(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 980 967**

(51) Int. Cl.:

G16B 25/00 (2009.01)
G06F 16/22 (2009.01)
G06F 16/28 (2009.01)
G16B 40/00 (2009.01)
G16B 99/00 (2009.01)
G16B 25/10 (2009.01)
G16B 40/30 (2009.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(86) Fecha de presentación y número de la solicitud internacional: **07.11.2017 PCT/US2017/060451**

(87) Fecha y número de publicación internacional: **17.05.2018 WO18089378**

(96) Fecha de presentación y número de la solicitud europea: **07.11.2017 E 17805344 (3)**

(97) Fecha y número de publicación de la concesión europea: **17.04.2024 EP 3539035**

(54) Título: **Métodos para la clasificación de perfiles de expresión**

(30) Prioridad:

08.11.2016 US 201662419291 P
13.01.2017 US 201762446227 P

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
03.10.2024

(73) Titular/es:

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

(72) Inventor/es:

FAN, JUE;
ZHANG, JESSE y
HU,JING

(74) Agente/Representante:

IZQUIERDO BLANCO, María Alicia

DESCRIPCIÓN

Métodos para la clasificación de perfiles de expresión

SOLICITUDES RELACIONADAS

La presente solicitud reivindica prioridad de la Solicitud Provisional de Estados Unidos Nº 62/419.291, presentada el 8 de noviembre de 2016; y de la Solicitud Provisional de Estados Unidos Nº 62/446.227, presentada el 13 de enero de 2017.

AVISO DE DERECHOS DE AUTOR Y MARCA COMERCIAL

ANTECEDENTES

Campo

La presente divulgación se refiere de manera general al campo de la clasificación de perfiles de expresión y, más particularmente, a la identificación de dianas para distinguir tipos de células.

Descripción de la técnica relacionada

Los métodos y técnicas como la codificación con códigos de barras (por ejemplo, codificación con códigos de barras estocástica) son útiles para el análisis celular. Por ejemplo, puede usarse la codificación con códigos de barras para descifrar perfiles de expresión génica de células individuales para determinar sus estados usando, por ejemplo, la transcripción inversa, la amplificación por reacción en cadena de la polimerasa (PCR) y la secuenciación de próxima generación (NGS). Sin embargo, la gran cantidad de datos generados por estos métodos y técnicas debe analizarse más a fondo para identificar marcadores para distinguir los tipos de células y para determinar los tipos de células analizadas. SORLIE T ET AL: "Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications", PROCEEDINGS NATIONAL ACADEMY OF SCIENCES PNAS, NATIONAL ACADEMY OF SCIENCES, US, vol. 98, Nº 19, 11 de septiembre de 2001 (2001-09-11), divulga un método para clasificar los carcinomas de mama basándose en la expresión génica usando al mismo tiempo agrupaciones y dendogramas. Este documento no divulga la fusión de conjuntos cuando la diferencia entre perfiles de expresión está entre una distancia de fusión.

SUMARIO

En la presente se divulgan métodos para identificar dianas para distinguir tipos de células. En algunas realizaciones, el método comprende: (a) recibir una estructura de datos de recuentos de dianas, en donde la estructura de datos de recuentos de dianas comprende perfiles de expresión de una pluralidad de células, y en donde los perfiles de expresión de la pluralidad de células comprenden un número de cada diana de una pluralidad de dianas para cada célula de la pluralidad de células; (b) agrupar jerárquicamente los perfiles de expresión de la pluralidad de células basándose en la estructura de datos de recuentos de dianas y las distancias entre los perfiles de expresión de la pluralidad de células para generar un dendrograma que represente los perfiles de expresión de la pluralidad de células, en donde el dendrograma comprende una pluralidad de nodos, en donde la pluralidad de nodos comprende un nodo raíz, una pluralidad de nodos hoja, y una pluralidad de nodos no raíz y no hoja, en donde cada nodo hoja de la pluralidad de nodos hoja representa un perfil de expresión de una célula diferente de la pluralidad de células, y en donde el nodo raíz representa perfiles de expresión de la pluralidad de células; (c) mientras se pasa a través de cada nodo de la pluralidad de nodos del dendrograma desde el nodo raíz del dendrograma hasta la pluralidad de nodos hoja del dendrograma: (1) determinar si una división del nodo en nodos hijos del nodo es válida o inválida (por ejemplo, las diferencias entre los nodos hijos no son significativas); y (2) si la división del nodo en los nodos hijos del nodo no es válida, añadir el nodo a un conjunto de grupos de fusión; (d) iterativamente, para cada primer nodo en el conjunto de grupos de fusión, si una distancia entre el primer nodo en el conjunto de grupos de fusión y un segundo nodo en el conjunto de grupos de fusión que está más cerca del primer nodo está dentro de un umbral de distancia de fusión, fusionar el primer nodo con el segundo nodo para generar un nodo fusionado que comprenda perfiles de expresión representados por el primer nodo y el segundo nodo; y (e) para cada nodo en el conjunto de grupos de fusión, identificar dianas para distinguir tipos de células basándose en los perfiles de expresión de la pluralidad de dianas de células representadas por el nodo.

En algunas realizaciones, la estructura de datos de recuentos de dinas comprende una matriz de recuentos de dianas. Cada fila o cada columna de la matriz de recuentos de dianas puede comprender un número de cada diana

de una pluralidad de dianas para una células individual diferente de la pluralidad de células.

En algunas realizaciones, cada uno de la pluralidad de nodos hoja y la pluralidad de nodos no raíz y no hoja puede asociarse con un nodo padre, y cada uno del nodo raíz y la pluralidad de nodos no raíz y no hoja puede asociarse con un nodo hijo izquierdo y un nodo hijo derecho y representa perfiles de expresión representados por el nodo hijo izquierdo y el nodo hijo derecho del nodo.

En algunas realizaciones, el método comprende, antes de recibir la estructura de datos de recuentos de dianas en (a): (f) codificar con códigos de barras la pluralidad de dianas en la pluralidad de células usando una pluralidad de códigos de barras para crear una pluralidad de dianas codificadas con códigos de barras, en donde cada uno de la pluralidad de códigos de barras comprende un marcador celular y un marcador molecular, en donde las dianas codificadas con códigos de barras creadas a partir de dianas de diferentes células tienen diferentes marcadores celulares, y en donde las dianas codificadas con códigos de barras creadas a partir de dianas de una célula de la pluralidad de células tienen diferentes marcadores moleculares; (g) obtener datos de secuenciación de la pluralidad de dianas codificadas con códigos de barras; y (h) para cada una de la pluralidad de células: (1) contar el número de marcadores moleculares con secuencias distintas asociadas con cada diana de la pluralidad de dianas en los datos de secuenciación para la célula; y (2) estimar el número de cada diana de la pluralidad de dianas para la célula basándose en el número de marcadores moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación contados en (h)(1). Por ejemplo, el método puede comprender, antes de recibir la estructura de datos de recuento de dianas en (a): el paso o pasos (f) codificar estocásticamente con códigos de barras la pluralidad de dianas en la pluralidad de células usando una pluralidad de códigos de barras estocásticos para crear una pluralidad de dianas codificadas estocásticamente con códigos de barras, en donde cada una de la pluralidad de códigos de barras estocásticos comprende un marcador celular y un marcador molecular, en donde las dianas codificadas estocásticamente con códigos de barras creadas a partir de dianas de diferentes células tienen diferentes marcadores celulares, y en donde las dianas codificadas estocásticamente con códigos de barras creadas a partir de dianas de una célula de la pluralidad de células tienen diferentes marcadores moleculares; y/o (g) obtener datos de secuenciación de la pluralidad de dianas codificadas estocásticamente con códigos de barras. La recepción de la estructura de datos de recuentos de dianas puede comprender: generar una estructura de datos de recuentos de dianas a partir del número de cada diana de la pluralidad de dianas para la célula estimado en (h)(2), en donde el perfil de expresión de la célula de la pluralidad de células comprende el número de cada diana de la pluralidad de dianas para la célula estimado en (h)(2).

En algunas realizaciones, el método comprende antes de agrupar jerárquicamente los perfiles de expresión de la pluralidad de células basándose en la estructura de datos de recuentos de dianas y las distancias entre los perfiles de expresión de la pluralidad de células para generar el dendrograma que representa los perfiles de expresión de la pluralidad de células en (b): (i) determinar una estructura de datos de distancias de elementos de la matriz de recuentos de dianas, en donde la estructura de datos de distancias comprende distancias entre los perfiles de expresión de la pluralidad de células. La estructura de datos de distancias puede comprender una matriz de distancia. Cada elemento diagonal de la matriz de distancias puede tener un valor de cero. La agrupación jerárquica de los perfiles de expresión de la pluralidad de células basada en la estructura de datos de recuentos de dianas y las distancias entre los perfiles de expresión de la pluralidad de células para generar el dendrograma que representa los perfiles de expresión de la pluralidad de células en (b) puede comprender agrupar jerárquicamente los perfiles de expresión de la pluralidad de células sobre la base de la estructura de datos de recuentos de dianas y la estructura de datos de distancia. Las distancias entre los perfiles de expresión de la pluralidad de células pueden comprender distancias de correlación por pares entre los perfiles de expresión de la pluralidad de células.

En algunas realizaciones, antes de determinar la estructura de datos de distancias de los elementos de la estructura de datos de recuentos de dianas en (i), se transforma logarítmicamente la estructura de datos de recuentos de dianas en una estructura de datos de recuentos de dianas transformada logarítmicamente, en donde determinar la estructura de datos de distancias de los elementos de la estructura de datos de recuentos de dianas comprende determinar la estructura de datos de distancias de la estructura de datos de recuentos de dianas transformada logarítmicamente, y en donde agrupar jerárquicamente los perfiles de expresión de la pluralidad de células basándose en la estructura de datos de recuentos de dianas y las distancias entre los perfiles de expresión de la pluralidad de células en (b) comprende: agrupar jerárquicamente los perfiles de expresión de la pluralidad de células basándose en la estructura de datos de recuentos de dianas transformada logarítmicamente y la estructura de datos de distancias para generar el dendrograma. La transformación logarítmica de la estructura de datos de recuentos de dianas en la estructura de datos de recuentos de dianas transformada logarítmicamente puede comprender aumentar el valor de cada elemento de la estructura de datos de recuentos de dianas en un incremento (como uno).

En algunas realizaciones, agrupar jerárquicamente los perfiles de expresión de la pluralidad de células basándose en la estructura de datos de recuentos de dianas y las distancias entre los perfiles de expresión de la pluralidad de células en (b) comprende: asignar cada perfil de expresión de la pluralidad de células a un nodo de hoja diferente; y combinar iterativamente un primer nodo y un segundo nodo de la pluralidad de nodos para generar un nodo padre del primer nodo y el segundo nodo si el segundo nodo es el nodo más cercano de la pluralidad de nodos al primer nodo. La distancia entre el primer nodo y el segundo nodo es la distancia máxima entre cualquier célula con

un perfil de expresión representado por el primer nodo y cualquier célula con un perfil de expresión representado por el segundo nodo.

En algunas realizaciones, el método comprende: en cada nodo al atravesar la pluralidad de nodos del dendrograma: si la división es válida, continuar atravesando desde el nodo hasta el nodo hijo izquierdo y el nodo hijo derecho del nodo; y si la división no es válida, interrumpir el recorrido desde el nodo hasta el nodo hijo izquierdo y el nodo hijo derecho del nodo. Por lo menos una de una correlación intranodo del primer nodo y una correlación intranodo del segundo nodo puede ser mayor que la correlación intranodo del primer nodo y del segundo nodo. Una medida o una indicación de una correlación intranodo del primer nodo y una correlación intranodo del segundo nodo puede ser mayor que una correlación internodo del primer nodo y el segundo nodo. La medida de la correlación intranodo del primer nodo y la correlación intranodo del segundo nodo puede basarse en por lo menos una de: una correlación intranodo máxima del primer nodo y el segundo nodo, una correlación intranodo media del primer nodo y el segundo nodo, una correlación intranodo mediana del primer nodo y el segundo nodo, una correlación mínima intranodo del primer nodo y el segundo nodo, y cualquier combinación de las mismas. La correlación intranodo del primer nodo puede basarse en por lo menos una de: una correlación máxima intranodo del primer nodo, una correlación media intranodo del primer nodo, una correlación mediana intranodo del primer nodo, una correlación mínima intranodo del primer nodo, y cualquier combinación de las mismas. La correlación intranodo del segundo nodo puede basarse en por lo menos una de: una correlación máxima intranodo del segundo nodo, una correlación media intranodo del segundo nodo, una correlación mediana intranodo del segundo nodo, una correlación mínima intranodo del segundo nodo, y cualquier combinación de las mismas. La correlación internodo del primer nodo y el segundo nodo puede basarse en por lo menos una de: una correlación máxima internodo del primer nodo y el segundo nodo, una correlación media internodo del primer nodo y el segundo nodo, una correlación mediana internodo del primer nodo y el segundo nodo, una correlación mínima internodo del primer nodo y el segundo nodo, y cualquier combinación de las mismas.

En algunas realizaciones, determinar si la división del nodo con los nodos hijos del nodo es válida o inválida comprende: determinar que la división es válida si la distancia entre el nodo hijo izquierdo y el nodo hijo derecho está por encima de un umbral de división, y en caso contrario es inválida. La distancia entre el nodo hijo izquierdo y el nodo hijo derecho puede determinarse basándose en una prueba estadística realizada en cada diana de la pluralidad de dianas entre los perfiles de expresión representados por el nodo hijo izquierdo y el nodo hijo derecho. La prueba estadística puede comprender una prueba t de Welch. La distancia entre el nodo hijo izquierdo y el nodo hijo derecho puede determinarse basándose en el valor p máximo de la prueba estadística realizada en cada diana de la pluralidad de dianas entre cada perfil de expresión representado por el nodo hijo izquierdo y cada perfil de expresión representado por el nodo hijo derecho.

En algunas realizaciones, determinar si la división del nodo con los nodos hijos del nodo es válida o inválida comprende: determinar que la división es válida si por lo menos una de una correlación intranodo del primer nodo y una correlación intranodo del segundo nodo es mayor que una correlación internodo del primer nodo y el segundo nodo, e inválida en caso contrario. En algunas realizaciones, determinar si la división del nodo con los nodos hijos del nodo es válida o inválida comprende: determinar que la división es válida si una medida o una indicación de una correlación intranodo del primer nodo y una correlación intranodo del segundo nodo es mayor que una correlación internodo del primer nodo y el segundo nodo, e inválida en caso contrario. La medida de la correlación intranodo del primer nodo y la correlación intranodo del segundo nodo puede basarse en por lo menos una de: una correlación máxima intranodo del primer nodo y el segundo nodo, una correlación media intranodo del primer nodo y el segundo nodo, una correlación mediana intranodo del primer nodo y el segundo nodo, una correlación mínima intranodo del primer nodo y el segundo nodo, y cualquier combinación de las mismas. La correlación intranodo del primer nodo puede basarse en por lo menos una de: una correlación máxima intranodo del primer nodo, una correlación media intranodo del primer nodo, una correlación mediana intranodo del primer nodo, una correlación mínima intranodo del primer nodo, y cualquier combinación de las mismas. La correlación intranodo del segundo nodo puede basarse en por lo menos una de: una correlación máxima intranodo del segundo nodo, una correlación media intranodo del segundo nodo, una correlación mediana intranodo del segundo nodo, una correlación mínima intranodo del segundo nodo, y cualquier combinación de las mismas. La correlación internodo del primer nodo y el segundo nodo puede basarse en por lo menos una de las siguientes: una correlación máxima internodo del primer nodo y el segundo nodo, una correlación media internodo del primer nodo y el segundo nodo, una correlación mediana internodo del primer nodo y el segundo nodo, una correlación mínima internodo del primer nodo y el segundo nodo, y cualquier combinación de las mismas.

