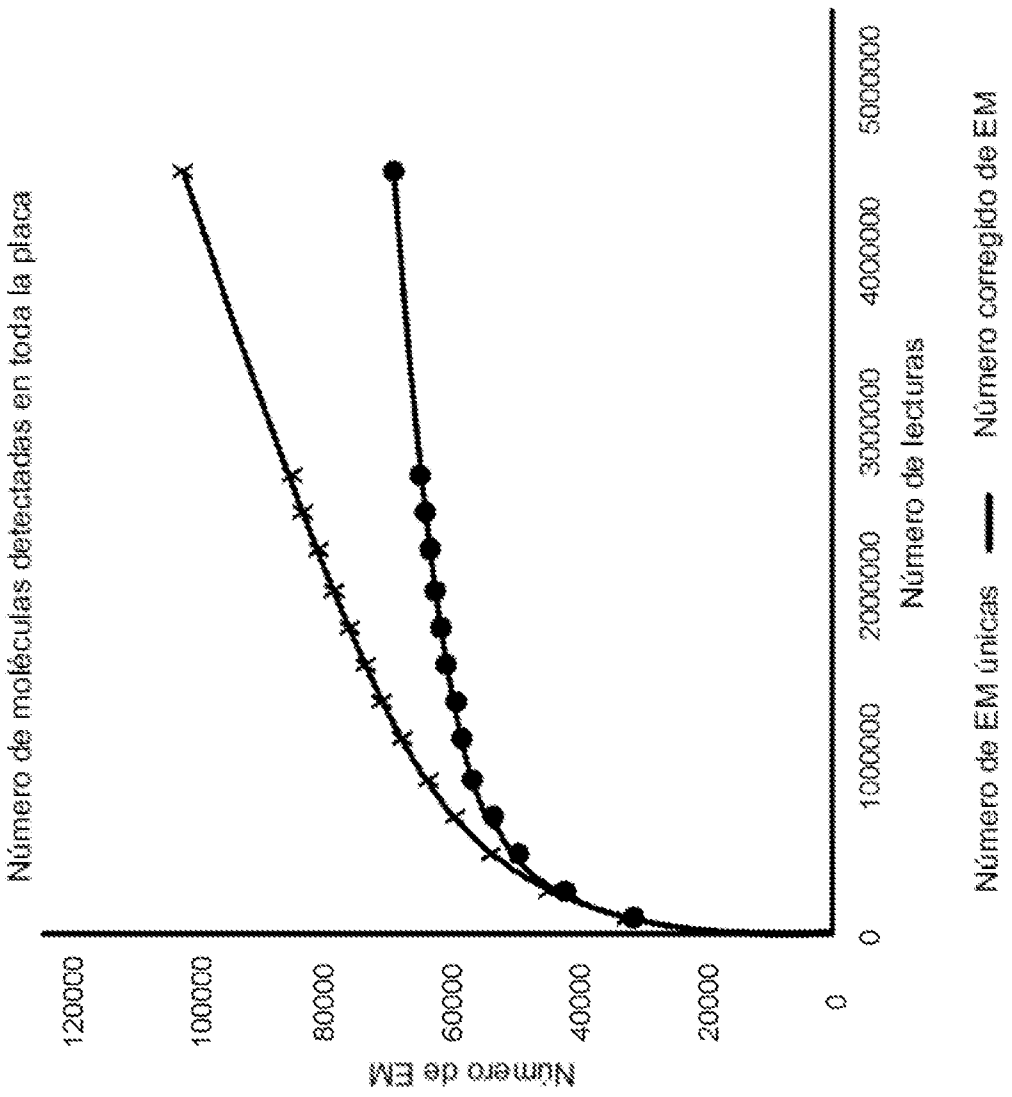

Número de moléculas detectadas en toda la placa
Número de EM
120000
100000
80000
60000
40000
20000
0
0
1000000
2000000
3000000
4000000
5000000
Número de lecturas
Número de EM únicas
Número corregido de EM

FIG. 30

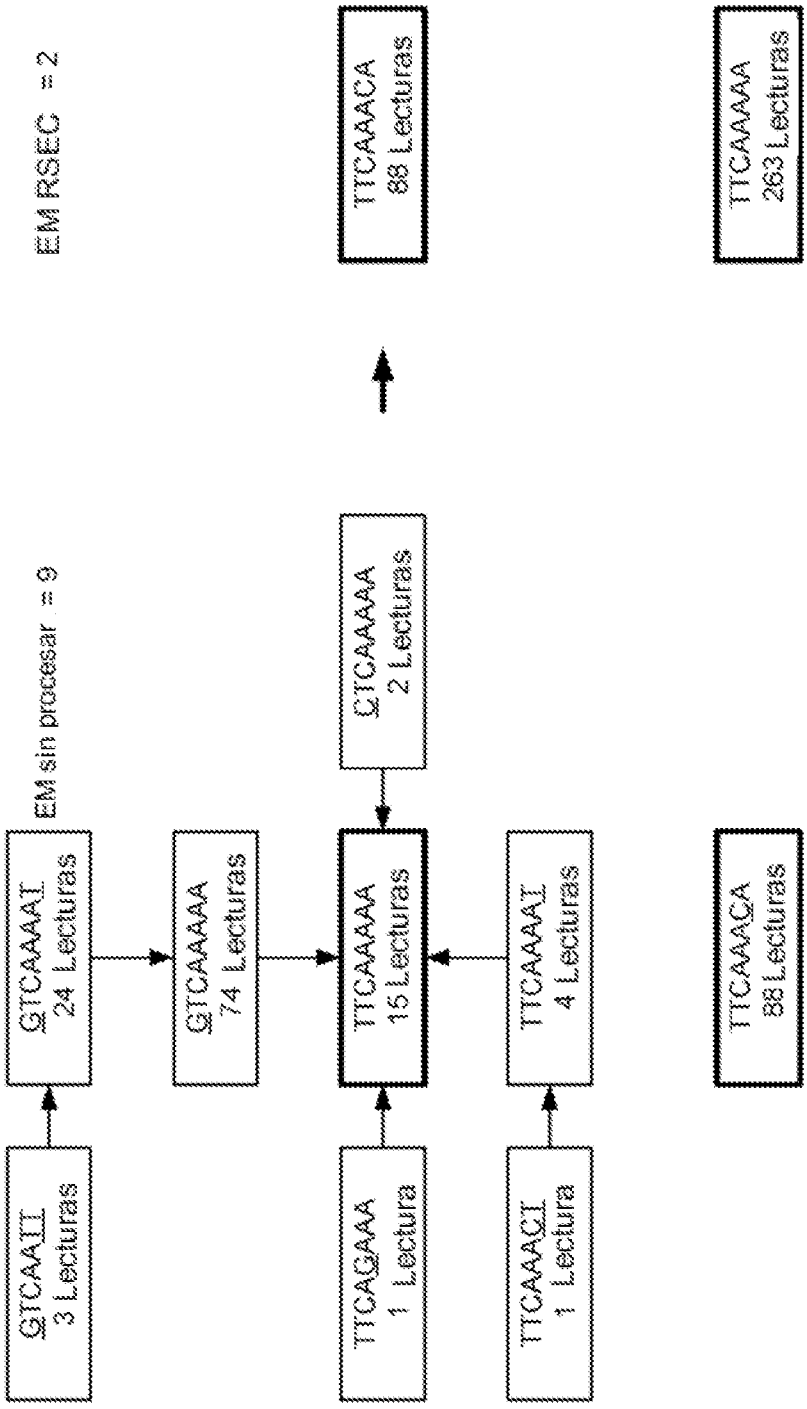

EM RSEC = 2
EM sin procesar = 9
GTCAAAT
3 Lecturas
GTCAAAAT
24 Lecturas
GTCAAAAA
74 Lecturas
TTCAGAAA
1 Lectura
TTCAAAAA
15 Lecturas
GTCAAAAA
2 Lecturas
TTCAAAGT
1 Lectura
TTCAAAAT
4 Lecturas
TTCAAACA
88 Lecturas
TTCAAAACA
88 Lecturas
TTCAAAAA
263 Lecturas

FIG. 31

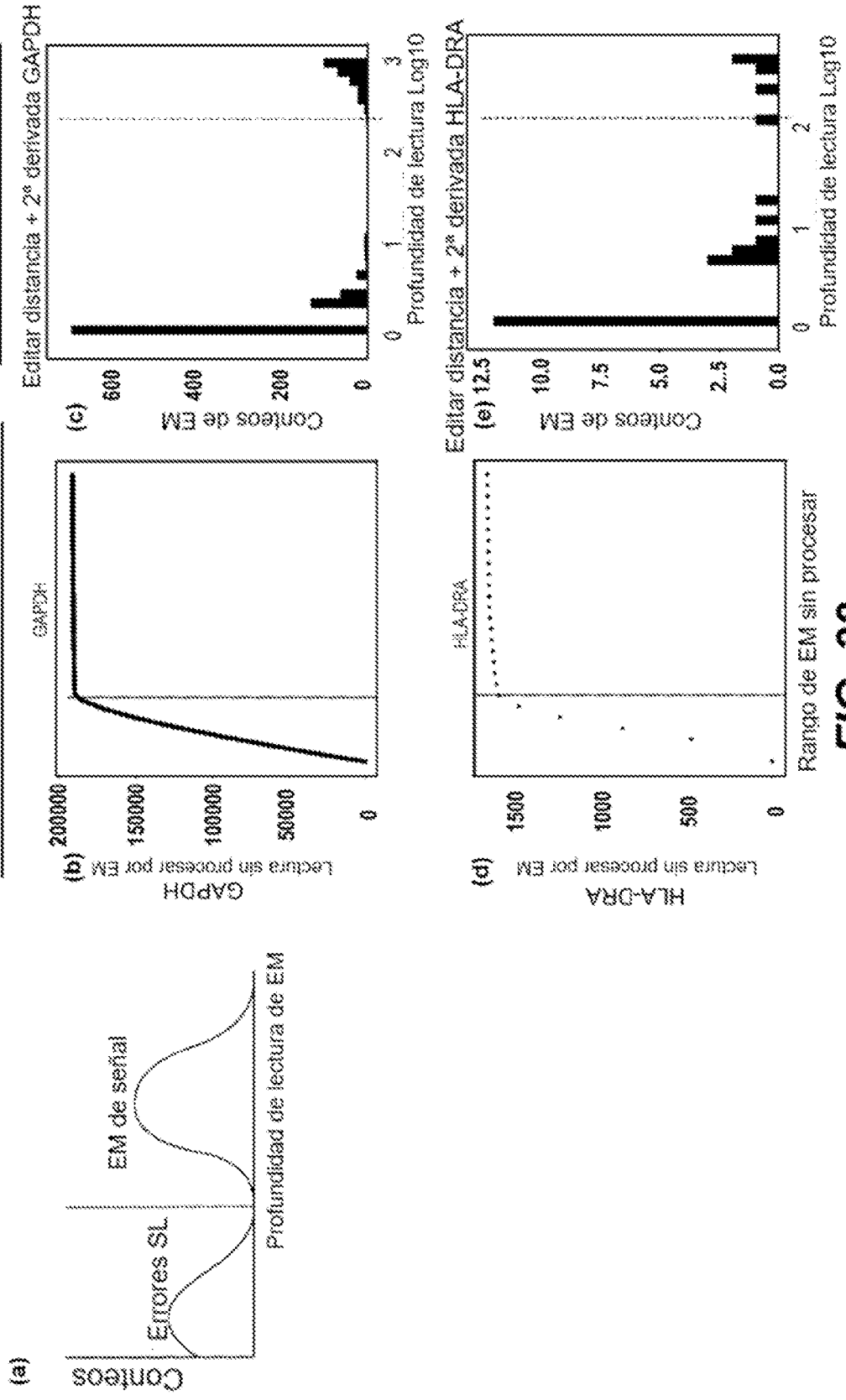

Gráficos de suma cumulativa de lecturas
Histograma de profundidad de EM
(a)
Conteos
Errores SL
EM de señal
Profundidad de lectura de EM
(b)
GAPDH
Lectura sin procesar por EM
200000
150000
100000
50000
0
(c)
Editar distancia + 2ª derivada GAPDH
Conteos de EM
600
400
200
0
0
1
2
3
Profundidad de lectura Log10
(d)
HLA-DRA
Lectura sin procesar por EM
1500
1000
500
0
Rango de EM sin procesar
(e)
Editar distancia + 2ª derivada HLA-DRA
Conteos de EM
12.5
10.0
7.5
5.0
2.5
0.0
0
1
2
Profundidad de lectura Log10

FIG. 32

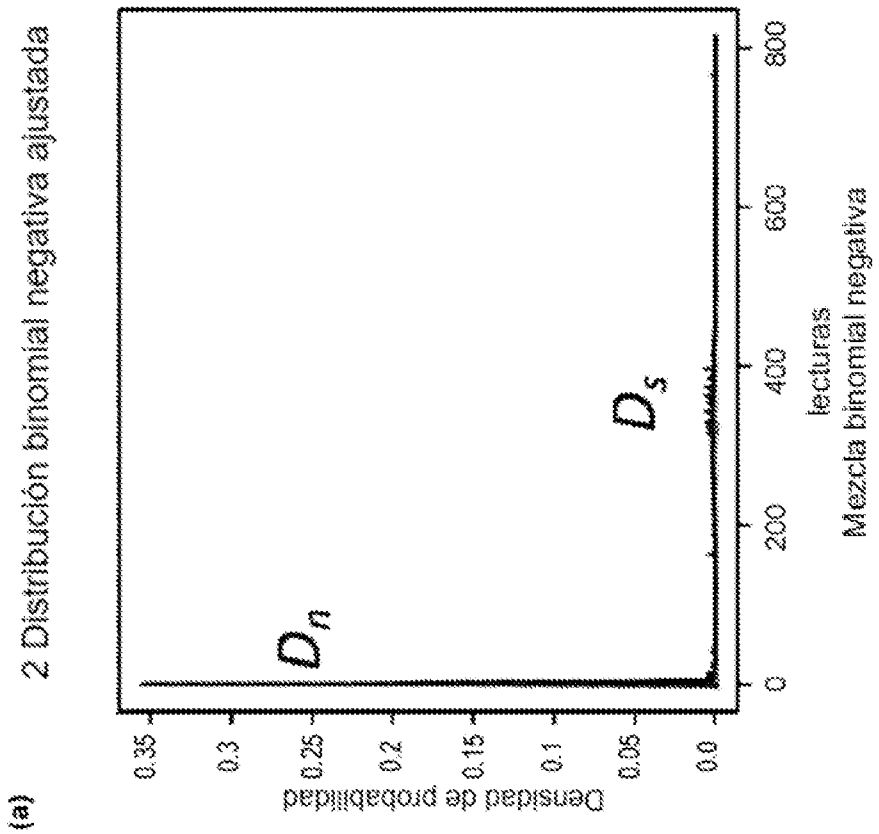

(a)
2 Distribución binomial negativa ajustada
Densidad de probabilidad
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0.0
D_n
D_s
0
200
400
600
800
lecturas
Mezcla binomial negativa

FIG. 33

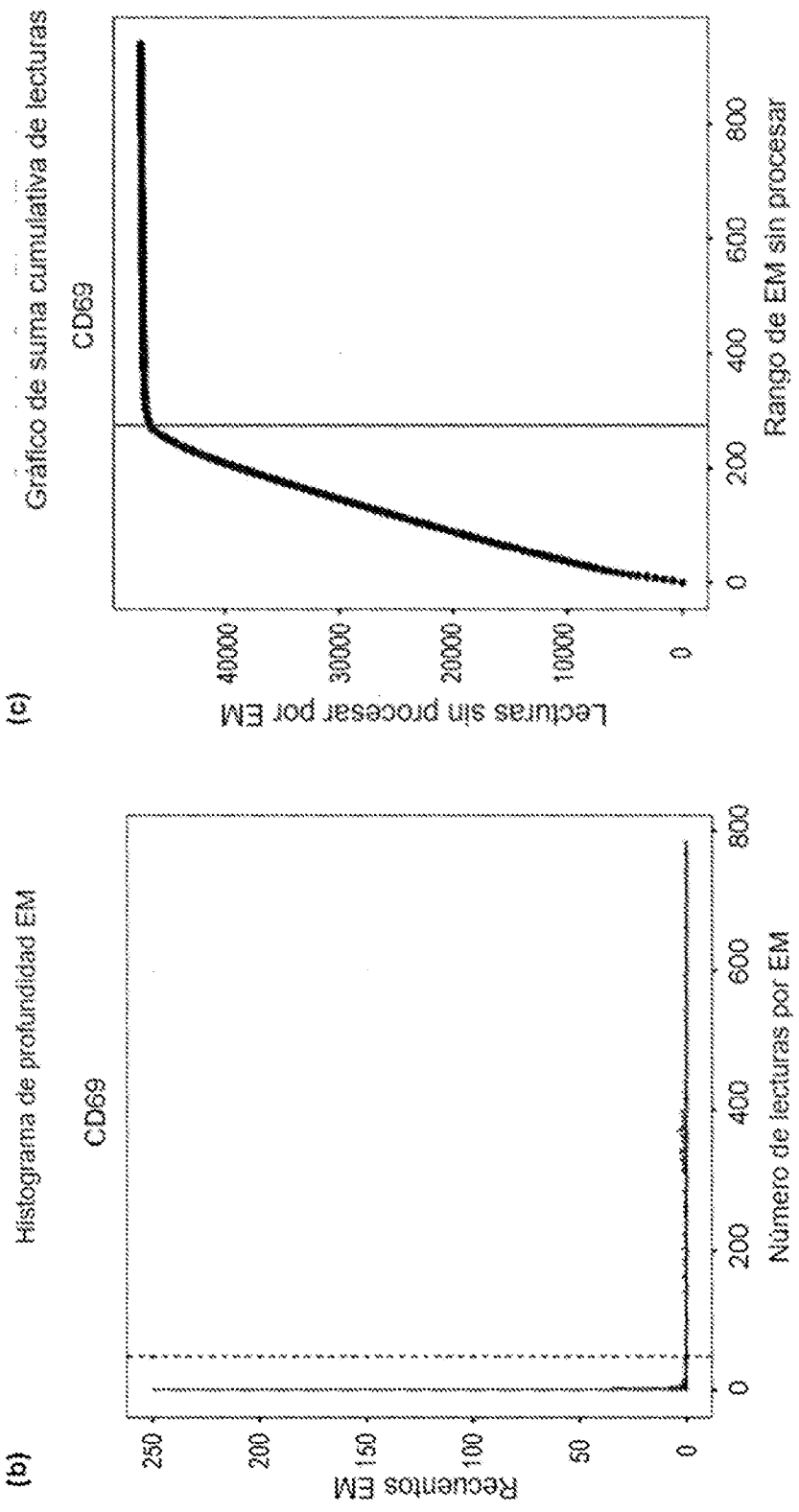


FIG. 33 (Continuada)

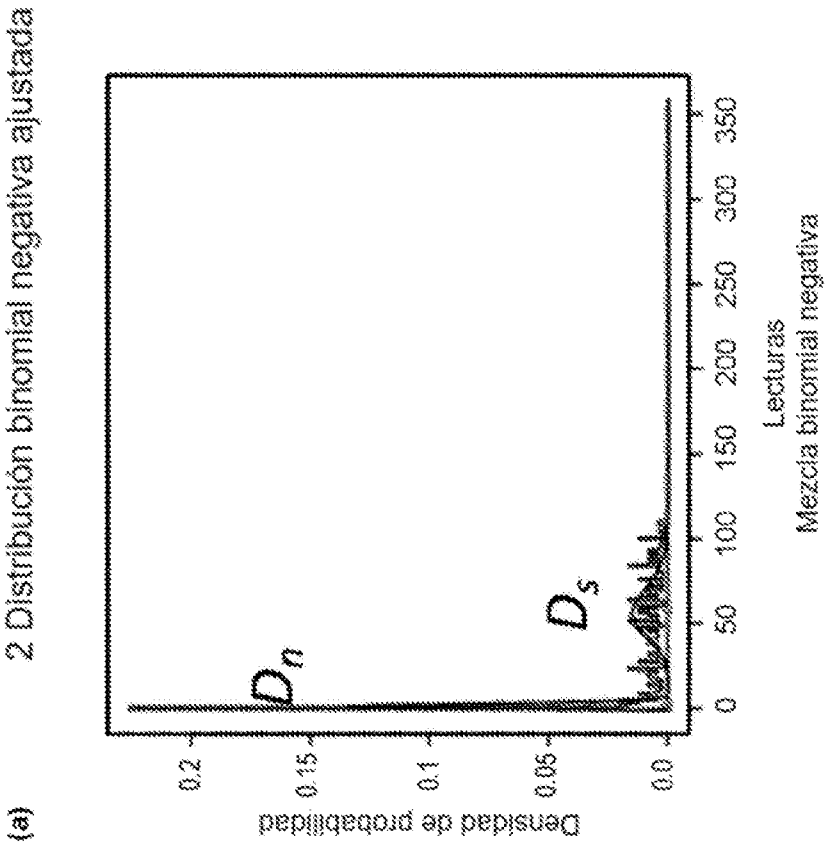

(a)
2 Distribución binomial negativa ajustada
Densidad de probabilidad
0.2
0.15
0.1
0.05
0.0
D_n
D_s
0
50
100
150
200
250
300
350
Lecturas
Mezcla binomial negativa