En algunas realizaciones, el método comprende: en cada nodo al atravesar la pluralidad de nodos del dendrograma: (3) añadir el nodo al conjunto de grupos de fusión si el nodo representa un perfil de expresión de una única célula. En algunas realizaciones, el método puede comprender en cada nodo al atravesar la pluralidad de nodos del dendrograma: asignar un marcador de nodo al nodo. Si el nodo representa un perfil de expresión de una única célula, el marcador de nodo del nodo comprende una designación de una única célula, de lo contrario, si el nodo es el nodo hijo izquierdo del nodo padre, el marcador de nodo del nodo comprende el marcador de nodo del nodo padre y una designación izquierda, y de lo contrario, el marcador de nodo del nodo comprende el marcador de nodo del nodo padre y una designación derecha.

En algunas realizaciones, para cada nodo en el conjunto de grupos de fusión, la identificación de las dianas para distinguir los tipos de células basándose en los perfiles de expresión de la pluralidad de dianas de las células representadas por el nodo comprende: determinar que una diferencia, entre los perfiles de expresión representados por el nodo y los perfiles de expresión representados por otro nodo en el conjunto de grupos de fusión, en los números de marcadores moleculares con secuencias distintas asociadas con las dianas para distinguir los tipos de células es mayor que un umbral de significancia.

En algunas realizaciones, el método comprende, antes de fusionar el primer nodo con el segundo nodo para generar el nodo fusionado en (d): fusionar cada tercer nodo en el conjunto de grupos de fusión que representa un perfil de expresión de una sola célula con un cuarto nodo en el conjunto de grupos de fusión si una distancia entre el tercer nodo y el cuarto nodo está dentro de un umbral de distancia de nodo. En algunas realizaciones, el método comprende clasificar la pluralidad de células basándose en los nodos del conjunto de grupos de fusión que representan perfiles de expresión de las células. El método puede comprender diseñar un ensayo de transcriptoma completo basado en las dianas para distinguir los tipos de células identificados. En algunas realizaciones, el método puede comprender diseñar un ensayo transcriptómico dirigido basado en las dianas para distinguir los tipos de células identificados.

En la presente se divulgan métodos para identificar dianas para distinguir tipos de células. En algunas realizaciones, el método comprende: (a) recibir perfiles de expresión de una pluralidad de células, en donde los perfiles de expresión comprenden un número de cada diana de una pluralidad de dianas para cada célula de la pluralidad de células; (b) agrupar jerárquicamente los perfiles de expresión de la pluralidad de células basándose en la matriz de recuentos de dianas y en las distancias entre los perfiles de expresión de la pluralidad de células para generar un dendrograma que represente los perfiles de expresión de la pluralidad de células, en donde el dendrograma comprenda una pluralidad de nodos, en donde la pluralidad de nodos comprende un nodo raíz, una pluralidad de nodos hoja, y una pluralidad de nodos no raíz y no hoja, en donde cada nodo hoja de la pluralidad de nodos hoja representa un perfil de expresión de una célula diferente de la pluralidad de células, y en donde el nodo raíz representa perfiles de expresión de la pluralidad de células; (c) mientras atraviesa cada nodo de la pluralidad de nodos del dendrograma desde el nodo raíz del dendrograma hasta la pluralidad de nodos hoja del dendrograma: (1) determinar si dos subramas del nodo (por ejemplo, representadas por los nodos hijos del nodo) son significativamente diferentes; y (2) si las dos subramas del nodo son significativamente diferentes, dividir el nodo en dos conjuntos de grupos (por ejemplo, atravesando a las dos subramas del nodo). En algunas realizaciones, el método comprende, (3) si la división del nodo en los nodos hijos del nodo no es válida, añadir el nodo a un conjunto de grupos de fusión. En algunas realizaciones, el método comprende: (d) iterativamente, para cada primer nodo en el conjunto de grupos de fusión, si una distancia entre el primer nodo en el conjunto de grupos de fusión y un segundo nodo en el conjunto de grupos de fusión que está más cercano al primer nodo está dentro de un umbral de distancia de fusión, fusionar el primer nodo con el segundo nodo para generar un nodo fusionado en el conjunto de grupos de fusión; y (e) para cada nodo en el conjunto de grupos de fusión, identificar dianas para distinguir tipos de células basándose en perfiles de expresión de la pluralidad de dianas de células representadas por el nodo.

En la presente se divulgan métodos para identificar dianas para distinguir tipos de células. En algunas realizaciones, el método comprende: (a) recibir perfiles de expresión de una pluralidad de células, en donde los perfiles de expresión comprenden un número de cada diana de una pluralidad de dianas para cada célula de la pluralidad de células; (b) agrupar los perfiles de expresión de la pluralidad de células para generar una pluralidad de grupos de perfiles de expresión sobre la base de las distancias entre los perfiles de expresión de la pluralidad de células, en donde cada grupo tiene una o más asociaciones con uno o ambos de (1) un grupo padre y (2) dos o más grupos hijos, en donde el grupo padre representa perfiles de expresión de una o más células de la pluralidad de células representadas por el grupo, y en donde el grupo representa perfiles de expresión representados por los dos o más grupos hijos; (c) para cada grupo con dos o más grupos hijo, si las asociaciones entre el grupo con los dos o más grupos hijos no son válidas (por ejemplo, las diferencias entre los dos o más grupos hijo no son significativas), añadir al grupo un conjunto de grupos de fusión; (d) iterativamente, para cada primer grupo en el conjunto de grupos fusionados, si una distancia entre el primer grupo en el conjunto de grupos de fusión y un segundo grupo en el conjunto de grupos de fusión que es el más cercano al primer grupo está dentro de un umbral de distancia de fusión, fusionar el primer grupo y el segundo grupo para generar un grupo fusionado, en donde el grupo fusionado comprende perfiles de expresión del primer grupo y del segundo grupo; y (e) para cada grupo del conjunto de grupos de fusión, identificar dianas para distinguir tipos de células basándose en los perfiles de expresión de la pluralidad de dianas de las células representadas por el grupo.

En algunas realizaciones, recibir perfiles de expresión de la pluralidad de células comprende recibir una estructura de datos de recuentos de dianas. La estructura de datos de recuentos de dianas puede comprender una matriz de recuentos de dianas. Cada fila o cada columna de la matriz de recuentos de dianas puede comprender un perfil de expresión de una célula individual diferente de la pluralidad de células. La agrupación de los perfiles de expresión de la pluralidad de células en la pluralidad de grupos de perfiles de expresión basándose en las distancias entre los perfiles de expresión de la pluralidad de células puede comprender: agrupar jerárquicamente los perfiles de expresión de la pluralidad de células para generar un dendrograma que represente los perfiles de expresión de la pluralidad de células basándose en las distancias entre los perfiles de expresión de la pluralidad de células. El dendrograma puede comprender una pluralidad de grupos. La pluralidad de grupos puede comprender un grupo de

raíz, una pluralidad de grupos de hoja y una pluralidad de grupos que no raíz y no hoja. Un grupo de hoja puede representar un perfil de expresión de una célula. Un grupo no raíz y no hoja puede representar perfiles de expresión de células representadas por los grupos hijos de los grupos no raíz y no hoja. El grupo raíz puede representar perfiles de expresión de sus grupos hijos. Cada uno de la pluralidad de grupos hoja y la pluralidad de grupos no raíz y no hoja puede tener una asociación con un grupo padre. Cada uno de los grupos raíz y la pluralidad de grupos no raíz y no hoja, puede tener asociaciones con un grupo hijo izquierdo y un grupo hijo derecho y representa los perfiles de expresión representados por el grupo hijo izquierdo y el grupo hijo derecho del grupo. El grupo raíz puede representar los perfiles de expresión de la pluralidad de células.

En algunas realizaciones, para cada grupo con dos o más grupos hijos, si las asociaciones entre el grupo con los dos o más grupos hijos no son válidas, añadir el grupo a un conjunto de grupos de fusión comprende: mientras se atraviesa cada grupo del dendrograma desde el grupo raíz del dendrograma hasta la pluralidad de grupos hoja del dendrograma: (1) determinar si las asociaciones del grupo con los grupos hijos del grupo son válidas o inválidas; y (2) si las asociaciones son inválidas, añadir el grupo a un conjunto de grupos de fusión.

En algunas realizaciones, el método comprende: antes de recibir los perfiles de expresión de la pluralidad de células en (a): (f) codificar con códigos de barras la pluralidad de dianas en la pluralidad de células usando una pluralidad de códigos de barras para crear una pluralidad de dianas codificadas con códigos de barras, en donde cada uno de la pluralidad de códigos de barras comprende un marcador celular y un marcador molecular, en donde las dianas codificadas con códigos de barras creadas a partir de dianas de diferentes células tienen diferentes marcadores celulares, y en donde las dianas codificadas con códigos de barras creadas a partir de dianas de una célula de la pluralidad de células tienen diferentes marcadores moleculares; (g) obtener datos de secuenciación de la pluralidad de dianas codificadas con códigos de barras; y (h) para cada una de la pluralidad de células: (1) contar el número de marcadores moleculares con secuencias distintas asociados con cada diana de la pluralidad de dianas en los datos de secuenciación para la célula; y (2) estimar el número de cada diana de la pluralidad de dianas para la célula basándose en el número de marcadores moleculares con secuencias distintas asociados con la diana en los datos de secuenciación contados en (h)(1). Por ejemplo, el método puede comprender: antes de recibir los perfiles de expresión de la pluralidad de células en (a): el paso o pasoso (f) codificar estocásticamente con códigos de barras la pluralidad de dianas en la pluralidad de células usando una pluralidad de códigos de barras estocásticos para crear una pluralidad de dianas codificadas estocásticamente con códigos de barras, en donde cada uno de la pluralidad de códigos de barras estocásticos comprende un marcador celular y un marcador molecular, en donde las dianas codificadas estocásticamente con códigos de barras creadas a partir de dianas de diferentes células tienen diferentes marcadores celulares, y en donde las dianas codificar estocásticamente con códigos de barras creadas a partir de dianas de una célula de la pluralidad de células tienen diferentes marcadores moleculares; y/o g) obtener datos de secuenciación de la pluralidad de dianas codificadas estocásticamente con códigos de barras.

En algunas realizaciones, el perfil de expresión de la célula de la pluralidad de células comprende el número de cada diana de la pluralidad de dianas para la célula estimado en (h)(2). En algunas realizaciones, el método comprende, antes de agrupar los perfiles de expresión de la pluralidad de células para generar la pluralidad de grupos de perfiles de expresión basándose en las distancias entre los perfiles de expresión de la pluralidad de células en (b): (i) determinar una estructura de datos de distancias de los perfiles de expresión de la pluralidad de células. La estructura de datos de distancias puede comprender una matriz de distancias de los perfiles de expresión de la pluralidad de células. Cada elemento diagonal de la matriz de distancias tiene un valor de cero. La agrupación de los perfiles de expresión de la pluralidad de células para generar la pluralidad de grupos de perfiles de expresión sobre la base de las distancias entre los perfiles de expresión de la pluralidad de células en (b) puede comprender: agrupar los perfiles de expresión de la pluralidad de células para generar la pluralidad de grupos de perfiles de expresión sobre la base de la matriz de distancia. Las distancias entre los perfiles de expresión de la pluralidad de células pueden ser distancias de correlación por pares entre los perfiles de expresión de la pluralidad de células.

En algunas realizaciones, el método comprende, antes de determinar la estructura de datos de distancias en (i), transformar logarítmicamente la estructura de datos de recuentos de dianas en una estructura de datos de recuentos de dianas transformada logarítmicamente, en donde determinar la estructura de datos de distancias de elementos de la estructura de datos de recuentos de dianas comprende determinar la estructura de datos de distancias de la estructura de datos de recuentos de dianas transformada logarítmicamente, y en donde la agrupación de los perfiles de expresión de la pluralidad de células para generar la pluralidad de grupos de perfiles de expresión sobre la base de las distancias entre los perfiles de expresión de la pluralidad de células en (b) comprende agrupar los perfiles de expresión de la pluralidad de células basándose en la estructura de datos de recuentos de dianas transformada logarítmicamente y la estructura de datos de distancias para generar la pluralidad de grupos. La transformación logarítmica de la estructura de datos de recuentos de dianas en la estructura de datos de recuentos de dianas transformada logarítmicamente puede comprender aumentar el valor de cada elemento de la estructura de datos de recuentos de dianas en un incremento. El incremento puede ser uno.

En algunas realizaciones, la agrupación de los perfiles de expresión de la pluralidad de células sobre la base de distancias entre los perfiles de expresión de la pluralidad de células en (b) comprende: asignar cada perfil de expresión de la pluralidad de células a un grupo de hoja diferente en la pluralidad de grupos; y combinar iterativamente

un primer grupo y un segundo grupo de la pluralidad de grupos para generar un grupo padre del primer grupo y del segundo grupo si el segundo grupo es el grupo más cercano de la pluralidad de grupos al primer grupo. La distancia entre el primer grupo y el segundo grupo puede ser la distancia máxima entre cualquier perfil de expresión representado por el primer grupo y cualquier perfil de expresión representado por el segundo grupo.

En algunas realizaciones, una correlación intragrupo del primer grupo y una correlación intragrupo del segundo grupo son mayores que una correlación intergrupo del primer grupo y del segundo grupo. Una medida o una indicación de una correlación intragrupo del primer grupo y una correlación intragrupo del segundo grupo es mayor que una correlación intergrupo del primer grupo y el segundo grupo. La medida de la correlación intragrupo del primer grupo y la correlación intragrupo del segundo grupo puede basarse en por lo menos una de: una correlación máxima intragrupo del primer grupo y el segundo grupo, una correlación media intragrupo del primer grupo y el segundo grupo, una correlación mediana intragrupo del primer grupo y el segundo grupo, una correlación mínima intragrupo del primer grupo y el segundo grupo, y cualquier combinación de las mismas. La correlación intragrupo del primer grupo puede basarse en por lo menos una de las siguientes: una correlación máxima intragrupo del primer grupo, una correlación media intragrupo del primer grupo, una correlación mediana intragrupo del primer grupo, una correlación mínima intragrupo del primer grupo y cualquier combinación de las mismas. La correlación intragrupo del segundo grupo puede basarse en por lo menos una de las siguientes: una correlación máxima intragrupo del segundo grupo, una correlación media intragrupo del segundo grupo, una correlación mediana intragrupo del segundo grupo, una correlación mínima intragrupo del segundo grupo y cualquier combinación de las mismas. La correlación intergrupo del primer grupo y el segundo grupo puede basarse en por lo menos una de: una correlación máxima intergrupo del primer grupo y el segundo grupo, una correlación media intergrupo del primer grupo y el segundo grupo, una correlación mediana intergrupo del primer grupo y el segundo grupo, una correlación mínima intergrupo del primer grupo y el segundo grupo, y cualquier combinación de las mismas.

En algunas realizaciones, el método comprende, en cada grupo al atravesar la pluralidad de grupos del dendrograma: si las asociaciones son válidas, continuar atravesando desde el grupo hasta el grupo hijo izquierdo y el grupo hijo derecho del grupo; y si las asociaciones no son válidas, interrumpir el recorrido desde el grupo hasta el grupo hijo izquierdo y el grupo hijo derecho del grupo. Determinar si las asociaciones del grupo con los grupos hijo del grupo son válidas o inválidas puede comprender: determinar que las asociaciones son válidas si la distancia entre el grupo hijo izquierdo y el grupo hijo derecho está por encima de un umbral de asociación, y si es lo contrario son inválidas.

En algunas realizaciones, la distancia entre el grupo hijo izquierdo y el grupo hijo derecho puede determinarse basándose en una prueba estadística realizada en cada diana de la pluralidad de dianas entre los perfiles de expresión representados por el grupo hijo izquierdo y el grupo hijo derecho. La prueba estadística puede comprender una prueba t de Welch. La distancia entre el grupo hijo izquierdo y el grupo hijo derecho puede determinarse basándose en el valor p máximo de la prueba estadística realizada en cada diana de la pluralidad de dianas entre el perfil de expresión representado por el grupo hijo izquierdo y cada perfil de expresión representado por el grupo hijo derecho.

En algunas realizaciones, determinar si las asociaciones del grupo con los grupos hijos del grupo son válidas o inválidas puede comprender: determinar que las asociaciones son válidas si por lo menos una de una correlación intragrupo del primer grupo y una correlación intragrupo del segundo grupo es mayor que una correlación intergrupo del primer grupo y el segundo grupo, y de lo contrario son inválidas. En algunas realizaciones, determinar si las asociaciones del grupo con los grupos hijos del grupo son válidas o inválidas puede comprender: determinar que las asociaciones son válidas si una medida o una indicación de una correlación intragrupo del primer grupo y una correlación intragrupo del segundo grupo es mayor que una correlación intergrupo del primer grupo y el segundo grupo. La medida de la correlación intragrupo del primer grupo y la correlación intragrupo del segundo grupo puede basarse en por lo menos una de: una correlación máxima intragrupo del primer grupo y el segundo grupo, una correlación media intragrupo del primer grupo y el segundo grupo, una correlación mediana intragrupo del primer grupo y el segundo grupo, una correlación mínima intragrupo del primer grupo y el segundo grupo, y cualquier combinación de las mismas. La correlación intragrupo del primer grupo puede basarse en por lo menos una de: una correlación máxima intragrupo del primer grupo, una correlación media intragrupo del primer grupo, una correlación mediana intragrupo del primer grupo, una correlación mínima intragrupo del primer grupo y cualquier combinación de las mismas. La correlación intragrupo del segundo grupo puede basarse en por lo menos una de: una correlación máxima intragrupo del segundo grupo, una correlación media intragrupo del segundo grupo, una correlación mediana intragrupo del segundo grupo, una correlación mínima intragrupo del segundo grupo, o y combinaciones de las mismas. La correlación intergrupo del primer grupo y el segundo grupo puede basarse en por lo menos una de: una correlación máxima intergrupo del primer grupo y el segundo grupo, una correlación media intergrupo del primer grupo y el segundo grupo, una correlación mediana intergrupo del primer grupo y el segundo grupo, una correlación mínima intergrupo del primer grupo y el segundo grupo, y cualquier combinación de las mismas.

En algunas realizaciones, el método comprende, en cada grupo cuando atraviesa la pluralidad de grupos del dendrograma: (3) añadir el grupo al conjunto de grupos de fusión si el grupo representa un perfil de expresión de una única célula. El método puede comprender, en cada grupo al atravesar la pluralidad de grupos del dendrograma: asignar un marcador de grupo al grupo. En algunas realizaciones, si el grupo representa un perfil de expresión de una

única célula, el marcador de grupo del grupo comprende una designación de una única célula, de lo contrario, si el grupo es el grupo hijo izquierdo del grupo padre, el marcador de grupo del grupo comprende el marcador de grupo del grupo padre y una designación izquierda, y de lo contrario, el marcador de grupo del grupo comprende el marcador de grupo del grupo padre y una designación derecha.

En algunas realizaciones, para cada grupo en el conjunto de grupos de fusión, la identificación de las dianas para distinguir los tipos de células basándose en los perfiles de expresión de la pluralidad de dianas de las células representadas por el grupo comprende: determinar una diferencia, entre los perfiles de expresión representados por el grupo y los perfiles de expresión representados por otro grupo en el conjunto de grupos de fusión, en números de marcadores moleculares con secuencias distintas asociadas con las dianas para distinguir los tipos de células es mayor que un umbral de significancia.

En algunas realizaciones, el método comprende, antes de fusionar el primer grupo con el segundo grupo para generar el grupo fusionado en (d): fusionar cada tercer grupo en el conjunto de grupos de fusión que representa un perfil de expresión de una única célula con un cuarto grupo en el conjunto de grupos de fusión si una distancia entre el tercer grupo y el cuarto grupo está dentro de un umbral de distancia de grupo. El método puede comprender clasificar la pluralidad de células basándose en los grupos del conjunto de grupos de fusión que representan perfiles de expresión de las células. El método puede comprender el diseño de un ensayo de transcriptoma completo basado en las dianas para distinguir los tipos de células identificados o el diseño de un ensayo de transcriptoma dirigido basado en las dianas para distinguir los tipos de células identificados.

En la presente se divulgan sistemas para identificar dianas para distinguir tipos de células. En algunas realizaciones, el sistema comprende: un procesador de hardware; y la memoria no transitoria que tiene instrucciones almacenadas en la misma, que cuando es ejecutada por el procesador de hardware hace que el procesador realice cualquiera de los métodos divulgados en la presente. En la presente se divulgan medios legibles por ordenador para identificar dianas para distinguir tipos de células. En algunas realizaciones, el medio legible por ordenador comprende código para realizar cualquiera de los métodos divulgados en la presente.

En la presente se divulgan realizaciones de un sistema para identificar dianas para distinguir tipos de células. En algunas realizaciones, el sistema comprende: memoria no transitoria configurada para almacenar instrucciones ejecutables, y un procesador de hardware en comunicación con la memoria no transitoria, el procesador de hardware programado por instrucciones ejecutables para: (a) recibir una estructura de datos de recuentos de dianas, en donde la estructura de datos de recuentos de dianas comprende perfiles de expresión de una pluralidad de células, y en donde los perfiles de expresión de la pluralidad de células comprenden un número de cada diana de una pluralidad de dianas para cada célula de la pluralidad de células; (b) agrupar jerárquicamente los perfiles de expresión de la pluralidad de células basándose en la estructura de datos de recuentos de dianas y en las distancias entre los perfiles de expresión de la pluralidad de células para generar un dendrograma que represente los perfiles de expresión de la pluralidad de células, en donde el dendrograma comprende una pluralidad de nodos, en donde la pluralidad de nodos comprende un nodo raíz, una pluralidad de nodos hoja, y una pluralidad de nodos no raíz y no hoja, en donde cada nodo hoja de la pluralidad de nodos hoja representa un perfil de expresión de una célula diferente de la pluralidad de células, y en donde el nodo raíz representa perfiles de expresión de la pluralidad de células; (c) mientras atraviesa a través de cada nodo de la pluralidad de nodos del dendrograma desde el nodo raíz del dendrograma hasta la pluralidad de nodos hoja del dendrograma: (1) determinar si una división del nodo en nodos hijos del nodo es válida o inválida; y (2) si la división del nodo en los nodos hijos del nodo es inválida, añadir el nodo a un conjunto de grupos de fusión; (d) iterativamente, para cada primer nodo en el conjunto de grupos de fusión, si una distancia entre el primer nodo en el conjunto de grupos de fusión y un segundo nodo en el conjunto de grupos de fusión que está más cerca del primer nodo está dentro de un umbral de distancia de fusión, fusionar el primer nodo con el segundo nodo para generar un nodo fusionado que comprende perfiles de expresión representados por el primer nodo y el segundo nodo; y (e) para cada nodo en el conjunto de grupos de fusión, identificar objetivos para distinguir tipos de células basándose en perfiles de expresión de la pluralidad de dianas de células representadas por el nodo.

En algunas realizaciones, la estructura de datos de recuentos de dianas comprende una matriz de recuentos de dianas. Cada fila o cada columna de la matriz de recuentos de dianas puede comprender un número de cada diana de una pluralidad de dianas para una célula individual diferente de la pluralidad de células. Cada uno de la pluralidad de nodos hoja y la pluralidad de nodos no raíz y no hoja puede asociarse con un nodo padre, y cada uno del nodo raíz y la pluralidad de nodos no raíz y no hoja puede asociarse con un nodo hijo izquierdo y un nodo hijo derecho y representa perfiles de expresión representados por el nodo hijo izquierdo y el nodo hijo derecho del nodo.

En algunas realizaciones, el procesador de hardware puede programarse para, antes de recibir la estructura de datos de recuento de dianas en (a): (f) provocar la codificación con códigos de barras de la pluralidad de dianas en la pluralidad de células usando una pluralidad de códigos de barras para crear una pluralidad de dianas codificadas con código de barras, en donde cada uno de la pluralidad de códigos de barras comprende un marcador celular y un marcador molecular, en donde las dianas codificadas con códigos de barras creadas a partir de dianas de diferentes células tienen diferentes marcadores celulares, y en donde las dianas codificadas con códigos de barras creadas a partir de dianas de una célula de la pluralidad de células tienen diferentes marcadores moleculares; (g) obtener datos

de secuenciación de la pluralidad de dianas codificadas con códigos de barras; y (h) para cada una de la pluralidad de células: (1) contar el número de marcadores moleculares con secuencias distintas asociadas con cada diana de la pluralidad de dianas en los datos de secuenciación para la célula; y (2) estimar el número de cada diana de la pluralidad de dianas para la célula basándose en el número de marcadores moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación contados en (h)(1). Para recibir la estructura de datos de recuentos de dianas, el procesador de hardware puede programarse para: generar una estructura de datos de recuento de dianas a partir del número de cada diana de la pluralidad de dianas para la célula estimado en (h)(2), en donde el perfil de expresión de la célula de la pluralidad de células comprende el número de cada diana de la pluralidad de dianas para la célula estimado en (h)(2).

En algunas realizaciones, el procesador de hardware puede programarse para, antes de agrupar jerárquicamente los perfiles de expresión de la pluralidad de células basándose en la estructura de datos de recuentos de dianas y las distancias entre los perfiles de expresión de la pluralidad de células para generar el dendrograma que representa los perfiles de expresión de la pluralidad de células en (b): (i) determinar una estructura de datos de distancias de elementos de la estructura de datos de recuentos de dianas, en donde la estructura de datos de distancias comprende distancias entre los perfiles de expresión de la pluralidad de células. La estructura de datos de distancias comprende una matriz de distancias. Cada elemento diagonal de la matriz de distancias tiene un valor de cero.

En algunas realizaciones, para agrupar jerárquicamente los perfiles de expresión de la pluralidad de células basándose en la estructura de datos de recuentos de dianas y las distancias entre los perfiles de expresión de la pluralidad de células para generar el dendrograma que representa los perfiles de expresión de la pluralidad de células en (b), el procesador de hardware puede programarse para: agrupar jerárquicamente los perfiles de expresión de la pluralidad de células basándose en la estructura de datos de recuentos de dianas y la estructura de datos de distancia. Las distancias entre los perfiles de expresión de la pluralidad de células pueden comprender distancias de correlación por pares entre los perfiles de expresión de la pluralidad de células.

En algunas realizaciones, el procesador de hardware puede programarse para, antes de determinar la estructura de datos de distancias de los elementos de la estructura de datos de recuentos de dianas en (i), transformar logarítmicamente la estructura de datos de recuentos de dianas en una estructura de datos de recuentos de dianas transformada logarítmicamente. Para determinar la estructura de datos de distancias de los elementos de la estructura de datos de recuentos de dianas, el procesador de hardware puede programarse para: determinar la estructura de datos de distancias de la estructura de datos de recuentos de dianas transformada logarítmicamente. Para agrupar jerárquicamente los perfiles de expresión de la pluralidad de células basándose en la estructura de datos de recuentos de dianas y las distancias entre los perfiles de expresión de la pluralidad de células en (b), el procesador de hardware puede programarse para: agrupar jerárquicamente los perfiles de expresión de la pluralidad de células basándose en la estructura de datos de recuentos de dianas transformada logarítmicamente y la estructura de datos de distancias para generar el dendrograma. Para transformar logarítmicamente la estructura de datos de recuentos de dianas en la estructura de datos de recuentos de dianas transformada logarítmicamente, el procesador de hardware puede programarse para: aumentar el valor de cada elemento de la estructura de datos de recuentos de dianas en un incremento. El incremento puede ser uno.

Para agrupar jerárquicamente los perfiles de expresión de la pluralidad de células basándose en la estructura de datos de recuentos de dianas y las distancias entre los perfiles de expresión de la pluralidad de células en (b), el procesador de hardware puede programarse para: asignar cada perfil de expresión de la pluralidad de células a un nodo de hoja diferente; y combinar iterativamente un primer nodo y un segundo nodo de la pluralidad de nodos para generar un nodo padre del primer nodo y el segundo nodo si el segundo nodo es el nodo más cercano de la pluralidad de nodos al primer nodo. La distancia entre el primer nodo y el segundo nodo puede ser la distancia máxima entre cualquier célula con un perfil de expresión representado por el primer nodo y cualquier célula con un perfil de expresión representado por el segundo nodo.

En algunas realizaciones, por lo menos una de las correlaciones intranodo del primer nodo y una correlación intranodo del segundo nodo puede ser mayor que una correlación internodo del primer nodo y el segundo nodo. Una medida o una indicación de una correlación intranodo del primer nodo y una correlación intranodo del segundo nodo puede ser mayor que una correlación internodo del primer nodo y el segundo nodo. La medida de la correlación intranodo del primer nodo y la correlación intranodo del segundo nodo puede basarse en por lo menos una de: una correlación máxima intranodo del primer nodo y el segundo nodo, una correlación media intranodo del primer nodo y el segundo nodo, una correlación mediana intranodo del primer nodo y el segundo nodo, una correlación mínima intranodo del primer nodo y el segundo nodo, y cualquier combinación de las mismas. La correlación intranodo del primer nodo puede basarse en por lo menos una de: una correlación máxima intranodo del primer nodo, una correlación media intranodo del primer nodo, una correlación mediana intranodo del primer nodo, una correlación mínima intranodo del primer nodo, y cualquier combinación de las mismas. La correlación intranodo del segundo nodo puede basarse en por lo menos una: una correlación máxima intranodo del segundo nodo, una correlación media intranodo del segundo nodo, una correlación mediana intranodo del segundo nodo, una correlación mínima intranodo del segundo nodo, y cualquier combinación de las mismas. La correlación internodo del primer nodo y el segundo

nodo puede basarse en por lo menos una de: una correlación máxima internodo del primer nodo y el segundo nodo, una correlación media internodo del primer nodo y el segundo nodo, una correlación mediana internodo del primer nodo y el segundo nodo, una correlación mínima internodo del primer nodo y el segundo nodo, y cualquier combinación de las mismas.

En algunas realizaciones, el procesador de hardware puede programarse para, en cada nodo cuando se atraviesa la pluralidad de nodos del dendrograma: si la división es válida, continuar atravesando desde el nodo hasta el nodo hijo izquierdo y el nodo hijo derecho del nodo; y si la división no es válida, interrumpir el recorrido desde el nodo hasta el nodo hijo izquierdo y el nodo hijo derecho del nodo. Para determinar si la división del nodo con los nodos hijos del nodo es válida o inválida, el procesador de hardware puede programarse para: determinar que la división es válida si la distancia entre el nodo hijo izquierdo y el nodo hijo derecho está por encima de un umbral de división, y de lo contrario es inválida. La distancia entre el nodo hijo izquierdo y el nodo hijo derecho puede determinarse basándose en una prueba estadística realizada en cada diana de la pluralidad de dianas entre los perfiles de expresión representados por el nodo hijo izquierdo y el nodo hijo derecho. La prueba estadística puede comprender una prueba t de Welch. La distancia entre el nodo hijo izquierdo y el nodo hijo derecho puede determinarse basándose en el valor p máximo de la prueba estadística realizada en cada diana de la pluralidad de dianas entre cada perfil de expresión representado por el nodo hijo izquierdo y cada perfil de expresión representado por el nodo hijo derecho.