FIG. 34

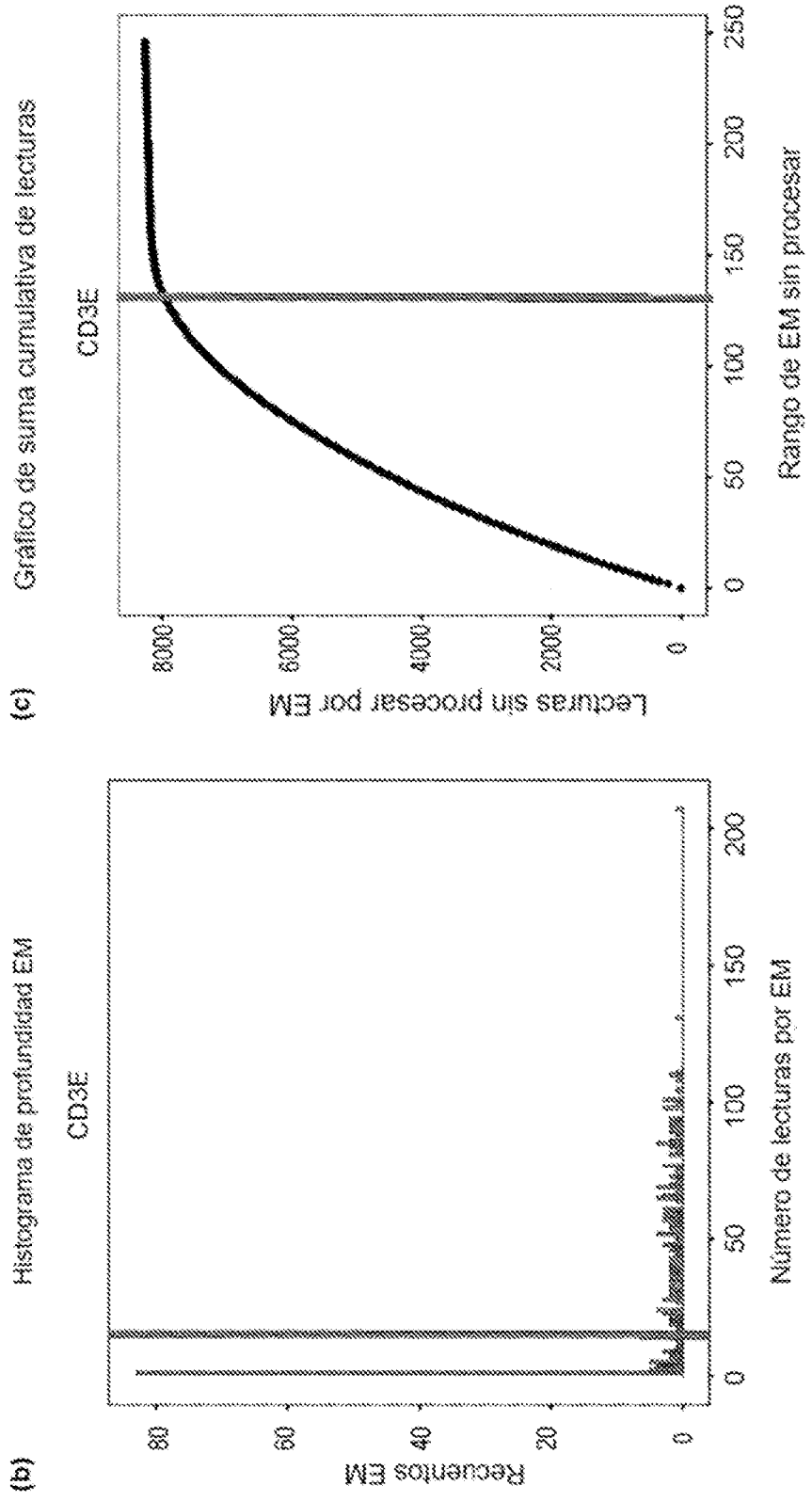


FIG. 34 (Continuada)

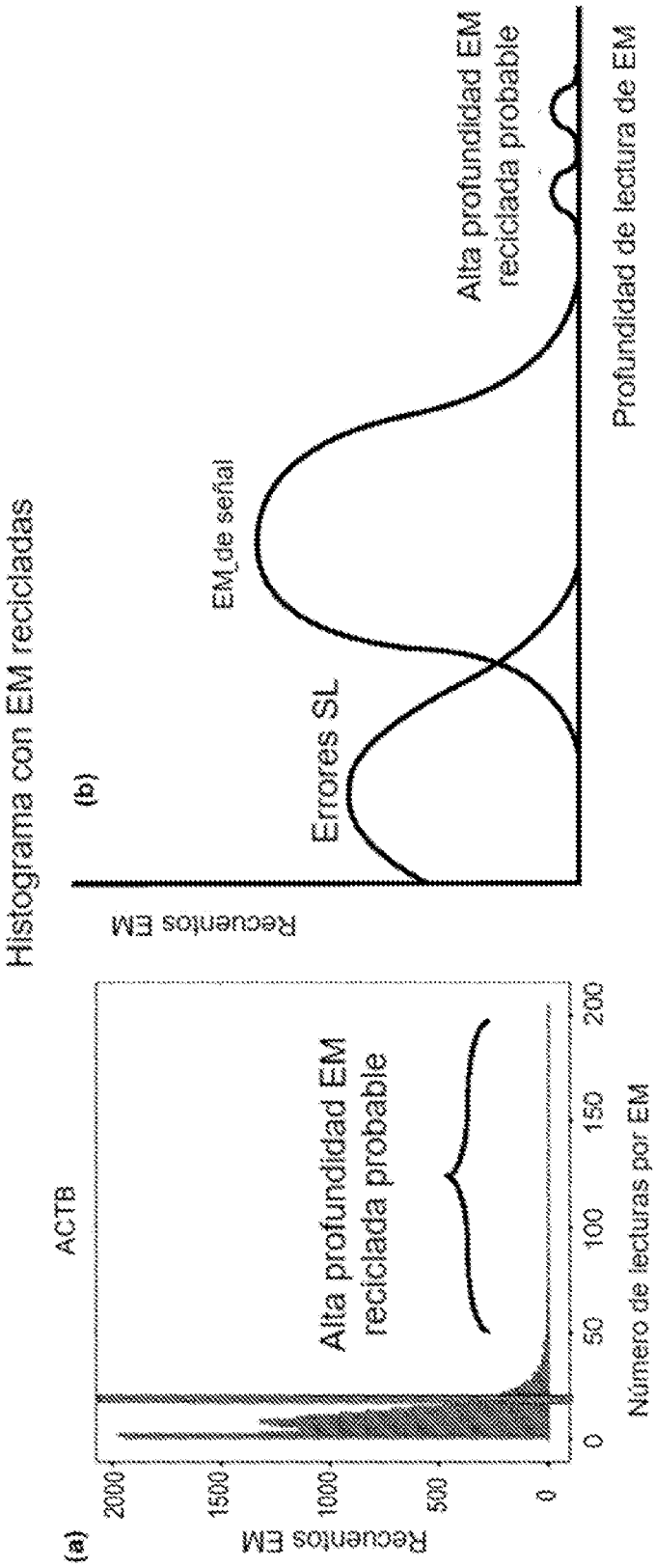

Histograma con EM recicladas
(a)
ACTB
Recuentos EM
2000
1500
1000
500
0
Alta profundidad EM reciclada probable
0
50
100
150
200
Número de lecturas por EM
(b)
Recuentos EM
Errores SL
EM_de señal
Alta profundidad EM reciclada probable
Profundidad de lectura de EM

FIG. 35

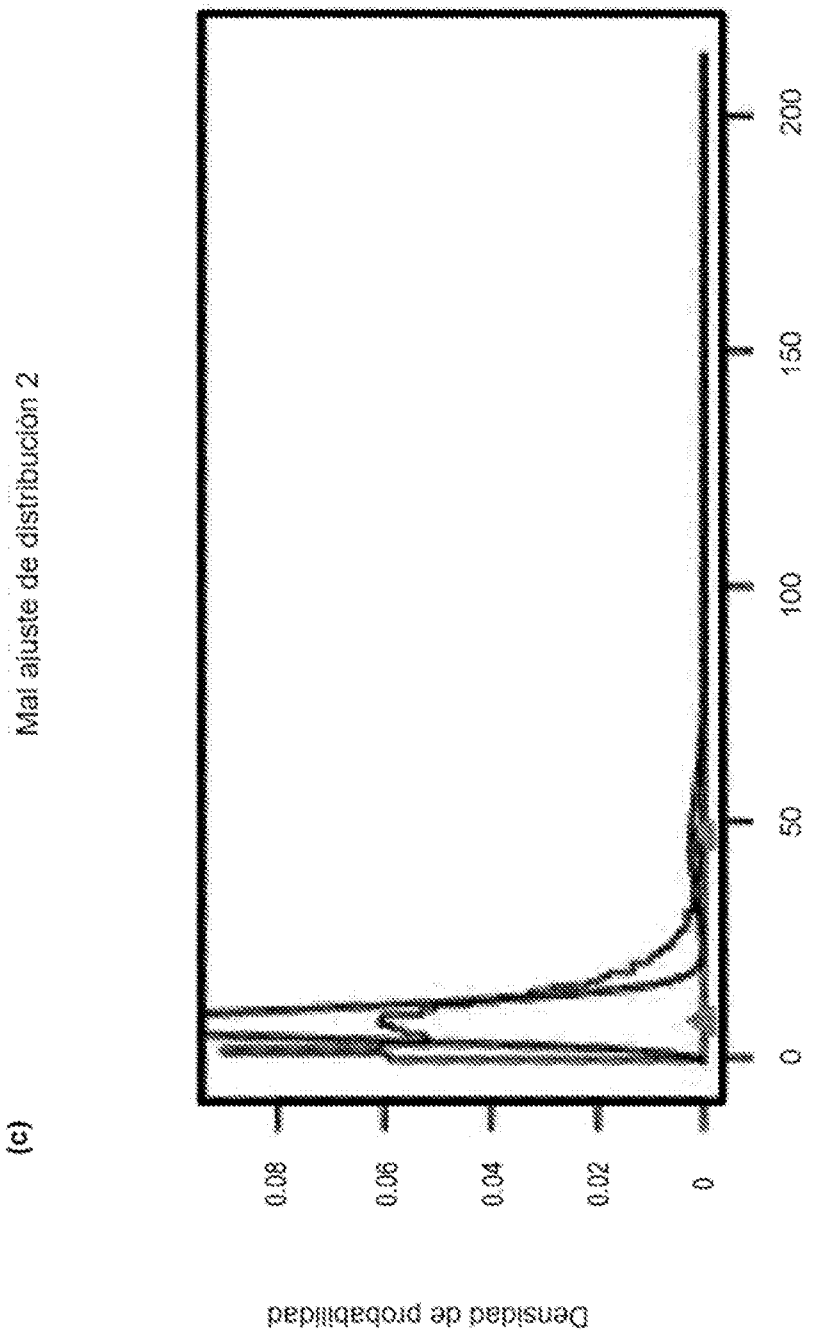


FIG. 35 (Continuada)