En algunas realizaciones, el procesador de hardware puede programarse para, en cada nodo al atravesar la pluralidad de nodos del dendrograma: (3) añadir el nodo al conjunto de grupos de fusión si el nodo representa un perfil de expresión de una única célula. En algunas realizaciones, en cada nodo al atravesar la pluralidad de nodos del dendrograma, el procesador de hardware puede programarse para: asignar un marcador de nodo al nodo. Si el nodo representa un perfil de expresión de una única célula, el marcador de nodo del nodo puede comprender una designación de una única célula, de lo contrario, si el nodo es el nodo hijo izquierdo del nodo padre, el marcador de nodo del nodo puede comprender el marcador de nodo del nodo padre y una designación izquierda, y de lo contrario, el marcador de nodo del nodo puede comprender el marcador de nodo del nodo padre y una designación derecha.

En algunas realizaciones, para cada nodo en el conjunto de grupos de fusión, para identificar las dianas para distinguir los tipos de células basándose en los perfiles de expresión de la pluralidad de dianas de las células representadas por el nodo, el procesador de hardware puede programarse para: determinar una diferencia, entre los perfiles de expresión representados por el nodo y los perfiles de expresión representados por otro nodo en el conjunto de grupos de fusión, en los números de marcadores moleculares con secuencias distintas asociadas con las dianas para distinguir los tipos de células que es mayor que un umbral de significancia.

En algunas realizaciones, el procesador de hardware puede programarse para: antes de fusionar el primer nodo con el segundo nodo para generar el nodo fusionado en (d): fusionar cada tercer nodo en el conjunto de grupos de fusión que representa un perfil de expresión de una única célula con un cuarto nodo en el conjunto de grupos de fusión si una distancia entre el tercer nodo y el cuarto nodo está dentro de un umbral de distancia de nodo. El procesador de hardware puede programarse para: clasificar la pluralidad de células basándose en los nodos del conjunto de agrupaciones de fusión que representan perfiles de expresión de las células. El procesador de hardware puede programarse para: diseñar un ensayo de transcriptoma completo sobre la base de los objetivos para distinguir los tipos de células identificados. El procesador de hardware puede programarse para: diseñar un ensayo de transcriptoma dirigido basado en las dianas para distinguir los tipos de células identificados.

En la presente se divulgan realizaciones de un sistema para identificar objetivos para distinguir tipos de células. En algunas realizaciones, el sistema comprende: memoria no transitoria configurada para almacenar instrucciones ejecutables, y un procesador de hardware en comunicación con la memoria no transitoria, el procesador de hardware programado por instrucciones ejecutables para: (a) recibir perfiles de expresión de una pluralidad de células, en donde los perfiles de expresión comprenden un número de cada diana de una pluralidad de dianas para cada célula de la pluralidad de células; (b) agrupar los perfiles de expresión de la pluralidad de células para generar una pluralidad de grupos de perfiles de expresión sobre la base de las distancias entre los perfiles de expresión de la pluralidad de células, en donde cada grupo tiene una o más asociaciones con uno o ambas de (1) un grupo padre y (2) dos o más grupos hijos, en donde el grupo padre representa perfiles de expresión de una o más células de la pluralidad de células representadas por el grupo, y en donde el grupo representa perfiles de expresión representados por los dos o más grupos hijos; (c) para cada grupo con dos o más grupos hijos, si las asociaciones entre el grupo con los dos o más grupos hijos no son válidas, añadir el grupo a un conjunto de grupos de fusión; (d) iterativamente, para cada primer grupo en el conjunto de grupos de fusión, si una distancia entre el primer grupo en el conjunto de grupos de fusión y un segundo grupo en el conjunto de grupos de fusión que es el más cercano al primer grupo está dentro de un umbral de distancia de fusión, fusionar el primer grupo y el segundo grupo para generar un grupo fusionado, en donde el grupo fusionado comprende perfiles de expresión del primer grupo y del segundo grupo; y (e) para cada grupo del conjunto de grupos de fusión, identificar dianas para distinguir tipos de células basándose en los perfiles de expresión de la pluralidad de dianas de las células representadas por el grupo.

En algunas realizaciones, el procesador de hardware puede programarse para: recibir perfiles de expresión de la pluralidad de células comprende recibir una estructura de datos de recuentos de dianas. La estructura de datos

de recuentos de dianas puede comprender una matriz de recuentos de dianas. Cada fila o cada columna de la matriz de recuentos de dianas puede comprender un perfil de expresión de una célula individual diferente de la pluralidad de células.

En algunas realizaciones, para agrupar los perfiles de expresión de la pluralidad de células en la pluralidad de grupos de perfiles de expresión sobre la base de las distancias entre los perfiles de expresión de la pluralidad de células, el procesador de hardware puede programarse para: agrupar jerárquicamente los perfiles de expresión de la pluralidad de células para generar un dendrograma que represente los perfiles de expresión de la pluralidad de células sobre la base de las distancias entre los perfiles de expresión de la pluralidad de células, en donde el dendrograma comprende la pluralidad de grupos, en donde la pluralidad de grupos comprende un grupo raíz, una pluralidad de grupos hoja, y una pluralidad de grupos no raíz y no hoja. Cada uno de los grupos hoja y la pluralidad de grupos no raíz y no hoja puede estar asociado a un grupo padre. Cada uno de los grupos raíz y la pluralidad de grupos no raíz y no hoja pueden tener asociaciones con un grupo hijo izquierdo y un grupo hijo derecho y representan perfiles de expresión representados por el grupo hijo izquierdo y el grupo hijo derecho del grupo. El grupo raíz puede representar los perfiles de expresión de la pluralidad de células. Para cada grupo con dos o más grupos hijos, si las asociaciones entre el grupo con los dos o más grupos hijos no son válidas, añadir el grupo a un conjunto de grupos de fusión, el procesador de hardware puede programarse para, mientras atraviesa cada grupo del dendrograma desde el grupo raíz del dendrograma hasta la pluralidad de grupos hoja del dendrograma (1) determinar si las asociaciones del grupo con los grupos hijos del grupo son válidas o inválidas; y (2) si las asociaciones son inválidas, añadir el grupo a un conjunto de grupos de fusión.

En algunas realizaciones, el procesador de hardware puede programarse para, antes de recibir los perfiles de expresión de la pluralidad de células en (a): (f) codificar con códigos de barras la pluralidad de dianas en la pluralidad de células usando una pluralidad de códigos de barras para crear una pluralidad de dianas codificadas con códigos de barras, en donde cada una de la pluralidad de códigos de barras comprende un marcador celular y un marcador molecular, en donde las dianas codificadas con códigos de barras creadas a partir de dianas de diferentes células tienen diferentes marcadores celulares, y en donde las dianas codificadas con códigos de barras creadas a partir de dianas de una célula de la pluralidad de células tienen diferentes marcadores moleculares; (g) obtener datos de secuenciación de la pluralidad de dianas codificadas con códigos de barras; y (h) para cada una de la pluralidad de células: (1) contar el número de marcadores moleculares con secuencias distintas asociadas con cada diana de la pluralidad de dianas en los datos de secuenciación para la célula; y (2) estimar el número de cada diana de la pluralidad de dianas para la célula basándose en el número de marcadores moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación contados en (h)(1). El perfil de expresión de la célula de la pluralidad de células puede comprender el número de cada diana de la pluralidad de dianas para la célula estimado en (h)(2).

En algunas realizaciones, el procesador de hardware puede programarse para, antes de agrupar los perfiles de expresión de la pluralidad de células para generar la pluralidad de grupos de perfiles de expresión sobre la base de las distancias entre los perfiles de expresión de la pluralidad de células en (b): (i) determinar una estructura de datos de distancias de los perfiles de expresión de la pluralidad de células. La estructura de datos de distancias puede comprender una matriz de distancias de los perfiles de expresión de la pluralidad de células. Cada elemento diagonal de la matriz de distancias puede tener un valor de cero. Para agrupar los perfiles de expresión de la pluralidad de células para generar la pluralidad de grupos de perfiles de expresión basados en las distancias entre los perfiles de expresión de la pluralidad de células en (b), el procesador de hardware puede programarse para: agrupar los perfiles de expresión de la pluralidad de células para generar la pluralidad de grupos de perfiles de expresión sobre la base de la matriz de distancia. Las distancias entre los perfiles de expresión de la pluralidad de células pueden ser distancias de correlación por pares entre los perfiles de expresión de la pluralidad de células.

En algunas realizaciones, el procesador de hardware puede programarse para, antes de determinar la estructura de datos de distancias en (i), transformar logarítmicamente la estructura de datos de recuentos de dianas en una estructura de datos de recuentos de dianas transformada logarítmicamente. Para determinar la estructura de datos de distancias de los elementos de la estructura de datos de recuentos de dianas, el procesador de hardware puede programarse para: determinar la estructura de datos de distancias de la estructura de datos de recuentos de dianas transformada logarítmicamente. Para agrupar los perfiles de expresión de la pluralidad de células para generar la pluralidad de grupos de perfiles de expresión basados en las distancias entre los perfiles de expresión de la pluralidad de células en (b), el procesador de hardware puede programarse para: agrupar los perfiles de expresión de la pluralidad de células sobre la base de la estructura de datos de recuentos de dianas transformada logarítmicamente y la estructura de datos de distancias para generar la pluralidad de grupos. Para transformar logarítmicamente la estructura de datos de recuentos de dianas en la estructura de datos de recuentos de dianas transformada logarítmicamente, el procesador de hardware puede programarse para: aumentar el valor de cada elemento de la estructura de datos de recuentos de dianas en un incremento. El incremento puede ser uno.

En algunas realizaciones, para agrupar los perfiles de expresión de la pluralidad de células sobre la base de las distancias entre los perfiles de expresión de la pluralidad de células en (b), el procesador de hardware puede programarse para: asignar cada perfil de expresión de la pluralidad de células a un grupo de hoja diferente en la pluralidad de grupos; y combinar iterativamente un primer grupo y un segundo grupo de la pluralidad de grupos para

generar un grupo padre del primer grupo y el segundo grupo si el segundo grupo es el grupo más cercano de la pluralidad de grupos al primer grupo. La distancia entre el primer grupo y el segundo grupo puede ser la distancia máxima entre cualquier perfil de expresión representado por el primer grupo y cualquier perfil de expresión representado por el segundo grupo.

En algunas realizaciones, una correlación intragrupo del primer grupo y una correlación intragrupo del segundo grupo son mayores que una correlación intergrupo del primer grupo y del segundo grupo. Una medida o una indicación de una correlación intragrupo del primer grupo y una correlación intragrupo del segundo grupo puede ser mayor que una correlación intergrupo del primer grupo y el segundo grupo. La medida de la correlación intragrupo del primer grupo y la correlación intragrupo del segundo grupo puede basarse en por lo menos una de: una correlación máxima intragrupo del primer grupo y el segundo grupo, una correlación media intragrupo del primer grupo y el segundo grupo, una correlación mediana intragrupo del primer grupo y el segundo grupo, una correlación mínima intragrupo del primer grupo y el segundo grupo, y cualquier combinación de las mismas. La correlación intragrupo del primer grupo puede basarse en por lo menos una de las siguientes: una correlación máxima intragrupo del primer grupo, una correlación media intragrupo del primer grupo, una correlación mediana intragrupo del primer grupo, una correlación mínima intragrupo del primer grupo y cualquier combinación de las mismas. La correlación intragrupo del segundo grupo puede basarse en por lo menos una de: una correlación máxima intragrupo del segundo grupo, una correlación media intragrupo del segundo grupo, una correlación mediana intragrupo del segundo grupo, una correlación mínima intragrupo del segundo grupo y cualquier combinación de las mismas. La correlación intergrupo del primer grupo y el segundo grupo puede basarse en por lo menos una de las siguientes: una correlación máxima intergrupo del primer grupo y el segundo grupo, una correlación media intergrupo del primer grupo y el segundo grupo, una correlación mediana intergrupo del primer grupo y el segundo grupo, una correlación mínima intergrupo del primer grupo y el segundo grupo, y cualquier combinación de las mismas.

En algunas realizaciones, puede programarse el procesador de hardware para, en cada grupo cuando atraviesa la pluralidad de grupos del dendrograma: si las asociaciones son válidas, continuar atravesando desde el grupo hasta el grupo hijo izquierdo y el grupo hijo derecho del grupo; y si las asociaciones no son válidas, interrumpir el recorrido desde el grupo hasta el grupo hijo izquierdo y el grupo hijo derecho del grupo. Para determinar si las asociaciones del grupo con los grupos hijos del grupo son válidas o inválidas, el procesador de hardware puede programarse para: determinar que las asociaciones son válidas si la distancia entre el grupo hijo izquierdo y el grupo hijo derecho está por encima de un umbral de asociación, y de lo contrario son inválidas. La distancia entre el grupo hijo izquierdo y el grupo hijo derecho puede determinarse basándose en una prueba estadística realizada en cada diana de la pluralidad de dianas entre los perfiles de expresión representados por el grupo hijo izquierdo y el grupo hijo derecho. La prueba estadística puede comprender una prueba t de Welch. La distancia entre el grupo hijo izquierdo y el grupo hijo derecho puede determinarse basándose en el valor p máximo de la prueba estadística realizada en cada diana de la pluralidad de dianas entre el perfil de expresión representado por el grupo hijo izquierdo y cada perfil de expresión representado por el grupo hijo derecho.

En algunas realizaciones, el procesador de hardware puede programarse para, en cada grupo al atravesar la pluralidad de grupos del dendrograma: (3) añadir el grupo al conjunto de grupos de fusión si el grupo representa un perfil de expresión de una única célula. El procesador de hardware puede programarse para: en cada grupo al atravesar la pluralidad de grupos del dendrograma: asignar un marcador de grupo al grupo. Si el grupo representa un perfil de expresión de una única célula, el marcador de grupo del grupo comprende una designación de una única célula, de lo contrario, si el grupo es el grupo hijo izquierdo del grupo padre, el marcador de grupo del grupo comprende el marcador de grupo del grupo padre y una designación izquierda, y de lo contrario, el marcador de grupo del grupo comprende el marcador de grupo del grupo padre y una designación derecha.

En algunas realizaciones, para cada grupo en el conjunto de grupos de fusión, al identificar las dianas para distinguir los tipos de células basándose en los perfiles de expresión de la pluralidad de dianas de las células representadas por el grupo, el procesador de hardware puede programarse para: determinar que una diferencia, entre los perfiles de expresión representados por el grupo y los perfiles de expresión representados por otro grupo en el conjunto de grupos de fusión, en números de marcadores moleculares con secuencias distintas asociadas con las dianas para distinguir los tipos de células es mayor que un umbral de significancia. El procesador de hardware puede programarse para comprender, antes de fusionar el primer grupo con el segundo grupo para generar el grupo fusionado en (d): fusionar cada tercer grupo en el conjunto de grupos de fusión que representa un perfil de expresión de una única célula con un cuarto grupo en el conjunto de grupos de fusión si una distancia entre el tercer grupo y el cuarto grupo está dentro de un umbral de distancia de grupos.

En algunas realizaciones, el procesador de hardware puede programarse para: clasificar la pluralidad de células basándose en los grupos del conjunto de grupos de fusión que representan perfiles de expresión de las células. El procesador de hardware puede programarse para: diseñar un ensayo de transcriptoma completo sobre la base de las dianas para distinguir los tipos de células identificados. El procesador de hardware puede programarse para: diseñar un ensayo de transcriptoma dirigido basado en las dianas para distinguir los tipos de células identificados

.

BREVE DESCRIPCIÓN DE LOS DIBUJOS

La FIG. 1 ilustra un código de barras ejemplar no limitativo (por ejemplo, un código de barras estocástico).
La FIG. 2 muestra un flujo de trabajo ejemplar no limitativo de codificación con códigos de barras y recuento digital (por ejemplo, codificación con códigos de barras estocásticos y recuento digital).
La FIG. 3 es una ilustración esquemática que muestra un proceso ejemplar no limitativo para generar una biblioteca indexada de dianas codificadas con códigos de barras (por ejemplo, dianas codificadas con códigos de barras estocásticos) a partir de una pluralidad de dianas.
La FIG. 4 es un diagrama de flujo que muestra un método ejemplar no limitativo de identificación de dianas para distinguir tipos de células mediante la agrupación de perfiles de expresión de células usando un dendrograma.
La FIG. 5 es una ilustración esquemática de un dendrograma ejemplar.
La FIG. 6 es un diagrama de flujo que muestra un método ejemplar no limitativo para identificar dianas para distinguir tipos de células mediante la agrupación de perfiles de expresión de células.
La FIG. 7 es un diagrama de bloques de un sistema informático ilustrativo configurado para implementar métodos de la divulgación.
La FIG. 8, paneles (a)-(d) muestran gráficos ejemplares no limitativos de perfiles de expresión en un espacio bidimensional después de dividir y fusionar los perfiles de expresión de células individuales.
La FIG. 9, paneles (a)-(x) son gráficos ejemplares no limitativos de perfiles de expresión en un espacio bidimensional que muestran cómo pueden decidirse las divisiones.
La FIG. 10 muestra un gráfico ejemplar no limitativo de los perfiles de expresión en un espacio bidimensional después del quinto ciclo de división.
La FIG. 11, paneles (a)-(l) son gráficos ejemplares no limitativos de perfiles de expresión en un espacio bidimensional que muestran por qué se mantuvieron ciertas divisiones en el dendrograma para el quinto ciclo de división mostrado en la FIG. 10.
La FIG. 12, paneles (a)-(i) son gráficos ejemplares no limitativos de perfiles de expresión en un espacio bidimensional que muestran cómo pueden decidirse las fusiones.
La FIG. 13 muestra un gráfico ejemplar no limitativo de los perfiles de expresión en un espacio bidimensional después del segundo ciclo de fusión.
La FIG. 14, paneles (a)-(d) son gráficos ejemplares no limitativos de perfiles de expresión en un espacio bidimensional que muestran cómo se decidió el segundo ciclo de fusiones mostrado en la FIG. 13.
La FIG. 15, paneles (a)-(f) son gráficos que muestran un tipo ejemplar no limitativo de análisis de expresión diferencial.
La FIG. 16, paneles (a)-(o) son gráficos que muestran otro tipo ejemplar no limitativo de análisis de expresión diferencial.
La FIG. 17, paneles (a)-(g) son gráficos ejemplares no limitativos que visualizan las distancias entre grupos.
La FIG. 18, paneles (a)-(e) muestran un dendrograma ejemplar no limitativo.
La FIG. 19, paneles (a)-(s) son gráficos ejemplares no limitativos que muestran el barrido de parámetros.
La FIG. 20 es un gráfico ejemplar no limitativo que muestra cómo puede usarse el barrido de parámetros para identificar un umbral.
La FIG. 21, paneles (a)-(j) son gráficos ejemplares no limitativos que muestran los resultados de la primera división.
La FIG. 22 es un gráfico ejemplar no limitativo que ilustra el resultado de la división de los perfiles de expresión en un espacio bidimensional
La FIG. 23, paneles (a)-(d) muestran un dendrograma ejemplar no limitativo que muestra perfiles de expresión clasificados en dos grupos.
La FIG. 24 es otro gráfico ejemplar no limitativo que muestra el barrido de parámetros.

DESCRIPCIÓN DETALLADA

En la siguiente descripción detallada se hace referencia a los dibujos acompañantes, que forman parte de la presente. En los dibujos, símbolos similares típicamente identifican componentes similares, a menos que el contexto indique lo contrario. Se entenderá fácilmente que los aspectos de la presente divulgación, como se describen de manera general en la presente y se ilustran en las figuras, pueden disponerse, sustituirse, combinarse, separarse y diseñarse en una amplia variedad de configuraciones diferentes, todas las cuales se contemplan explícitamente en la presente y forman parte de la divulgación de la presente.

La cuantificación de pequeñas cantidades de ácidos nucleicos o dianas, 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, puede ser muy difícil determinar el número absoluto de moléculas de ácido nucleico (por ejemplo, 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 cadena de la polimerasa (PCR) digital. Para contar el número de moléculas pueden usarse códigos de barras (por ejemplo, códigos de barras estocásticos) con marcadores moleculares únicos (ML, también denominadas índices moleculares (MI)). Los códigos de barras con marcadores moleculares que son únicos para cada marcador celular pueden usarse para contar el número de moléculas en cada célula. Los ensayos ejemplares no limitativos para codificar con códigos de barras (por ejemplo, codificar con códigos de barras estocásticos) incluyen el ensayo Precise™ (Cellular Research, Inc. (Palo Alto, CA)), el ensayo Resolve™

(Cellular Research, Inc. (Palo Alto, CA)) o el ensayo Rhapsody™ (Cellular Research, Inc. (Palo Alto, CA)).

El ensayo Rhapsody™ puede utilizar un grupo no agotable de códigos de barras (por ejemplo, códigos de barras estocásticos) con un gran número, por ejemplo de 6561 a 65536, marcadores moleculares únicos en oligonucleótidos poli(T) para hibridar con todos los ARNm poli(A) de una muestra durante el paso de RT. Además de los marcadores moleculares, pueden usarse marcadores celulares de códigos de barras para identificar cada célula individual en cada pocillo de una placa de micropocillos. Un código de barras (por ejemplo, un código de barras estocástico) puede comprender un sitio de cebado de PCR universal. Durante la RT, las moléculas del gen diana reaccionan aleatoriamente con los códigos de barras. Cada molécula diana puede hibridar con un código de barras resultante para generar moléculas de ácido ribonucleótido complementario (ADNc) codificado con código de barras (por ejemplo, moléculas de ADNc codificadas con códigos de barras estocásticos). Después del marcado, las moléculas de ADNc codificadas con código de barras de los micropocillos de una placa de micropocillos pueden agruparse en un único tubo para amplificación por PCR y secuenciación. Pueden analizarse los datos brutos de secuenciación para obtener el número de códigos de barras (por ejemplo, códigos de barras estocásticos) con marcadores moleculares únicos.

En la presente se divulgan métodos para identificar dianas para distinguir tipos de células. En algunas realizaciones, el método comprende: (a) recibir una estructura de datos de recuentos de dianas, en donde la estructura de datos de recuentos de dianas comprende perfiles de expresión de una pluralidad de células, y en donde los perfiles de expresión de la pluralidad de células comprenden un número de cada diana de una pluralidad de dianas para cada célula de la pluralidad de células; (b) agrupar jerárquicamente los perfiles de expresión de la pluralidad de células basándose en la estructura de datos de recuentos de dianas y las distancias entre los perfiles de expresión de la pluralidad de células para generar un dendrograma que represente los perfiles de expresión de la pluralidad de células, en donde el dendrograma comprende una pluralidad de nodos, en donde la pluralidad de nodos comprende un nodo raíz, una pluralidad de nodos hoja, y una pluralidad de nodos no raíz y no hoja, en donde cada nodo hoja de la pluralidad de nodos hoja representa un perfil de expresión de una célula diferente de la pluralidad de células, y en donde el nodo raíz representa perfiles de expresión de la pluralidad de células; (c) mientras atraviesa cada nodo de la pluralidad de nodos del dendrograma desde el nodo raíz del dendrograma hasta la pluralidad de nodos hoja del dendrograma: (1) determinar si una división del nodo en nodos hijos del nodo es válida o inválida (por ejemplo, las diferencias entre los nodos hijos no son significativas); y (2) si la división del nodo en los nodos hijos del nodo no es válida, añadir el nodo a un conjunto de grupos de fusión; (d) iterativamente, para cada primer nodo en el conjunto de grupos de fusión, si una distancia entre el primer nodo en el conjunto de grupos de fusión y un segundo nodo en el conjunto de grupos de fusión que está más cerca del primer nodo está dentro de un umbral de distancia de fusión, fusionar el primer nodo con el segundo nodo para generar un nodo fusionado que comprenda perfiles de expresión representados por el primer nodo y el segundo nodo; y (e) para cada nodo en el conjunto de grupo de fusión, identificar objetivos para distinguir tipos de células basándose en los perfiles de expresión de la pluralidad de dianas de las células representadas por el nodo.

En la presente se describen sistemas de identificación de dianas para distinguir tipos de células. En algunas realizaciones, el sistema comprende: un procesador de hardware; y una memoria no transitoria que tiene instrucciones almacenadas en la misma, que cuando son ejecutadas por el procesador de hardware hacen que el procesador realice cualquiera de los métodos divulgados en la presente. En la presente se divulgan medios legibles por ordenador para identificar objetivos para distinguir tipos de células. En algunas realizaciones, el medio legible por ordenador comprende código para realizar cualquiera de los métodos divulgados en la presente.

Definiciones

A menos que se definan de otro modo, los términos técnicos y científicos usados en la presente tienen el mismo significado que el entendido comúnmente por un experto en la técnica a la que pertenece la presente divulgación. Consultar, por ejemplo, Singleton et al., Dictionary of Microbiology and Molecular Biology 2ª ed., J. Wiley & Sons (New York, N.Y. 1994); Sambrook et al., Molecular Cloning, A Laboratory Manual, Cold Spring Harbor Press (Cold Spring Harbor, N.Y. 1989). Para los propósitos de la presente divulgación, a continuación se definen los siguientes términos.

Como se usa en la presente, el término "adaptador" puede significar una secuencia que facilita 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 los marcadores espaciales, marcadores diana, marcadores de muestra, marcador de indexación, códigos de barras, códigos de barras estocásticos, o marcadores moleculares. Los adaptadores pueden ser lineales. Los adaptadores pueden adaptadores ser preadenilados. Los adaptadores pueden ser de cadena doble o sencilla. Uno o más adaptadores pueden estar situados 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 localizado en los extremos 5′ y/o 3′ de un polinucleótido puede ser capaz de hibridar con uno o más oligonucleótidos inmovilizados en una superficie. En algunas realizaciones, un adaptador puede comprender una secuencia universal. Una secuencia universal puede ser una región de secuencia de nucleótidos que es común para 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 sea complementario a la secuencia universal. De manera similar, por lo 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 por lo 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 hibridar con dicha secuencia universal. Las moléculas portadoras de secuencias 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 ácidos nucleicos diana. El uno o más cebadores universales unidos al ácido nucleico diana pueden proporcionar sitios para la hibridación de cebadores universales. El uno o más cebadores universales unidos al ácido nucleico diana pueden ser iguales o diferentes entre sí.

Como se usa en la presente, el término "asociado" o "asociado a" puede significar que dos o más especies pueden identificarse como colocalizadas en un punto en el tiempo. Una asociación puede significar que dos o más especies están o estaban dentro de un recipiente similar. Una asociación puede ser una asociación informática donde, por ejemplo, la información digital relativa a dos o más especies se almacena y puede usarse para determinar que una o más de las especies estaban colocalizadas en un punto en el tiempo. Una asociación puede ser una asociación física. En algunas realizaciones, 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 marcadores a soportes sólidos o semisólidos como perlas. Una asociación puede ser una unión covalente entre una diana y un marcador.

Como se usa en la presente, el término "complementario" puede referirse a la capacidad de emparejamiento preciso entre dos nucleótidos. Por ejemplo, si un nucleótido en una posición dada de un ácido nucleico es capaz de formar un enlace 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 de cadena sencilla puede ser "parcial", en la que sólo se unen algunos de los nucleótidos, o puede ser completa cuando existe complementariedad total entre las moléculas de cadena sencilla. Puede decirse 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. Puede decirse que una primera secuencia de nucleótidos es el "complemento inverso" de una segunda secuencia, si la primera secuencia de nucleótidos es complementaria a una secuencia que es la inversa (es decir, el orden de los nucleótidos está invertido) de la segunda secuencia. Como se usa en la presente, los términos "complemento", "complementario" y "complemento inverso" pueden usarse indistintamente. De la divulgación se entiende que si una molécula puede hibridar con otra molécula puede ser el complemento de la molécula que está hibridando.

Como se usa en la presente, 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 un número de marcadores únicos que han sido asociados con dianas en una muestra. Esta metodología estocástica transforma el problema del recuento de moléculas de uno de localizar e identificar moléculas idénticas a una serie de preguntas digitales de sí/no referentes a la detección de un conjunto de marcadores predefinidos.

Como se usa en la presente, el término "marcador" o "marcadores" 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. Un marcador puede ser un marcador total o parcialmente amplificable. Un marcador puede ser un marcador total o parcialmente secuenciable. Un marcador puede ser una porción de un ácido nucleico nativo identificable como distinto. Un marcador puede ser una secuencia conocida. Un marcador puede comprender una unión de secuencias de ácido nucleico, por ejemplo una unión de una secuencia nativa y no nativa. Como se usa en la presente, el término "marcador" puede usarse indistintamente con los términos "índice", "etiqueta" o "marcador-etiqueta". Los marcadores pueden transportar información. Por ejemplo, en varias realizaciones, los marcadores pueden usarse para determinar la identidad de una muestra, una fuente de una muestra, una identidad de una célula y/o una diana.

Como se usa en la presente, el término "depósitos no agotables" puede referirse a un grupo de códigos de barras estocásticos compuesto de muchos marcadores diferentes. Un depósito no agotable puede comprender un gran número de códigos de barras estocásticos diferentes, de tal manera que cuando el depósito no agotable se asocia a un grupo de dianas, es probable que cada diana esté asociada a un código de barras estocástico único. La unicidad de cada molécula diana marcada puede determinarse mediante la estadística de elección aleatoria, y depende del número de copias de moléculas diana idénticas en la colección en comparación con la diversidad de marcadores. El tamaño del conjunto resultante de moléculas diana marcadas puede determinarse por la naturaleza estocástica del proceso de codificación con códigos de barras, y el análisis del número de códigos de barras estocásticos detectados permite después calcular el número de moléculas diana presentes en la colección o muestra original. Cuando la

proporción 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, hay una probabilidad muy baja de que se haya marcado más de una molécula diana con un marcador dado).