ID de gen	EM ajustada	Nº de lecturas por EM	ID de gen	EM ajustada	Nº de lecturas por EM	ID de gen	EM ajustada	Nº de lecturas por EM
GAPDH	GGGGGGGG	1417	ACTB	TTTTTTTA	5395	HSP90AB1	TTTTTTGA	982
GAPDH	GGTTGGTA	702	ACTB	GGGCCGGG	4014	HSP90AB2	TTTGGGGG	432
GAPDH	GCGGCGGG	630	ACTB	GGGGGGCG	3890	HSP90AB3	GGGTGGCG	359
GAPDH	GGTGGGGG	626	ACTB	GGGGTTCG	3749	HSP90AB4	CGTTTTTA	348
GAPDH	TGCGCCTC	598	ACTB	GGGGGGGG	3660	HSP90AB5	TGTTGTTA	335
GAPDH	GGCGGGGA	598	ACTB	TTTTTTTA	3435	HSP90AB6	TTTGGGGC	313
GAPDH	CTTTTTA	581	ACTB	GGCGGGGG	3430	HSP90AB7	TTCTTTTA	305
GAPDH	GGCGAGGG	575	ACTB	GGGCGGCG	3347	HSP90AB8	TGGTTCGC	300
GAPDH	CGGGGGGC	563	ACTB	GTGGGGGG	3282	HSP90AB9	TTTTTTTG	291
GAPDH	CGGGTCGG	551	ACTB	TGGCGGGG	3170	HSP90AB10	CGTGGTGC	283
GAPDH	GGCGTGGG	515	ACTB	GGGGGGGG	3070	HSP90AB11	GGGCGCGA	270
GAPDH	GTGTTGTA	511	ACTB	TTCGGTTG	2913	HSP90AB12	TGGCTGGC	268
GAPDH	TTTGGGGG	479	ACTB	GTCGGGGA	2861	HSP90AB13	GTGGGTGA	264
GAPDH	GTGTGTGA	474	ACTB	GGGGGTGG	2817	HSP90AB14	TTTTTGCG	252
GAPDH	GGGGGGGG	444	ACTB	TTTTTTTG	2734	HSP90AB15	GGCGGGGG	252
GAPDH	TGCGGGGG	442	ACTB	TTGGGCGC	2692	HSP90AB16	TGTCCTTG	249
GAPDH	TTTTTTGC	430	ACTB	TGGCTGGG	2667	HSP90AB17	TGGGGGGG	245
GAPDH	GGGGGGTC	430	ACTB	GTGTGTTG	2642	HSP90AB18	GGGGTTTG	244
GAPDH	GGTGGGGG	423	ACTB	TTGGTTTG	2540	HSP90AB19	TTGGTTTA	243
GAPDH	TGCGCGGA	419	ACTB	GGGGTTGG	2508	HSP90AB20	GGGGTGCC	241

FIG. 36

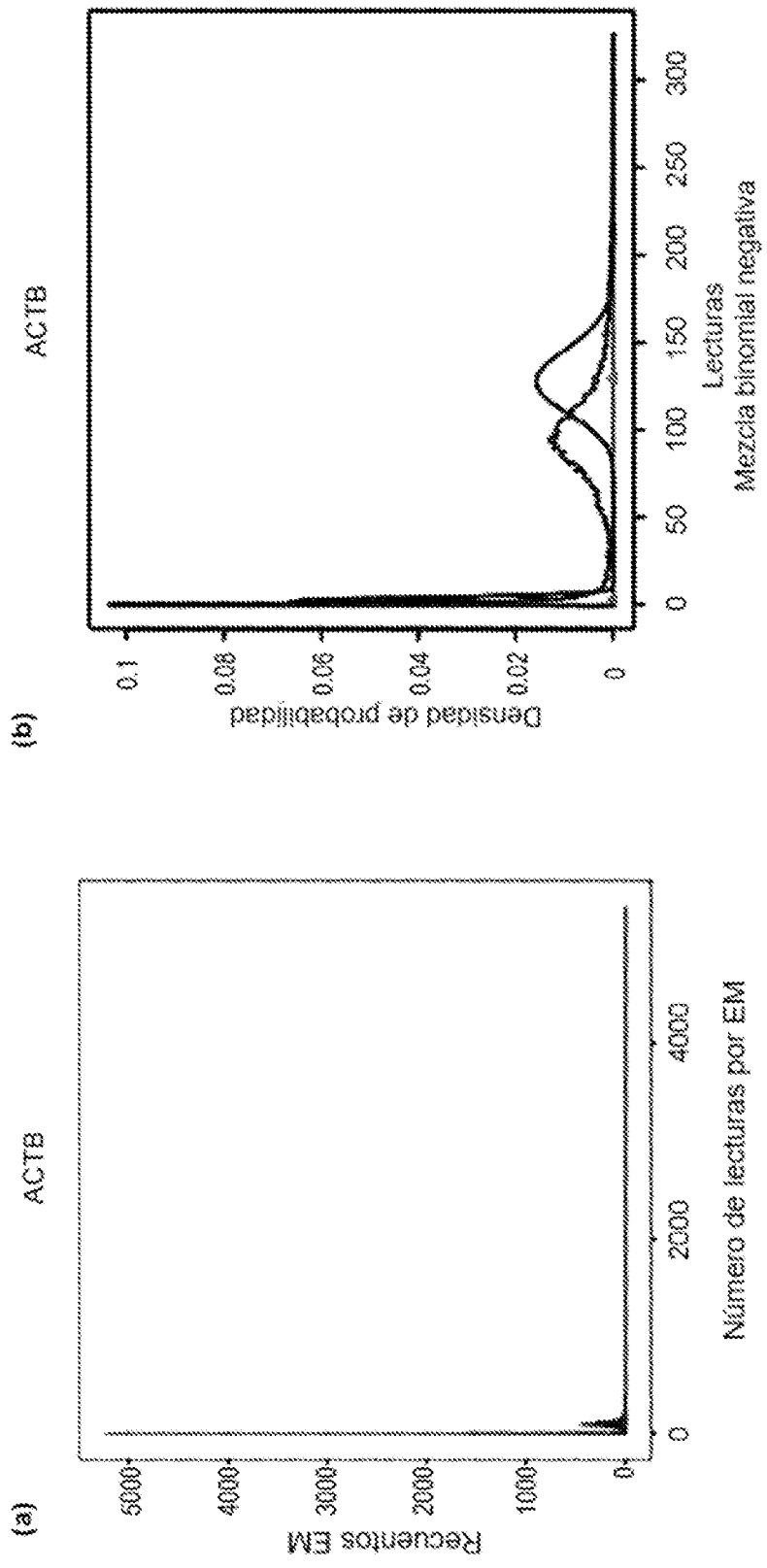

(a)
ACTB
Recuentos EM
5000
4000
3000
2000
1000
0
0
2000
4000
Número de lecturas por EM
(b)
ACTB
Densidad de probabilidad
0.1
0.08
0.06
0.04
0.02
0
0
50
100
150
200
250
300
Lecturas
Mezcla binomial negativa

FIG. 37

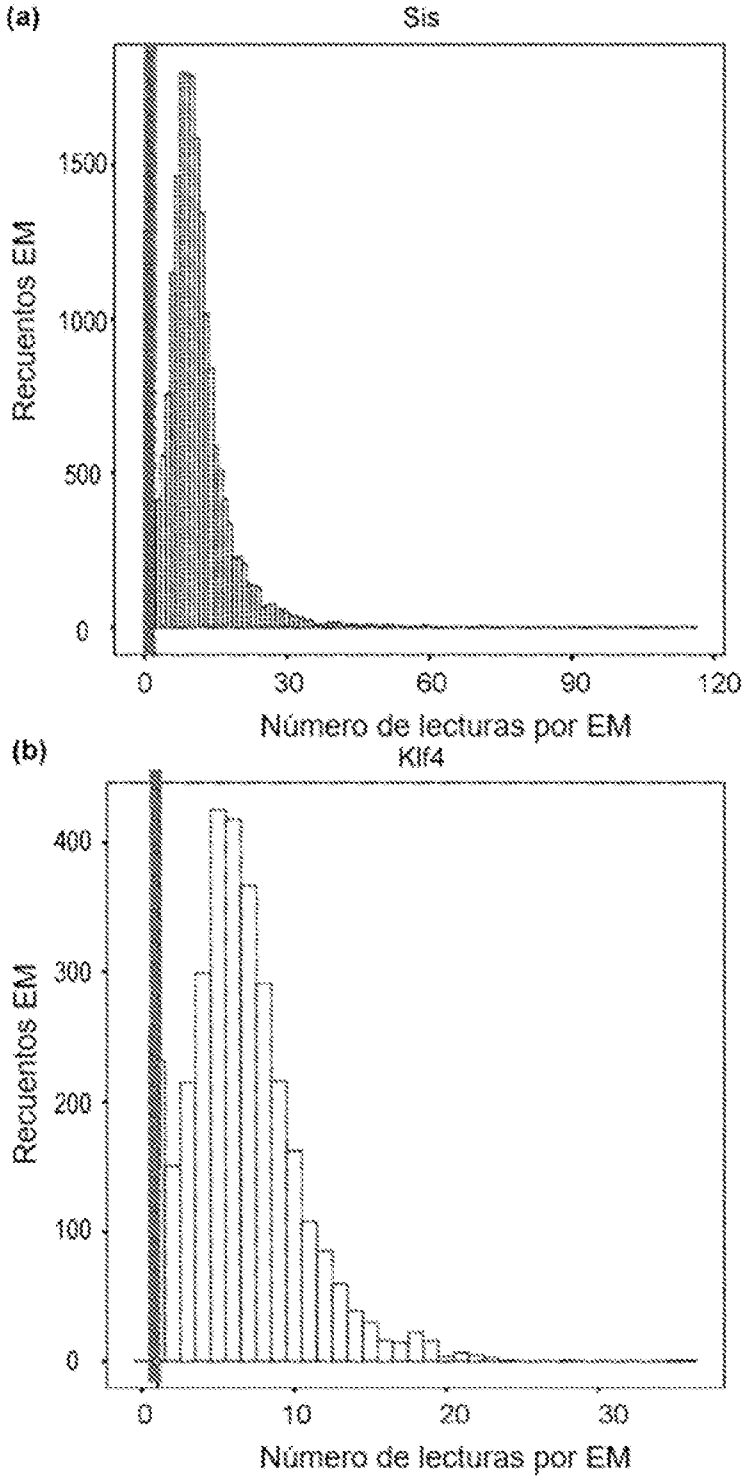

(a)
Sis
Recuentos EM
1500
1000
500
0
0
30
60
90
120
Número de lecturas por EM
(b)
Klf4
Recuentos EM
400
300
200
100
0
0
10
20
30
Número de lecturas por EM