Como se usa en la presente, el término "ácido nucleico" se refiere a una secuencia de polinucleótidos o a un 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, una estructura principal, azúcar o nucleobase alterados). Algunos ejemplos no limitativos de análogos incluyen: 5-bromouracilo, ácido nucleico peptídico, ácido xeno nucleico, morfolinos, ácidos nucleicos bloqueados, ácidos nucleicos glicólicos, ácidos nucleicos de treosa, dideoxinucleótidos, cordicepina, 7-deaza-GTP, fluoróforos (por ejemplo rodamina o fluoresceína enlazada al azúcar), nucleótidos que contienen tioles, nucleótidos enlazados a biotina, análogos de bases fluorescentes, islas de CpG, metil-7-guanosina, nucleótidos metilados, inosina, tiouridina, pseudouridina, dihidrouridina, queuosina y wiosina. Los términos "ácido nucleico", "polinucleótido", "polinucleótido diana" y "ácido nucleico diana" pueden usarse indistintamente.

Un ácido nucleico puede comprender una o más modificaciones (por ejemplo, una modificación de base, una modificación de la estructura principal), 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-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 además incluyen un grupo fosfato enlazado covalentemente a la porción de azúcar del nucleósido. En el caso de los nucleósidos que incluyen un azúcar pentofuranosílico, el grupo fosfato puede estar enlazado a la fracción hidroxilo 2′, 3′ o 5′ del azúcar. Al formar ácidos nucleicos, los grupos fosfato pueden enlazar covalentemente nucleósidos adyacentes entre sí para formar un compuesto polimérico lineal. A su vez, los extremos respectivos de este compuesto polimérico lineal pueden unirse adicionalmente para formar un compuesto circular; sin embargo, los compuestos lineales son generalmente adecuados. Además, los compuestos lineales pueden tener complementariedad interna de bases nucleotídicas y, por lo tanto, pueden plegarse de manera que produzcan un compuesto total o parcialmente de cadena doble. Dentro de los ácidos nucleicos, los grupos fosfato pueden denominarse comúnmente como formadores de la estructura principal internucleosídica del ácido nucleico. El enlace o estructura principal puede ser un enlace fosfodiéster 3′ a 5′.

Un ácido nucleico puede comprender una estructura principal modificada y/o enlaces internucleosídicos modificados. Las estructuras principales modificadas pueden incluir las que conservan un átomo de fósforo en la estructura principal y las que no tienen un átomo de fósforo en la estructura principal. Las estructuras principales de ácidos nucleicos modificadas adecuadas que contienen un átomo de fósforo en las mismas pueden incluir, por ejemplo, fosforotioatos, fosforotioatos quirales, fosforoditioatos, fosfotriesteres, aminoalquilfosfotriesteres, fosfonatos de metilo y otros de alquilo, como fosfonatos de 3′-alquileno, fosfonatos de 5′-alquileno, fosfonatos quirales, fosfinatos, fosforamidatos, incluyendo 3′-aminofosforamidato y aminoalquilfosforamidatos, fosforodiamidatos, tionofosforamidatos, tionoalquilfosfonatos, tionoalquilfosfotriesteres, selenofosfatos, y boranofosfatos que tienen enlaces normales 3′-5′, análogos con enlaces 2′-5′ y aquellos que tienenpolaridad invertida en los que uno o más enlaces internucleotídicos son un enlace 3′-3′, 5′-5′ o 2′-2′.

Un ácido nucleico puede comprender estructuras principales de polinucleótidos formadas por enlaces internucleosídicos de alquilo o cicloalquilo de cadena corta, enlaces internucleosídicos alquilo o cicloalquilo y heteroatómicos mixtos, o uno o más enlaces internucleosídicos heteroatómicos o heterocíclicos de cadena corta. Estos pueden incluir los que tienen enlaces morfolinos (formados en parte a partir de la porción de azúcar de un nucleósido); estructuras principales de siloxano; estructuras principales de sulfuro, sulfóxido y sulfona; estructuras principales de formacetilo y tioformacetilo; estructuras principales de metileno formacetilo y tioformacetilo; estructuras principales de riboacetilo; estructuras principales que contienen alquenos; estructuras principales de sulfamato; estructuras principales de metilenimino y metilenhidrazino; estructuras principales de sulfonato y sulfonamida; estructuras principales de amida; y otras que tienen partes componentes mixtas de N, O, S y CH2.

Un ácido nucleico puede comprender un mimético de ácido nucleico. El término "mimético" puede referirse a que incluye polinucleótidos en los que sólo el anillo de furanosa o tanto el anillo de furanosa como el enlace internucleotídico se sustituyen por grupos no de furanosa, la sustitución de solamente el anillo de furanosa puede denominarse como un sustituto de azúcar. La fracción de base heterocíclica o una fracción de base heterocíclica modificada puede mantenerse para la hibridación con un ácido nucleico diana apropiado. Uno de tales ácidos nucleicos puede ser un ácido nucleico peptídico (PNA). En un PNA, la estructura principal de azúcar de un polinucleótido puede sustituirse por una estructura principal que contenga amida, en particular una estructura principal de aminoetilglicina. Los nucleótidos pueden conservarse y unirse directa o indirectamente a los átomos de nitrógeno aza de la porción amida de la estructura principal. La estructura principal de los compuestos de PNA puede comprender dos o más unidades de aminoetilglicina enlazadas, lo que confiere al PNA una estructura principal que contiene amida. Las fracciones de base heterocíclica pueden unirse directa o indirectamente a los átomos de nitrógeno aza de la porción

amida de la estructura principal.

Un ácido nucleico puede tener una estructura 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 realizaciones, un fosforodiamidato u otro enlace internucleosídico no fosfodiéster puede sustituir a un enlace fosfodiéster.

Un ácido nucleico puede comprender unidades morfolino enlazadas (es decir, ácido nucleico morfolino) que tengan bases heterocíclicas unidas al anillo de morfolino. Los grupos de enlace pueden enlazar las unidades monoméricas de morfolino en un ácido nucleico de morfolino. Los compuestos oligoméricos no iónicos a base de morfolinos pueden tener menos interacciones no deseadas con las proteínas celulares. Los polinucleótidos a base de morfolinos pueden ser imitadores no iónicos de ácidos nucleicos. Una variedad de compuestos dentro de la clase morfolino pueden unirse usando diferentes grupos de enlace. Otra clase de miméticos de polinucleótidos puede denominarse ácidos nucleicos de ciclohexenilo (CeNA). El anillo de furanosa normalmente presente en una molécula de ácido nucleico puede sustituirse por un anillo de ciclohexenilo. Pueden prepararse y usarse monómeros de fosforamidita protegidos con DMT de CeNA para la síntesis de compuestos oligoméricos mediante química de fosforamidita. La incorporación de monómeros de CeNA en una cadena de ácido nucleico puede aumentar la estabilidad de un híbrido ADN/ARN. Los oligoadenilatos de CeNA pueden formar complejos con complementos de ácido nucleico 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 enlazado al átomo de carbono 4′ del anillo de azúcar formando de este modo un enlace 2′-C, 4′-C-oximetileno formando de este modo una fracción de azúcar bicíclica. El enlace puede ser un grupo metileno (-CH2-), que forma un puente entre el átomo de oxígeno 2′ y el átomo de carbono 4′ en donde n es 1 o 2. El LNA y los análogos del LNA pueden mostrar estabilidades térmicas dúplex muy altas con el ácido nucleico complementario (Tm=+3 a +10º C), estabilidad frente a la degradación 3′-exonucleolítica y buenas propiedades de solubilidad.

Un ácido nucleico también puede incluir modificaciones o sustituciones de nucleobases (a menudo denominadas simplemente como "bases"). Como se usan en la presente, las nucleobases "no modificadas" o "naturales" pueden incluir las bases de purina, (por ejemplo, adenina (A) y guanina (G)), y las bases de pirimidina, (por ejemplo, timina (T), citosina (C) y uracilo (U)). Las nucleobases modificadas pueden incluir otras nucleobases sintéticas y naturales, como la 5-metilcitosina (5-me-C), la 5-hidroximetilcitosina, la xantina, la hipoxantina, la 2-aminoadenina, 6-metilo y otros derivados alquílicos de la adenina y la guanina, 2-propilo y otros derivados alquílicos de la adenina y la 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 de pirimidina, 6-azo uracilo, 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-deazaguanina y 7-deazaadenina y 3-deazaguanina y 3-deazaadenina. 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 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 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′,2′:4,5)pirrolo[2,3-d]pirimidin-2-ona).

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

Como se usa en la presente, 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 en un sustrato. Un dispositivo de muestra puede referirse, por ejemplo, a una máquina de clasificación celular activada por fluorescencia (FACS), una máquina de clasificación celular, una aguja de biopsia, un dispositivo de biopsia, un dispositivo de seccionamiento de tejidos, un dispositivo de microfluidos, una rejilla de cuchillas y/o un micrótomo.

Como se usa en la presente, el término "soporte sólido" puede referirse a superficies sólidas o semisólidas discretas a las que puede unirse 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 plástico, cerámica, metal o material polimérico (por ejemplo, hidrogel) sobre la que puede inmovilizarse un ácido nucleico (por ejemplo, de manera 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, 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".

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 que pueden realizarse los métodos de la

divulgación. Un sustrato puede referirse a una matriz, un cartucho, un chip, un dispositivo y un portaobjetos, por ejemplo.

Como se usa en la presente, el término "marcador espacial" puede referirse a un marcador que puede asociarse a una posición en el espacio.

Como se usa en la presente, el término "código de barras estocástico" puede referirse a una secuencia de polinucleótidos que comprende marcadores. Un código de barras estocástico puede ser una secuencia de polinucleótidos que puede usarse para codificar con códigos de barras estocásticos. Los códigos de barras estocásticos pueden usarse para cuantificar dianas dentro de una muestra. Los códigos de barras estocásticos pueden usarse para controlar los errores que pueden producirse después de que un marcador se haya asociado con una diana. Por ejemplo, un código de barras estocástico puede usarse para evaluar errores de amplificación o secuenciación. Un código de barras estocástico asociado a una diana puede denominarse código de barras estocástico-diana o código de barras estocástico-etiqueta-diana.

Como se usa en la presente, el término "código de barras estocástico específico de gen" puede referirse a una secuencia de polinucleótidos que comprende marcadores y una región de unión a diana que es específica del gen. Un código de barras estocástico puede ser una secuencia de polinucleótidos que puede usarse para codificar con códigos de barras estocásticos. Los códigos de barras estocásticos pueden usarse para cuantificar dianas dentro de una muestra. Los códigos de barras estocásticos pueden usarse para controlar los errores que pueden producirse después de que un marcador se haya asociado con una diana. Por ejemplo, un código de barras estocástico puede usarse para evaluar errores de amplificación o secuenciación. Un código de barras estocástico asociado a una diana puede denominarse código de barras estocástico-diana o código de barras estocástico-etiqueta-diana.

Como se usa en la presente, el término "codificación con códigos de barras estocásticos" puede referirse al marcado aleatorio (por ejemplo, codificación con códigos de barras) de ácidos nucleicos. La codificación con códigos de barras estocásticos puede utilizar una estrategia de Poisson recursiva para asociar y cuantificar marcadores asociados a dianas. Como se usa en la presente, el término "codificación con códigos de barras estocásticos" puede usarse indistintamente con "codificación con códigos de barras estocásticos específica de gen".

Como se usa en la presente, el término "diana" puede referirse a una composición que puede asociarse con un código de barras estocástico. Las dianas ejemplares adecuadas para el análisis mediante los métodos, dispositivos y sistemas divulgados incluyen oligonucleótidos, ADN, ARN, ARNm, microARN, ARNt y similares. Las dianas pueden ser de cadena sencilla o doble. En algunas realizaciones, las dianas pueden ser proteínas. En algunas realizaciones, las dianas son lípidos.

Como se usa en la presente, el término "transcriptasas inversas" puede referirse a un grupo de enzimas que tienen actividad de transcriptasa inversa (es decir, que catalizan la síntesis de ADN a partir de una plantilla de ARN). En general, tales enzimas incluyen, entre otras, la transcriptasa inversa retroviral, la transcriptasa inversa de retrotransposones, las transcriptasas inversas de retroplásmidos, las transcriptasas inversas de retrones, las transcriptasas inversas bacterianas, las transcriptasas inversas derivadas de intrones del grupo II y mutantes, variantes o derivados de las mismas. Las transcriptasas inversas no retrovirales incluyen transcriptasas inversas de retrotransposones no LTR, transcriptasas inversas de retroplásmidos, 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 LI.LtrB de *Lactococcus lactis*, la transcriptasa inversa de intrones TeI4c de *Thrmosynechococcus elongatus* o la transcriptasa inversa de intrones GsI-IIC de *Geobacillus stearothermophilus*. 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).

En la presente se divulgan sistemas y métodos para identificar dianas para distinguir tipos de células. En algunas realizaciones, el método comprende: (a) recibir una estructura de datos de recuentos de dianas (por ejemplo, una matriz de recuentos de dianas) que comprende perfiles de expresión; (b) agrupar jerárquicamente los perfiles de expresión de la pluralidad de células para generar un dendrograma que represente los perfiles de expresión; (c) mientras se atraviesa cada nodo del dendrograma desde el nodo raíz del dendrograma hasta los nodos de hoja del dendrograma: (1) determinar si una división del nodo en nodos hijos del nodo es válida o inválida (por ejemplo, las diferencias entre los nodos hijos no son significativas); y (2) si la división del nodo en los nodos hijos del nodo no es válida, añadir el nodo a un conjunto de grupos de fusión; (d) iterativamente, para cada primer nodo en el conjunto de grupos de fusión, si una distancia entre el primer nodo en el conjunto de grupos de fusión y un segundo nodo en el conjunto de grupos de fusión que está más cerca del primer nodo está dentro de un umbral de distancia de fusión, fusionar el primer nodo con el segundo nodo para generar un nodo fusionado que comprenda perfiles de expresión representados por el primer nodo y el segundo nodo; y (e) para cada nodo en el conjunto de grupo de fusión, identificar dianas para distinguir tipos de células basándose en los perfiles de expresión de la pluralidad de dianas de las células representadas por el nodo.

Códigos de barras

La codificación con códigos de barras, como la codificación con códigos de barras estocástica, se ha descrito en, por ejemplo, la US20150299784, la WO2015031691, y Fu et al, Proc Natl Acad Sci U.S.A. 31 de mayo de 2011; 108(22):9026-31 y Fan et al., Science (2015) 347(6222):1258367. En algunas realizaciones, el código de barras divulgado en la presente puede ser un código de barras estocástico que puede ser una secuencia de polinucleótidos que puede usarse para marcar estocásticamente (por ejemplo, código de barras, etiqueta) una diana. Los códigos de barras pueden denominarse códigos de barras estocásticos si la relación entre el número de secuencias de códigos de barras diferentes de los códigos de barras estocásticos y el número de apariciones de cualquiera de las dianas que se van a marcar puede ser de, o de aproximadamente, 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, 90:1, 100:1, o un número o intervalo entre dos cualesquiera de estos valores. Una diana puede ser, por ejemplo, una especie de ARNm que comprende moléculas de ARNm con secuencias idénticas o casi idénticas. Los códigos de barras pueden denominarse códigos de barras estocásticos si la proporción entre el número de secuencias de códigos de barras diferentes de los códigos de barras estocásticos y el número de apariciones de cualquiera de las dianas que deben marcarse es por lo menos, o como máximo, de 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, 90:1 o 100:1. Las secuencias de códigos de barras de los códigos de barras estocásticos pueden denominarse marcadores moleculares.