FIG. 38

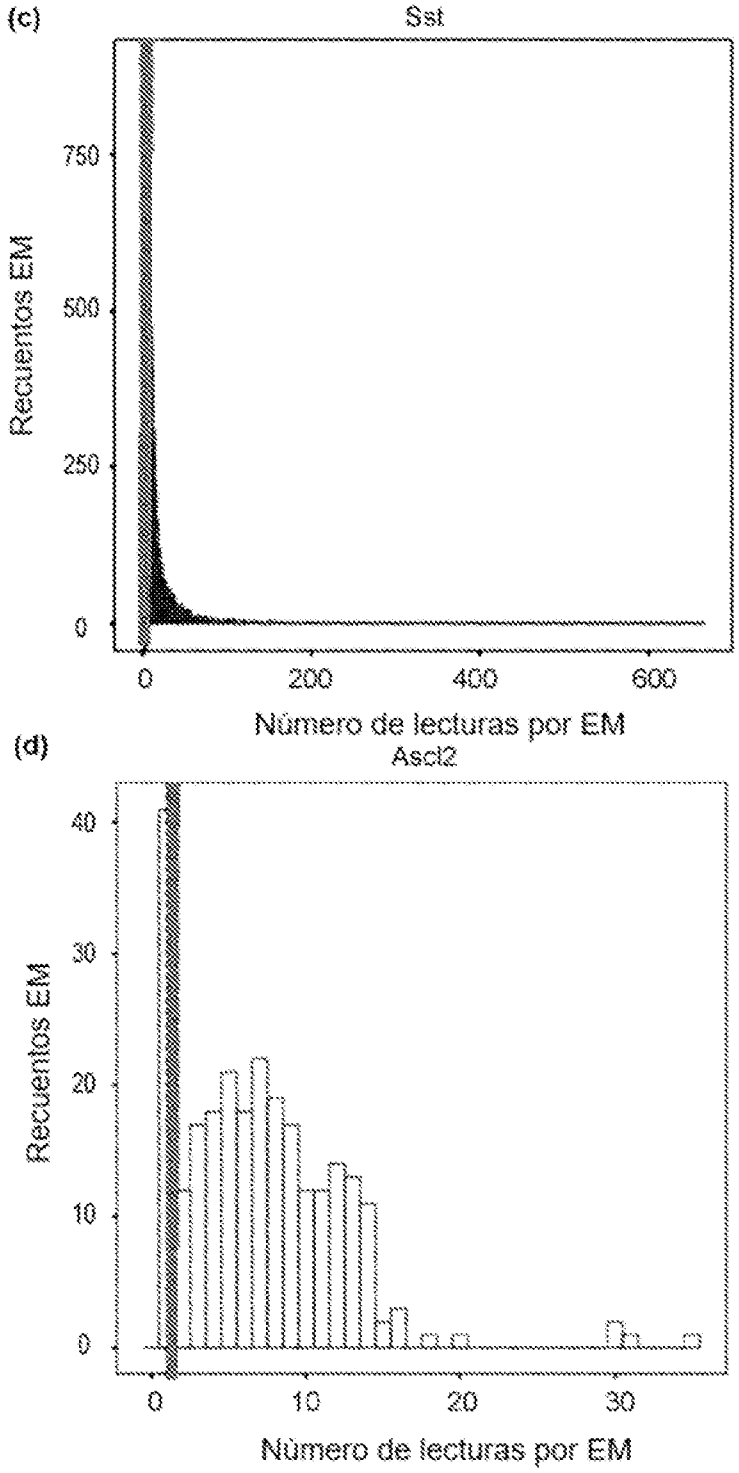


FIG. 38 (Continuada)

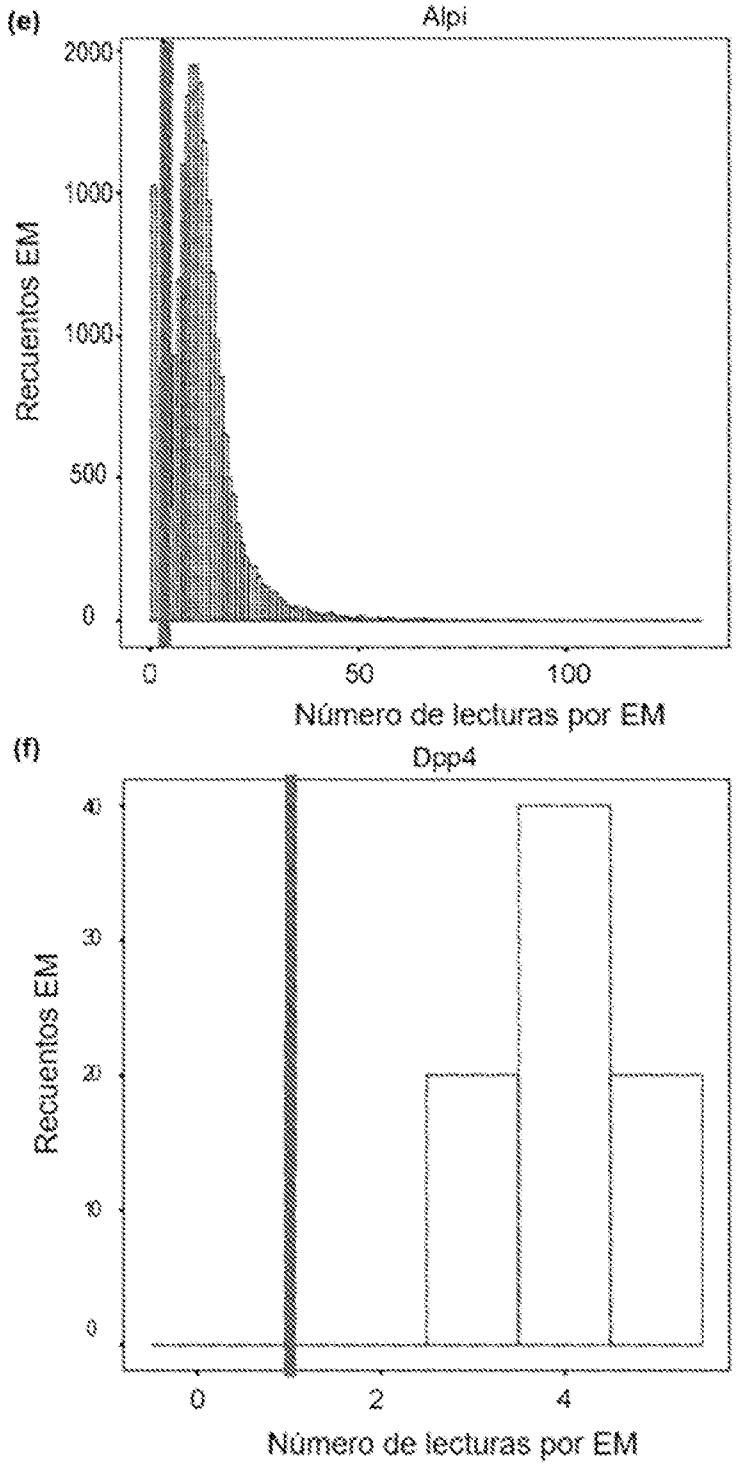


FIG. 38 (Continuada)

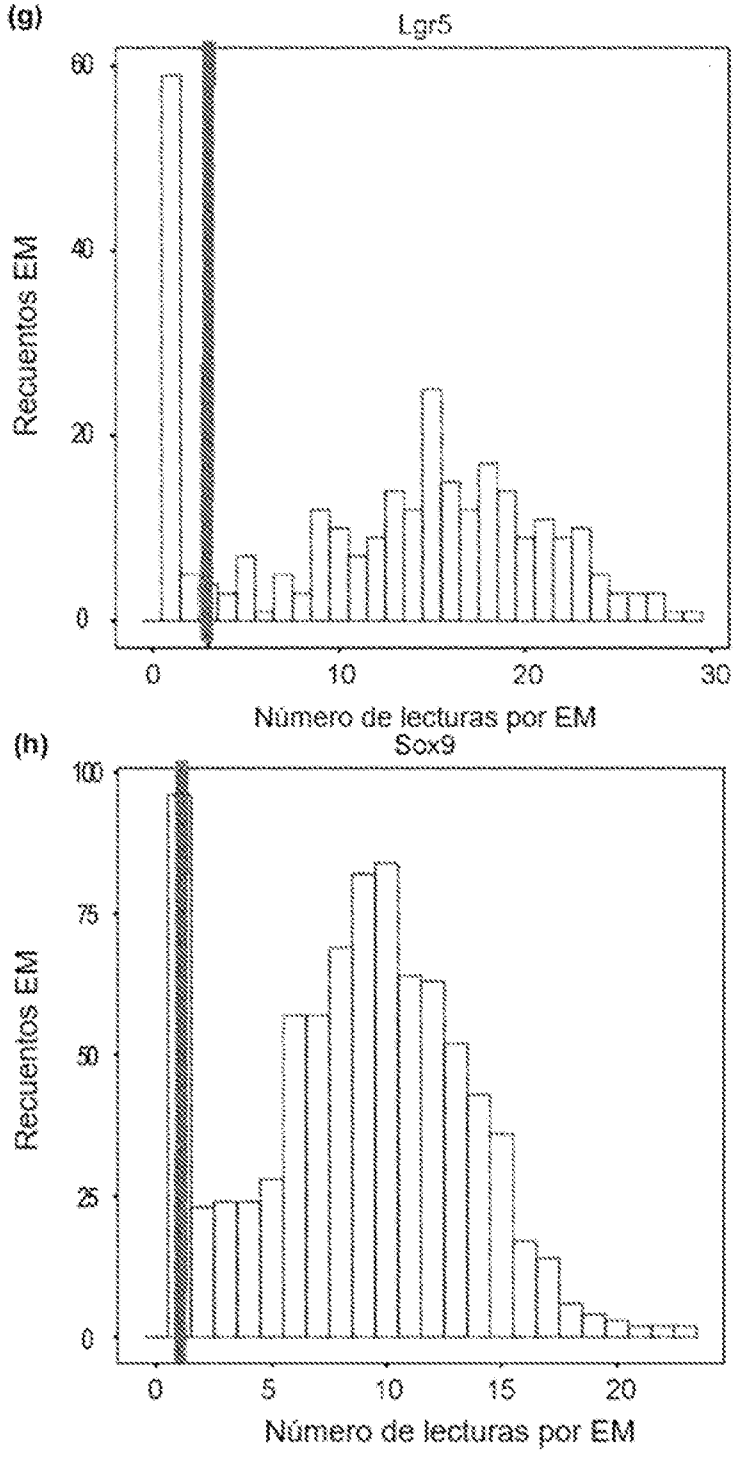


FIG. 38 (Continuada)

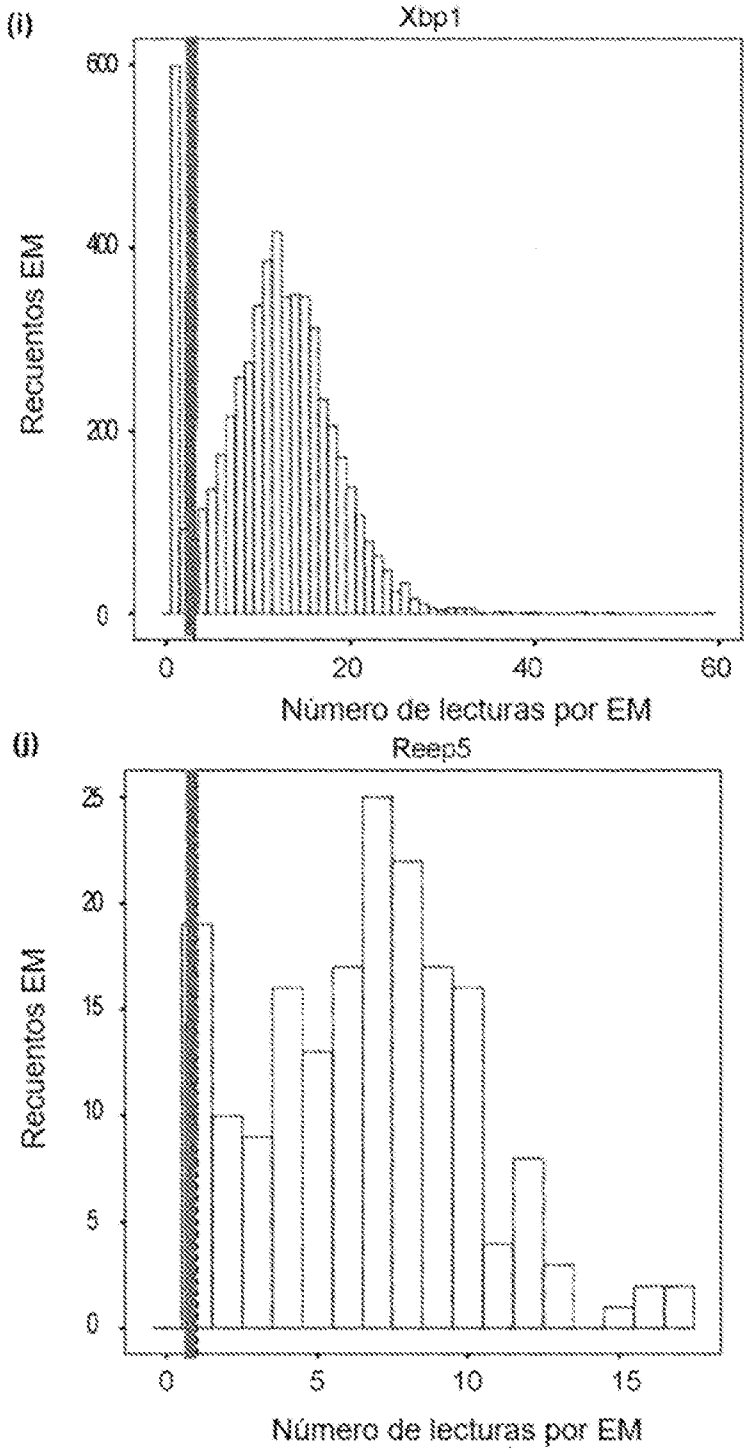


FIG. 38 (Continuada)

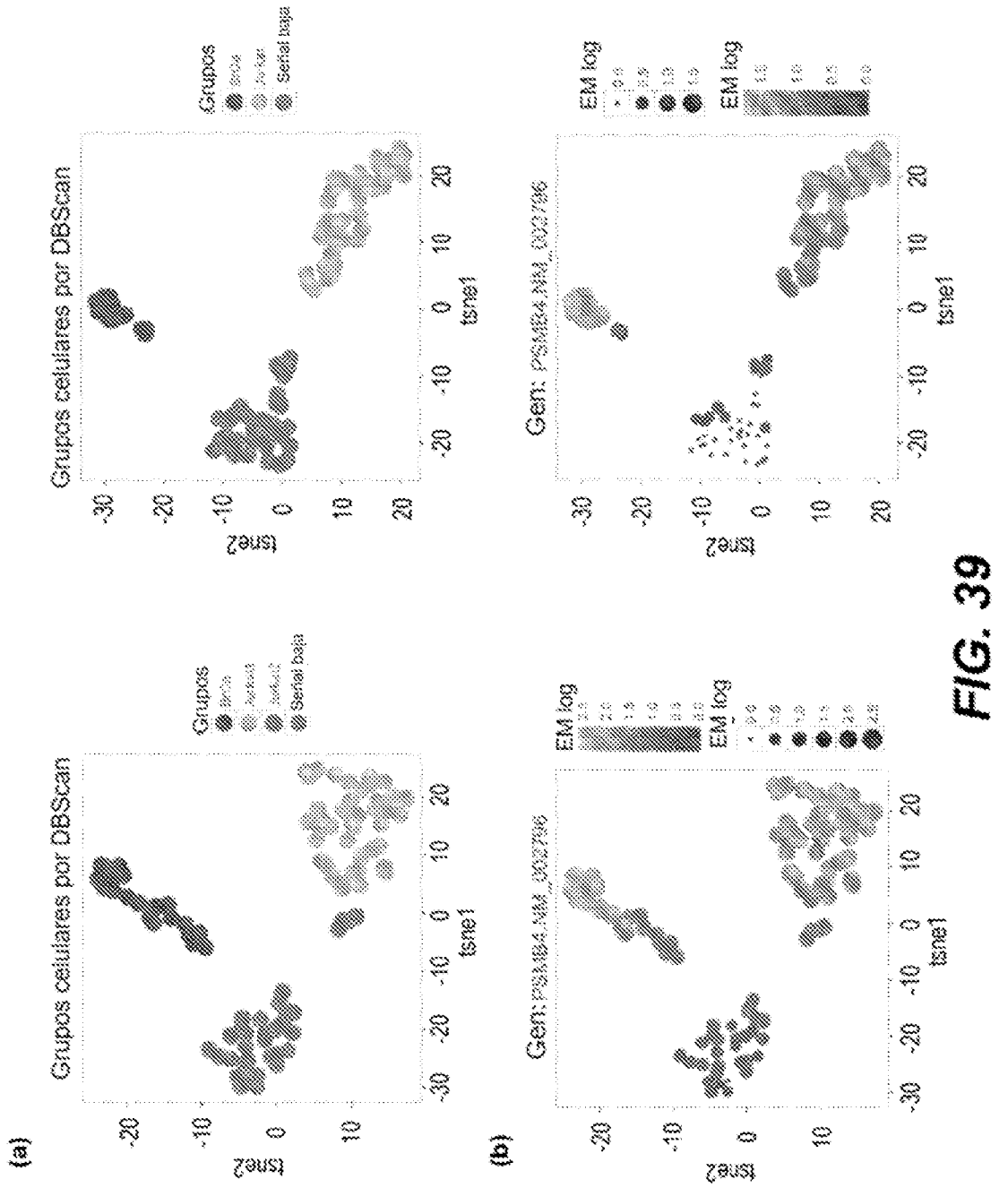

(a)
Grupos celulares por DBScan
Grupos
Señal baja
tsne1
tsne2
(b)
Gen: PSMB4-NM_002796
EM log
tsne1
tsne2

FIG. 39

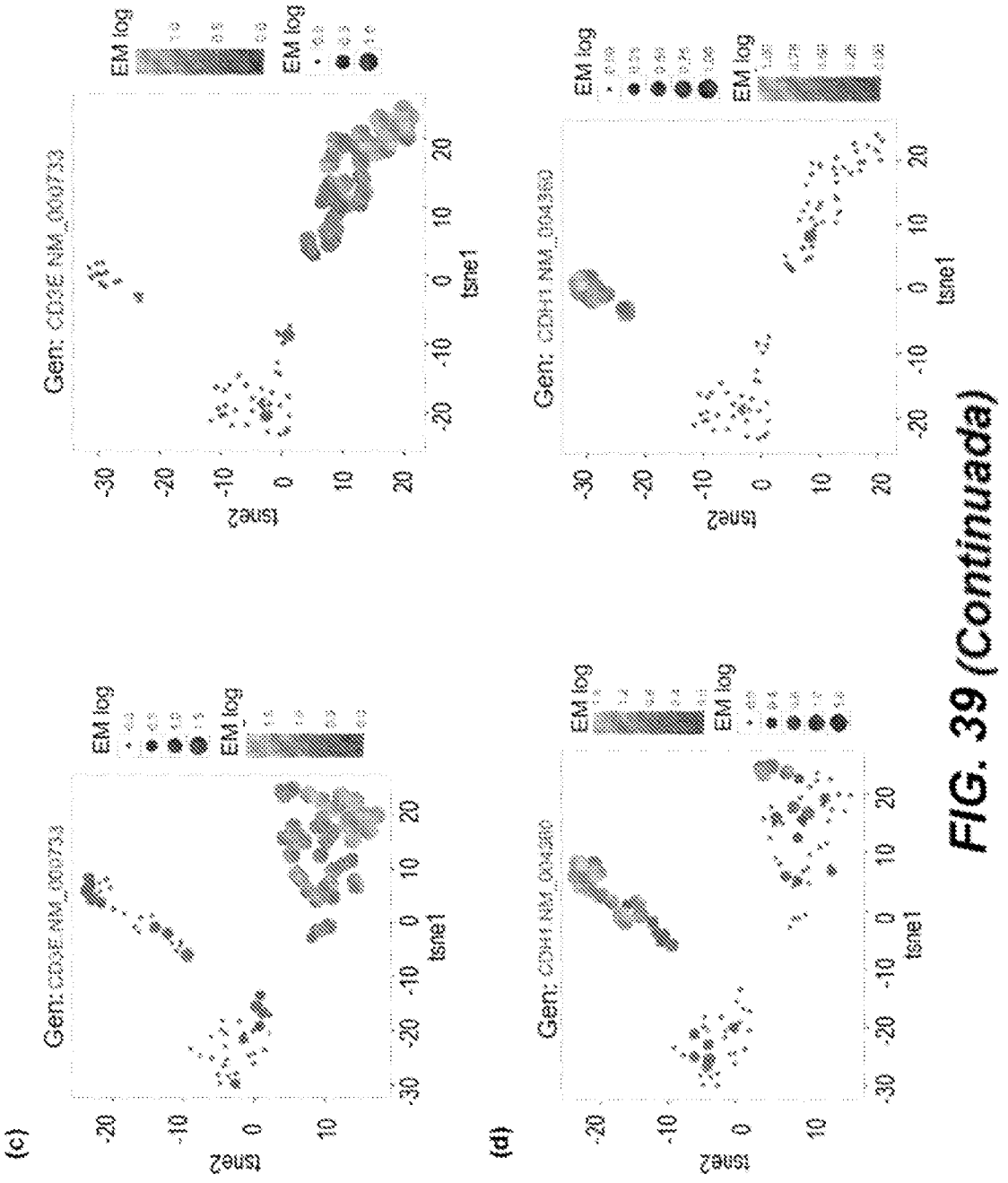


FIG. 39 (Continuada)