Un código de barras, por ejemplo un código de barras estocástico, puede comprender uno o más marcadores. Los marcadores ejemplares pueden incluir un marcador universal, un marcador celular, una secuencia de código de barras (por ejemplo, un marcador molecular), un marcador de muestra, un marcador de placa, un marcador espacial y/o un marcador preespacial. La FIG. 1 ilustra un código de barras 104 ejemplar con un marcador espacial. El código de barras 104 puede comprender una amina 5′ que puede enlazar el código de barras a un soporte sólido 105. El código de barras puede comprender un marcador universal, un marcador de dimensión, un marcador espacial, un marcador celular y/o un marcador molecular. El orden de los diferentes marcadores (incluyendo, entre otros, el marcador universal, el marcador de dimensión, el marcador espacial, el marcador celular y el marcador molecular) en el código de barras puede variar. Por ejemplo, como se muestra en la FIG. 1, el marcador universal puede ser el marcador más 5', y el marcador molecular puede ser el marcador más3′. El marcador espacial, el marcador de dimensión y el marcador celular pueden estar en cualquier orden. En algunas realizaciones, el marcador universal, el marcador espacial, el marcador de dimensión, el marcador celular y el marcador molecular están en cualquier orden. El código de barras puede incluir una región de unión a la diana. La región de unión a la diana puede interactuar con una diana (por ejemplo, ácido nucleico, ARN, ARNm, ADN diana) en una muestra. Por ejemplo, una región de unión a la diana puede comprender una secuencia oligo(dT) que puede interactuar con colas poli(A) de ARNm. En algunos casos, los marcadores del código de barras (por ejemplo, el marcador universal, el marcador de dimensión, el marcador espacial, el marcador celular y la secuencia del código de barras) pueden estar separadas por 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19 o 20 nucleótidos o más.

Un marcador, por ejemplo el marcador 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 usados en algunos códigos de corrección de errores de Hamming), que pueden diseñarse para proporcionar capacidad de corrección de errores. El conjunto de subsecuencias de corrección de errores comprende siete secuencias de nucleótidos y puede diseñarse de tal manera que cualquier combinación de secuencias por pares en el conjunto muestre una "distancia genética" definida (o número de bases malapareadas); por ejemplo, puede diseñarse un conjunto de subsecuencias de corrección de errores para que muestre 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 detalladamente a continuación) puede permitir detectar o corregir errores de amplificación o secuenciación. En algunas realizaciones, la longitud de las subsecuencias de ácido nucleico usadas para crear códigos de corrección de errores puede variar, por ejemplo, pueden tener, o tener aproximadamente 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 30, 31, 40, 50, o un número o un intervalo entre dos cualesquiera de estos valores, nucleótidos de longitud. En algunas realizaciones, para crear los códigos de corrección de errores pueden usarse subsecuencias de ácido nucleico de otras longitudes.

El código de barras puede incluir una región de unión a la diana. La región de unión a la diana puede interactuar con una diana en una muestra. La diana puede ser, o comprender, ácidos ribonucleicos (ARN), ARN mensajeros (ARNm), microARN, ARN interferetes pequeños (ARNip), productos de degradación del ARN, ARN cada uno de los cuales comprende una cola poli(A), o cualquier combinación de los mismos. En algunas realizaciones, la pluralidad de dianas puede incluir ácidos desoxirribonucleicos (ADN).

En algunas realizaciones, una región de unión a la diana puede comprender una secuencia oligo(dT) que puede interactuar con colas poli(A) de ARNm. Uno o más de lo marcadores del código de barras (por ejemplo, el marcador universal, el marcador de dimensión, el marcador espacial, el marcador celular y la secuencia del código de barras (por ejemplo, un marcador molecular)) pueden estar separadas por un espaciador de otro u otros marcadores restantes del código de barras. El espaciador puede tener, 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 realizaciones, ninguno de los marcadores del código de barras está separado por un espaciador.

Marcadores universales

Un código de barras puede comprender uno o más marcadores universales. En algunas realizaciones, el uno o más marcadores universales pueden ser los mismos para todos los códigos de barras del conjunto de códigos de barras unidos a un soporte sólido dado. En algunas realizaciones, el uno o más marcadores universales pueden ser los mismos para todos los códigos de barras unidos a una pluralidad de perlas. En algunas realizaciones, un marcador universal puede comprender una secuencia de ácido nucleico que es capaz de hibridar con un cebador de secuenciación. Los cebadores de secuenciación pueden usarse para secuenciar códigos de barras que comprenden un marcador 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 realizaciones, un marcador universal puede comprender una secuencia de ácido nucleico que es capaz de hibridar con un cebador de PCR. En algunas realizaciones, el marcador universal puede comprender una secuencia de ácido nucleico capaz de hibridar con un cebador de secuenciación y un cebador de PCR. La secuencia de ácido nucleico del marcador universal que es capaz de hibridar con un cebador de secuenciación o de PCR puede denominarse sitio de unión a cebador. Un marcador universal puede comprender una secuencia que puede usarse para iniciar la transcripción del código de barras. Un marcador universal puede comprender una secuencia que puede usarse para la extensión del código de barras o de una región dentro del código de barras. Un marcador universal puede tener, o tener aproximadamente, 1, 2, 3, 4, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, o un número o un intervalo entre dos cualesquiera de estos valores, nucleótidos de longitud. Por ejemplo, un marcador universal puede comprender por lo menos aproximadamente 10 nucleótidos. Un marcador universal puede tener por lo 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 realizaciones, puede formar parte de la secuencia del marcador universal un conector escindible o nucleótido modificado para permitir que el código de barras se escinda del soporte.

Marcadores de dimensión

Un código de barras puede comprender uno o más marcadores de dimensión. En algunas realizaciones, un marcador de dimensión puede comprender una secuencia de ácido nucleico que proporciona información sobre una dimensión en la que se produjo el marcado (por ejemplo, marcado estocástico). Por ejemplo, un marcador de dimensión puede proporcionar información sobre el momento en el que la diana se codificó estocásticamente con código de barras. Un marcador de dimensión puede asociarse a un momento de codificación con código de barras (por ejemplo, codificación con código de barras estocástico) en una muestra. Un marcador de dimensión puede activarse en el momento del marcado. En diferentes momentos pueden activarse diferentes marcadores de dimensión. El marcador de dimensión proporciona información sobre el orden en el que se codificaron estocásticamente con códigos de barras las dianas, grupos de dianas y/o muestras. Por ejemplo, una población de células puede codificarse estocásticamente con códigos de barras en la fase GO del ciclo celular. Las células pueden ser pulsadas de nuevo con códigos de barras (por ejemplo, códigos de barras estocásticos) en la fase G1 del ciclo celular. Las células pueden ser pulsadas de nuevo con códigos de barras en la fase S del ciclo celular, y así sucesivamente. Los códigos de barras en cada pulso (por ejemplo, cada fase del ciclo celular), pueden comprender diferentes marcadores de dimensión. De esta manera, el marcador de dimensión proporciona información sobre qué dianas se marcaron y en qué fase del ciclo celular. Los marcadores de dimensión pueden interrogar muchos momentos biológicos diferentes. Los momentos 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, puede marcarse estocásticamente una muestra (por ejemplo, una célula, una población de células) antes y/o después del tratamiento con un fármaco y/o terapia. Los cambios en el número de copias de distintas dianas pueden ser indicativos de la respuesta de la muestra al fármaco y/o terapia.

Un marcador de dimensión puede ser activable. Un marcador de dimensión activable puede activarse en un momento específico. El marcador activable puede activarse, por ejemplo, constitutivamente (por ejemplo, no apagarse). El marcador de dimensión activable puede activarse, por ejemplo, reversiblemente (por ejemplo, el marcador de dimensión activable puede encenderse y apagarse). El marcador de dimensión puede ser activable, por ejemplo, reversiblemente por lo menos 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, o más veces. El marcador de dimensión puede ser activable reversiblemente, por ejemplo, por lo menos 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 o más veces. En algunas realizaciones, el marcador de dimensión puede activarse con fluorescencia, luz, un evento químico (por ejemplo, escisión, ligadura de otra molécula, adición de modificaciones (por ejemplo, pegilarse, sumoilarse, acetilarse, metilarse, desacetilarse, desmetilarse), un evento fotoquímico (por ejemplo, fotocaptura), y la introducción de un nucleótido no natural.

El marcador de dimensión puede, en algunas realizaciones, ser idéntico para todos los códigos de barras (por ejemplo, códigos de barras estocásticos) unidos a un soporte sólido dado (por ejemplo, una perla), pero diferente para diferentes soportes sólidos (por ejemplo, perlas). En algunas realizaciones, por lo menos el 60%, 70%, 80%, 85%, 90%, 95%, 97%, 99% o 100% de los códigos de barras en el mismo soporte sólido pueden comprender el mismo marcador de dimensión. En algunas realizaciones, por lo menos el 60% de los códigos de barras en el mismo soporte sólido pueden comprender el mismo marcador de dimensión. En algunas realizaciones, por lo menos el 95% de los códigos de barras en el mismo soporte sólido pueden comprender el mismo marcador de dimensión.

En una pluralidad de soportes sólidos (por ejemplo, perlas) puede haber representadas tantas como 10^6 o más secuencias de marcadores de dimensión únicas. Un marcador de dimensión puede tener, o tener aproximadamente 1, 2, 3, 4, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, o un número o un intervalo entre dos cualesquiera de estos valores, nucleótidos de longitud. Un marcador de dimensión puede tener por lo 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. Un marcador de dimensión puede comprender entre aproximadamente 5 y aproximadamente 200 nucleótidos. Un marcador de dimensión puede comprender entre aproximadamente 10 y aproximadamente 150 nucleótidos. Un marcador de dimensión puede comprender entre aproximadamente 20 y aproximadamente 125 nucleótidos de longitud.

Marcadores espaciales

Un código de barras puede comprender uno o más marcadores espaciales. En algunas realizaciones, un marcador espacial puede comprender una secuencia de ácido nucleico que proporciona información sobre la orientación espacial de una molécula diana asociada al código de barras. Un marcador espacial puede asociarse con una coordenada en una muestra. La coordenada puede ser una coordenada fija. Por ejemplo, una coordenada puede ser fija con respecto a un sustrato. Un marcador espacial puede estar en referencia a una cuadrícula bi- o tridimensional. Una coordenada puede fijarse en referencia a un punto de referencia. El punto de referencia puede ser identificable en el espacio. Un punto de referencia puede ser una estructura de la que pueden obtenerse imágenes. Un punto de referencia puede ser una estructura biológica, por ejemplo un punto de referencia 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 colores, código de barras, propiedad magnética, fluorescentes, radiactividad, o un tamaño o forma únicos. Un marcador espacial puede asociarse a una división física (por ejemplo, un pocillo, un recipiente o una gotita). En algunas realizaciones, se usan múltiples marcadores espaciales juntos para codificar una o más posiciones en el espacio.

El marcador espacial puede ser idéntico para todos los códigos de barras unidos a un soporte sólido determinado (por ejemplo, una perla), pero diferente para diferentes soportes sólidos (por ejemplo, perlas). En algunas realizaciones, el porcentaje de códigos de barras en el mismo soporte sólido que comprenden el mismo marcador espacial puede ser, o ser aproximadamente, del 60%, 70%, 80%, 85%, 90%, 95%, 97%, 99%, 100%, o un número o un intervalo entre dos cualesquiera de estos valores. En algunas realizaciones, el porcentaje de códigos de barras en el mismo soporte sólido que comprenden el mismo marcador espacial puede ser como mínimo, o como máximo, del 60%, 70%, 80%, 85%, 90%, 95%, 97%, 99% o 100%. En algunas realizaciones, por lo menos el 60% de los códigos de barras en el mismo soporte sólido pueden comprender el mismo marcador espacial. En algunas realizaciones, por lo menos el 95% de los códigos de barras del mismo soporte sólido pueden comprender el mismo marcador espacial.

En una pluralidad de soportes sólidos (por ejemplo, perlas) puede haber representadas tantas como 10^6 o más secuencias de marcadores espaciales únicas. Un marcador espacial puede tener, o tener aproximadamente, 1, 2, 3, 4, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, o un número o un intervalo entre dos cualesquiera de estos valores, nucleótidos de longitud. Un marcador espacial puede tener por lo 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. Un marcador espacial puede comprender entre aproximadamente 5 y aproximadamente 200 nucleótidos, por ejemplo, entre aproximadamente 10 y aproximadamente 150 nucleótidos. Un marcador espacial puede comprender entre aproximadamente 20 y aproximadamente 125 nucleótidos de longitud.

Marcadores celulares

Un código de barras puede comprender uno o más marcadores celulares. En algunas realizaciones, 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 realizaciones, el marcador celular es idéntico para todos los códigos de barras unidos a un soporte sólido dado (por ejemplo, perla), pero diferente para diferentes soportes sólidos (por ejemplo, perlas). En algunas realizaciones, el porcentaje de códigos de barras en el mismo soporte sólido que comprenden el mismo marcador celular puede ser, o ser de aproximadamente el 60%, 70%, 80%, 85%, 90%, 95%, 97%, 99%, 100%, o un número o un intervalo entre dos cualesquiera de estos valores. En algunas realizaciones, el porcentaje de códigos de barras en el mismo soporte sólido que comprenden el mismo marcador celular puede ser, o ser de aproximadamente el 60%, 70%, 80%, 85%, 90%, 95%, 97%, 99%, o 100%. Por ejemplo, por lo menos el 60% de los códigos de barras en el mismo soporte sólido pueden comprender el mismo marcador celular. Como otro ejemplo, por lo menos el 95% de los códigos de barras en el mismo soporte sólido pueden comprender el mismo marcador celular.

En una pluralidad de soportes sólidos (por ejemplo, perlas) puede haber representadas tantas como 10^6 o más secuencias de marcadores celulares únicas. Un marcador celular puede tener, o tener aproximadamente, 1, 2, 3, 4, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, o un número o un intervalo entre dos cualesquiera de estos valores, nucleótidos de longitud. Un marcador celular puede tener por lo 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.

Secuencias de códigos de barras

Un código de barras puede comprender una o más secuencias de códigos de barras. En algunas realizaciones, una secuencia de código de barras 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. Una secuencia de código de barras puede comprender una secuencia de ácido nucleico que proporciona un contador (por ejemplo, que proporciona una aproximación aproximada) para la aparición específica de la especie de ácido nucleico diana hibridada con el código de barras (por ejemplo, región de unión a la diana).

En algunas realizaciones, se une un conjunto diverso de secuencias de códigos de barras a un soporte sólido dado (por ejemplo, una perla). En algunas realizaciones, puede haber, o haber aproximadamente, 10^2, 10^3, 10^4, 10^5, 10^6, 10^7, 10^8, 10^9, o un número o un intervalo entre dos cualesquiera de estos valores, secuencias de marcadores moleculares únicas. Por ejemplo, una pluralidad de códigos de barras puede comprender aproximadamente 6561 secuencias de códigos de barras con secuencias distintas. Como otro ejemplo, una pluralidad de códigos de barras puede comprender aproximadamente 65536 secuencias de códigos de barras con secuencias distintas. En algunas realizaciones, puede haber por lo menos, o haber como máximo, 10^2, 10^3, 10^4, 10^5, 10^6, 10^7, 10^8, o 10^9, secuencias de código de barras únicas. Las secuencias de marcadores moleculares únicas pueden unirse a un soporte sólido determinado (por ejemplo, una perla).

Un código de barras puede tener, o tener aproximadamente, 1, 2, 3, 4, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, o un número o un intervalo entre dos cualesquiera de estos valores, nucleótidos de longitud. Un código de barras puede tener por lo 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.

Marcadores moleculares

Un código de barras estocástico puede comprender uno o más marcadores moleculares. Los marcadores moleculares pueden incluir secuencias de código de barras. En algunas realizaciones, un marcador 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. Un marcador molecular puede comprender una secuencia de ácido nucleico que proporciona un contador para la aparición específica de la especie de ácido nucleico diana hibridada con el código de barras (por ejemplo, región de unión a la diana).