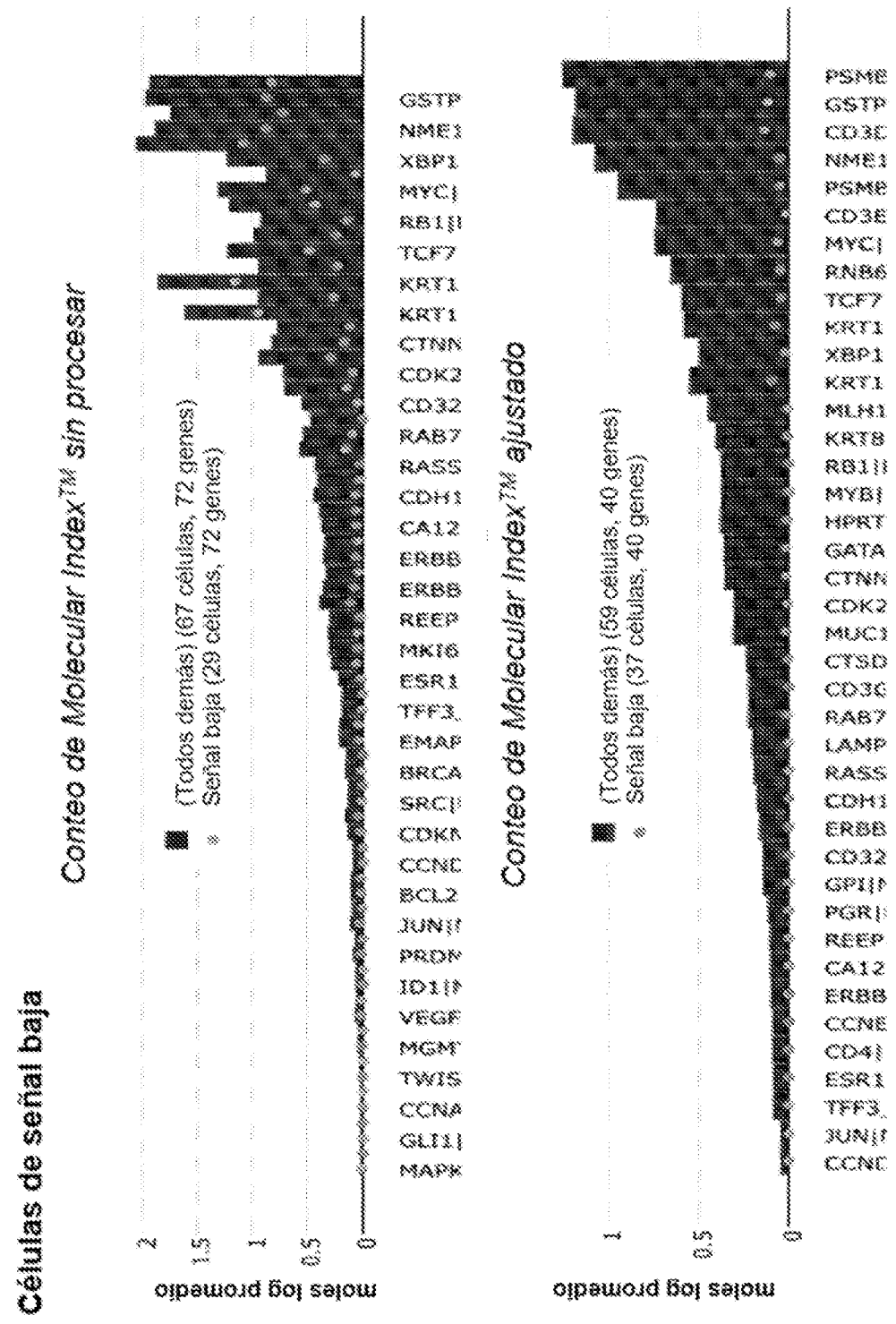

(a) Células de señal baja
Conteo de Molecular Index™ sin procesar
(Todos demás) (67 células, 72 genes)
Señal baja (29 células, 72 genes)
moles log promedio
2
1.5
1
0.5
0
GSTP
NME1
XBP1
MYC
RB1
TCF7
KRT1
KRT1
CTNN
CDK2
CD32
RAB7
RASS
CDH1
CA12
ERBB
ERBB
REEP
MK16
ESR1
TFF3
EMAP
BRCA
SRC
CDKN
CCNE
BCL2
JUN
PRDM
ID1
VEGF
MGMT
TWIS
CCNA
GLI1
MAPK
Conteo de Molecular Index™ ajustado
(Todos demás) (59 células, 40 genes)
Señal baja (37 células, 40 genes)
moles log promedio
1
0.5
0
PSME
GSTP
CD3E
NME1
PSME
CD3E
MYC
RNB6
TCF7
KRT1
XBP1
KRT1
MLH1
KRT8
RB1
MYB
HPRT
GATA
CTNN
CDK2
MUC1
CTSD
CD3E
RAB7
LAMP
RASS
CDH1
ERBB
CD32
GPI
PGR
REEP
CA12
ERBB
CCNE
CD4
ESR1
TFF3
JUN
CCNE

FIG. 40

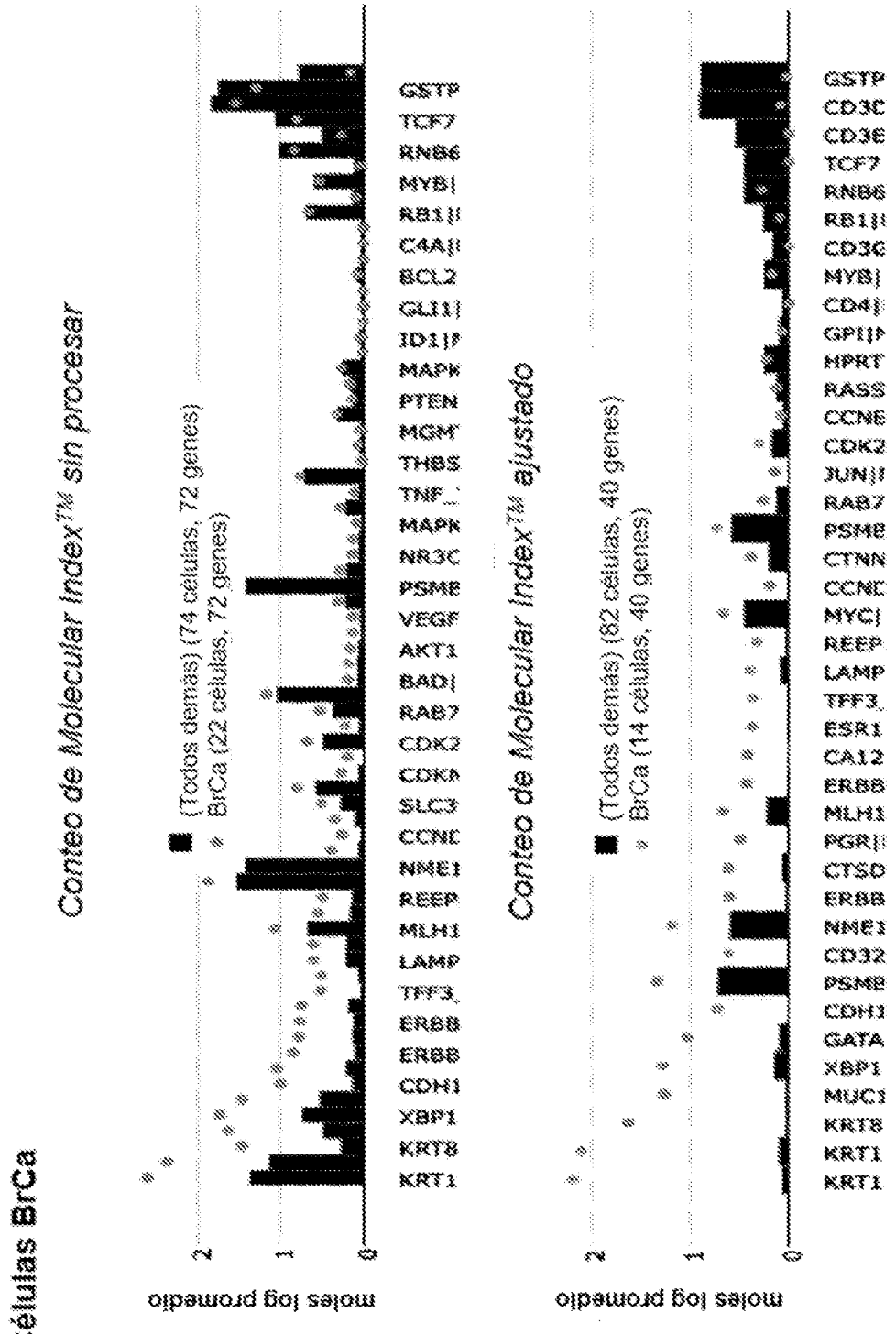


FIG. 40 (Continuada)

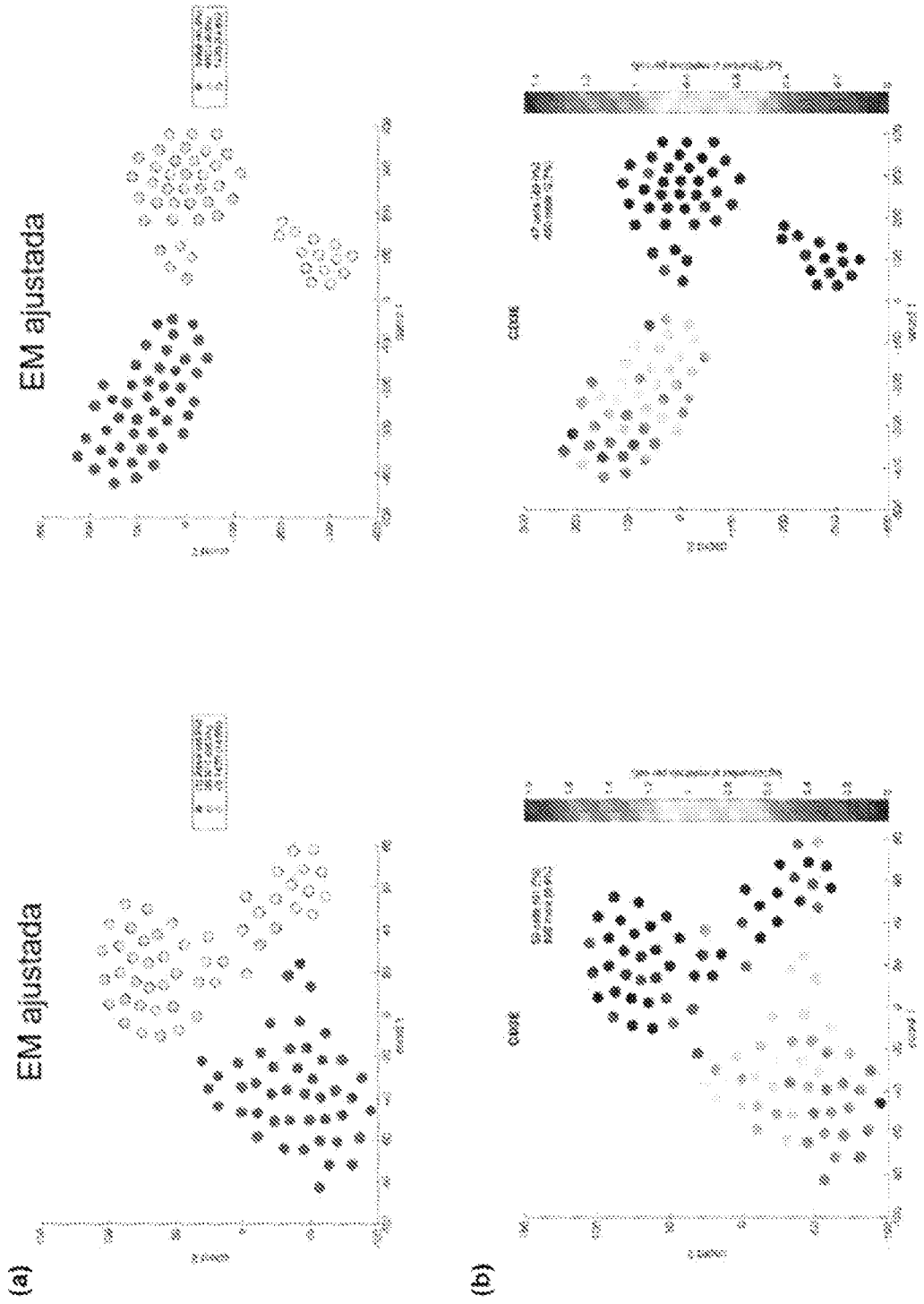


FIG. 41

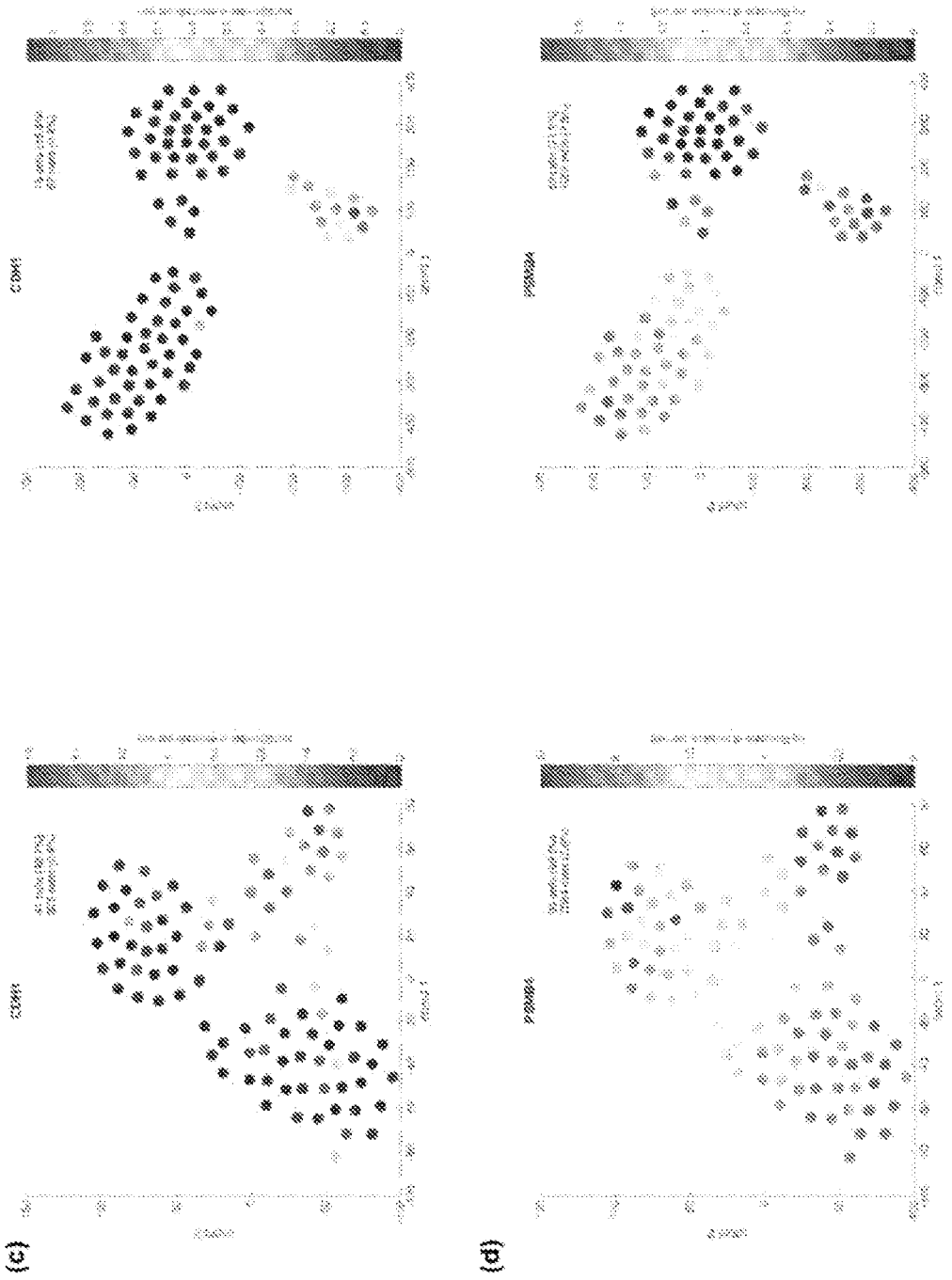


FIG. 41 (Continuada)

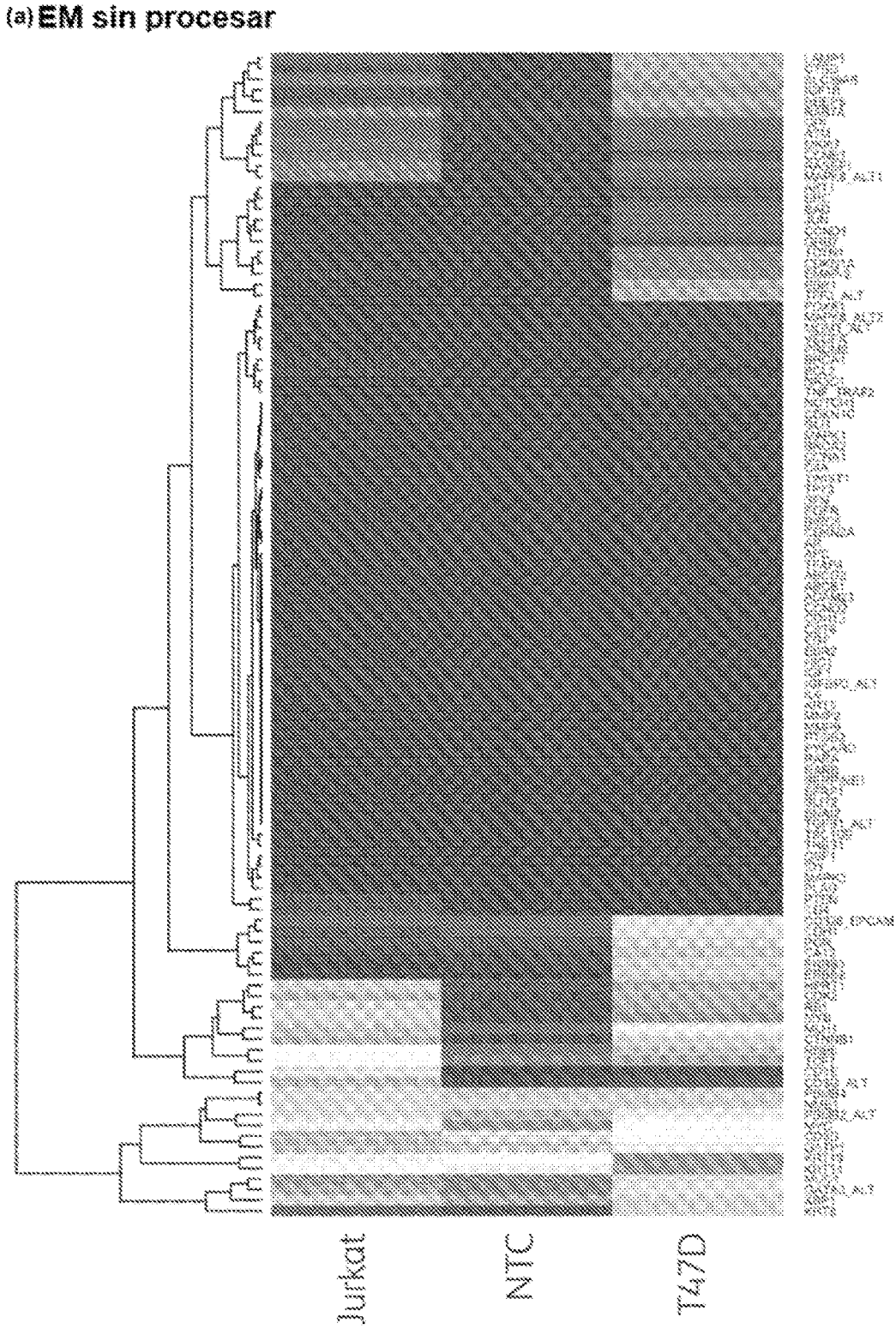


FIG. 42

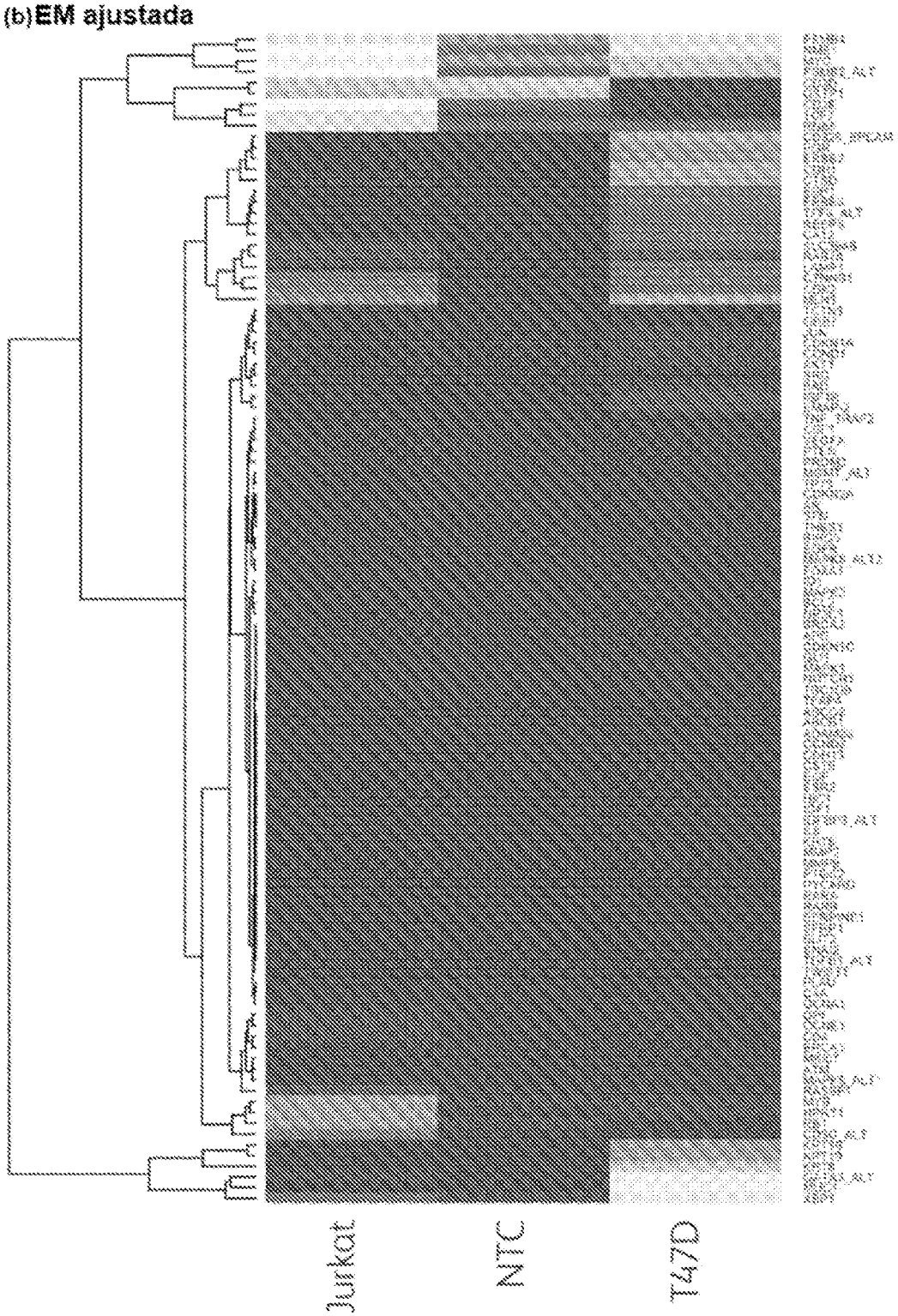


FIG. 42 (Continuada)