En algunas realizaciones, se une un conjunto diverso de marcadores moleculares a un soporte sólido dado (por ejemplo, perla). En algunas realizaciones, puede haber, o haber aproximadamente, 10^2, 10^3, 10^4, 10^5, 10^6, 10^7, 10^8, 10^9, o un número o un intervalo de secuencias de marcadores moleculares únicas. Por ejemplo, una pluralidad de códigos de barras puede comprender aproximadamente 6561 marcadores moleculares con secuencias distintas. Como otro ejemplo, una pluralidad de códigos de barras puede comprender aproximadamente 65536 marcadores moleculares con secuencias distintas. En algunas realizaciones, puede haber por lo menos, o como máximo, 10^2, 10^3, 10^4, 10^5, 10^6, 10^7, 10^8, o 10^9, secuencias de marcadores moleculares únicas. Los códigos de barras con secuencias de marcadores moleculares únicas pueden unirse a un soporte sólido dado (por ejemplo, perla).

Para la codificación con códigos de barras estocásticos usando una pluralidad de códigos de barras estocásticos, la proporción entre el número de secuencias de marcadores moleculares diferentes y el número de apariciones de cualquiera de las dianas puede ser, o ser de aproximadamente, 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, 90:1, 100:1, o un número o intervalo entre dos cualesquiera de estos valores. Una diana puede ser una especie de ARNm que comprende moléculas de ARNm con secuencias idénticas o casi idénticas. En algunas realizaciones, la proporción entre el número de secuencias de marcadores moleculares diferentes y el número de apariciones de cualquiera de las dianas es por lo menos, o como máximo, de 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, 90:1, o 100:1.

Un marcador molecular puede tener, o tener aproximadamente, 1, 2, 3, 4, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, o un número o un intervalo entre dos cualesquiera de estos valores, nucleótidos de longitud. Un marcador molecular puede tener por lo 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 unión a la diana

Un código de barras puede comprender una o más regiones de unión a la diana, como sondas de captura. En algunas realizaciones, una región de unión a la diana puede hibridar con una diana de interés. En algunas

realizaciones, las regiones de unión a la diana pueden comprender una secuencia de ácido nucleico que hibrida específicamente con una diana (por ejemplo, ácido nucleico diana, molécula diana, por ejemplo, un ácido nucleico celular que se va a analizar), por ejemplo con una secuencia génica específica. En algunas realizaciones, una región de unión a la diana puede comprender una secuencia de ácido nucleico que puede unirse (por ejemplo, hibridar) con una localización específica de un ácido nucleico diana específico. En algunas realizaciones, la región de unión a la diana puede comprender una secuencia de ácido nucleico que es capaz de hibridación específica con un saliente de sitio de enzima de restricción (por ejemplo, un saliente de extremo pegajoso de EcoRI). El código de barras puede entonces ligarse a cualquier molécula de ácido nucleico que comprenda una secuencia complementaria al saliente del sitio de restricción.

En algunas realizaciones, una región de unión a la 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 a la diana puede comprender una secuencia multimérica aleatoria, o una secuencia oligo(dT) que hibrida con la cola poli(A) de las moléculas de ARNm. Una secuencia multimérica aleatoria puede ser, por ejemplo, un dímero, trímero, cuatrámero, pentámero, hexámero, septámero, octámero, nonámero, decámero o una secuencia multimérica superior aleatoria de cualquier longitud. En algunas realizaciones, la región de unión a la diana es la misma para todos los códigos de barras unidos a una perla dada. En algunas realizaciones, las regiones de unión a la diana para la pluralidad de códigos de barras unidos a una perla dada pueden comprender dos o más secuencias de unión a la diana diferentes. Una región de unión a la diana puede tener, o tener aproximadamente, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, o un número o intervalo entre dos cualesquiera de estos valores, nucleótidos de longitud. Una región de unión a la diana puede tener como máximo aproximadamente 5, 10, 15, 20, 25, 30, 35, 40, 45, 50 o más nucleótidos de longitud.

En algunas realizaciones, una región de unión a la diana puede comprender un oligo(dT) que puede hibridar con ARNm que comprenden extremos poliadenilados. Una región de unión a la diana puede ser específica de un gen. Por ejemplo, una región de unión a la diana puede configurarse para que hibride con una región específica de una diana. Una región de unión a diana puede tener, o tener 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 intervalo entre dos cualesquiera de estos valores, nucleótidos de longitud. Una región de unión a la diana puede tener por lo 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 a la diana puede tener aproximadamente 5-30 nucleótidos de longitud. Cuando un código de barras comprende una región de unión a la diana específica de un gen, el código de barras puede denominarse en la presente código de barras específico de un gen.

Propiedad de orientación

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

Un código de barras puede comprender una o más propiedades de afinidad. Por ejemplo, un marcador espacial puede incluir una propiedad de afinidad. Una propiedad de afinidad puede incluir una fracción química y/o biológica que puede facilitar la unión del código de barras a otra entidad (por ejemplo, un receptor celular). Por ejemplo, una propiedad de afinidad puede comprender un anticuerpo, por ejemplo, un anticuerpo específico para una fracción específica (por ejemplo, receptor) en una muestra. En algunas realizaciones, el anticuerpo puede guiar al código de barras a un tipo de célula o molécula específica. Las dianas en y/o cerca del tipo de célula o molécula específica pueden estar marcadas estocásticamente. En algunas realizaciones, la propiedad de afinidad puede proporcionar información espacial además de la secuencia de nucleótidos del marcador espacial, ya que el anticuerpo puede guiar al código de barras a una localización 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.

El anticuerpo puede ser una molécula de inmunoglobulina de longitud completa (es decir, de origen natural o formada por procesos recombinatorios de fragmentos de genes de inmunoglobulina normales) (por ejemplo, un anticuerpo de IgG) o una porción inmunológicamente activa (es decir, de unión específica) de una molécula de

inmunoglobulina, como un fragmento de anticuerpo.

El fragmento de anticuerpo puede ser, por ejemplo, una porción de un anticuerpo como F(ab')2, Fab', Fab, Fv, sFv y similares. En algunas realizaciones, el fragmento de anticuerpo puede unirse al mismo antígeno que es reconocido por el anticuerpo de longitud completa. El fragmento de anticuerpo puede incluir fragmentos aislados que consisten en las regiones variables de los anticuerpos, como los fragmentos "Fv" que consisten en las regiones variables de las cadenas pesada y ligera y moléculas de polipéptidos recombinantes de cadena sencilla en las que las regiones variables ligera y pesada están conectadas por un conector peptídico ("proteínas scFv"). Los anticuerpos ejemplares pueden incluir, pero no se limitan a, anticuerpos contra células cancerosas, anticuerpos contra virus, anticuerpos que se unen a receptores de la superficie celular (CD8, CD34, CD45) y anticuerpos terapéuticos.

Cebador de adaptador universal

Un código de barras puede comprender uno o más cebadores de adaptadores universales. Por ejemplo, un código de barras específico de un gen, como un código de barras estocástico específico de un gen, puede comprender un cebador de adaptador universal. Un cebador de adaptador universal puede referirse a una secuencia de nucleótidos que es universal en todos los códigos de barras. Un cebador de adaptador universal puede usarse para construir códigos de barras específicos de genes. Un cebador de adaptador universal puede tener, o tener 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 intervalo entre dos cualesquiera de estos nucleótidos de longitud. Un cebador de adaptador universal puede tener por lo 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 de adaptador universal puede tener una longitud de 5-30 nucleótidos.

Conector

Cuando un código de barras comprende más de un tipo de marcador (por ejemplo, más de un marcador celular o más de una secuencia de código de barras, como un marcador molecular), los marcadores pueden estar intercalados con una secuencia de marcador de conector. Una secuencia de marcador de conector puede tener por lo menos aproximadamente 5, 10, 15, 20, 25, 30, 35, 40, 45, 50 o más nucleótidos de longitud. Una secuencia de marcador de conector puede tener como máximo aproximadamente 5, 10, 15, 20, 25, 30, 35, 40, 45, 50 o más nucleótidos de longitud. En algunos casos, una secuencia de marcador de conector tiene 12 nucleótidos de longitud. Una secuencia de marcador de conector puede usarse para facilitar la síntesis del código de barras. El marcador de conector puede comprender un código de corrección de errores (por ejemplo, Hamming).

Soportes sólidos

En algunas realizaciones, los códigos de barras, como los códigos de barras estocásticos, divulgados en la presente pueden asociarse con un soporte sólido. El soporte sólido puede ser, por ejemplo, una partícula sintética. En algunas realizaciones, algo o parte de la secuencia de código de barras, como marcadores moleculares para códigos de barras estocásticos (por ejemplo, las primeras secuencias de códigos de barras) de una pluralidad de códigos de barras (por ejemplo, la primera pluralidad de códigos de barras) sobre un soporte sólido difieren en por lo menos un nucleótido. Los marcadores celulares de los códigos de barras en el mismo soporte sólido pueden ser iguales. Los marcadores celulares de los códigos de barras en diferentes soportes sólidos pueden diferir en por lo menos un nucleótido. Por ejemplo, los primeros marcadores celulares de una primera pluralidad de códigos de barras en un primer soporte sólido pueden tener la misma secuencia, y los segundos marcadores celulares de una segunda pluralidad de códigos de barras en un segundo soporte sólido pueden tener la misma secuencia. Los primeros marcadores celulares de la primera pluralidad de códigos de barras en el primer soporte sólido y los segundos marcadores celulares de la segunda pluralidad de códigos de barras en el segundo soporte sólido pueden diferir en por lo menos un nucleótido. Un marcador celular puede tener, por ejemplo, aproximadamente 5-20 nucleótidos de longitud. Una secuencia de código de barras puede tener, por ejemplo, aproximadamente 5-20 nucleótidos de longitud. La partícula sintética puede ser, por ejemplo, una perla.

La perla puede ser, por ejemplo, una perla de gel de sílice, una perla de vidrio de poro controlado, una perla magnética, una 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 como polidimetilsiloxano (PDMS), poliestireno, vidrio, polipropileno, agarosa, gelatina, hidrogel, paramagnético, cerámica, plástico, vidrio, metilestireno, polímero acrílico, titanio, látex, Sepharose, celulosa, nailon, silicona, o cualquier combinación de los mismos.

En algunas realizaciones, la perla puede ser una perla polimérica, por ejemplo una perla deformable o una perla de gel, funcionalizada con códigos de barras o códigos de barras estocásticos (como las perlas de gel de 10X Genomics (San Francisco, CA). En algunas implementaciones, una perla de gel puede comprender un gel a base de polímeros. Las perlas de gel pueden generarse, por ejemplo, encapsulando uno o más precursores poliméricos en gotitas. Tras la exposición de los precursores poliméricos a un acelerador (por ejemplo, tetrametiletilendiamina (TMED)), puede generarse una perla de gel.

En algunas realizaciones, la partícula puede ser degradable. Por ejemplo, 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, fundirse 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.

Los analitos y/o reactivos, como los códigos de barras de oligonucleótidos, por ejemplo, pueden acoplarse/inmovilizarse en la superficie interior de una perla de gel (por ejemplo, el interior accesible mediante difusión de un código de barras de oligonucleótidos y/o los materiales usados para generar un código de barras de oligonucleótidos) y/o la superficie exterior de una perla de gel o cualquier otra microcápsula descrita en la presente. El acoplamiento/inmovilización puede realizarse mediante cualquier forma de enlace químico (por ejemplo, enlace covalente, enlace iónico) o fenómeno físico (por ejemplo, fuerzas de Van der Waals, interacciones dipolo-dipolo, etc.). En algunas realizaciones, el acoplamiento/inmovilización de un reactivo a una perla de gel o a cualquier otra microcápsula descrita en la presente puede ser reversible como, por ejemplo, a través de una fracción lábil (por ejemplo, a través de un reticulante químico, incluyendo los reticulantes químicos descritos en la presente). Tras la aplicación de un estímulo, la fracción lábil puede escindirse y se libera el reactivo inmovilizado. En algunas realizaciones, la fracción lábil es un enlace disulfuro. Por ejemplo, en el caso de que un código de barras de oligonucleótidos se inmovilice 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 oligonucleótidos de la perla. La fracción lábil puede incluirse como parte de una perla o microcápsula de gel, como parte de un conector químico que une un reactivo o analito a una perla o microcápsula de gel, y/o como parte de un reactivo o analito. En algunas realizaciones, por lo menos un código de barras de la pluralidad de códigos de barras puede inmovilizarse en la partícula, inmovilizarse parcialmente en la partícula, encerrarse en la partícula, encerrarse parcialmente en la partícula, o cualquier combinación de los mismos.

En algunas realizaciones, una perla de gel puede comprender una amplia variedad de polímeros diferentes, que incluyen pero no se limitan a: polímeros, polímeros sensibles al calor, polímeros fotosensibles, polímeros magnéticos, polímeros sensibles al pH, polímeros sensibles a la sal, polímeros químicamente sensibles, polielectrolitos, polisacáridos, péptidos, proteínas y/o plásticos. Los polímeros pueden incluir, pero no se limitan a, materiales como poli(N-isopropilacrilamida) (PNIPAAm), poli(sulfonato de estireno) (PSS), poli(alil amina) (PAAm), poli(ácido acrílico) (PAA), poli(etileno imina) (PEI), poli(cloruro de dialildimetil-amonio) (PDADMAC), poli(pirrol) (PPy), poli(vinilpirrolidona) (PVPON), poli(vinilpiridina) (PVP), poli(ácido metacrílico) (PMAA), poli(metacrilato de metilo) (PMMA), poliestireno (PS), poli(tetrahidrofurano) (PTHF), poli(ftaladehído) (PTHF), poli(hexil viologeno) (PHV), poli(L-lisina) (PLL), poli(L-arginina) (PARG), poli(ácido láctico-co-glicólico) (PLGA).

Para desencadenar la alteración, disolución o degradación de las perlas pueden usarse numerosos estímulos químicos. Los ejemplos de estos cambios químicos pueden incluir, pero no se limitan a, cambios mediados por el pH en la pared de la perla, disgregación de la pared de la perla mediante la escisión química de los enlaces cruzados, despolimerización desencadenada de la pared de la perla y reacciones de cambio en la pared de la perla. Para desencadenar la disgregación de las perlas también pueden usarse cambios de volumen.

Los cambios físicos o de volumen de la microcápsula mediante varios estímulos también ofrecen muchas ventajas en el diseño de cápsulas para liberar reactivos. Los cambios físicos o de volumen se producen 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, pero no se limitan a, la ruptura inducida por presión, la fusión de la pared de la perla o cambios en la porosidad de la pared de la perla.

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

También puede inducirse que las perlas liberen su contenido mediante la aplicación de un estímulo térmico. Un cambio de temperatura puede provocar una variedad de cambios en las perlas. Un cambio de calor puede provocar la fusión de una perla de tal manera que se disgregue la pared de la perla. En otros casos, el calor puede aumentar la presión interna de los componentes internos de la perla de tal manera que la perla se rompa o explote. En otros casos más, el calor puede transformar la perla en un estado deshidratado encogido. El calor también puede actuar sobre los

polímeros sensibles al calor de la pared de la perla para provocar la alteración de la perla.

La inclusión de nanopartículas magnéticas en la pared de perlas de las microcápsulas puede permitir la ruptura desencadenada de las perlas, así como guiar las perlas en una matriz. Un dispositivo de esta divulgación puede comprender perlas magnéticas para ambos propósitos. En un ejemplo, la incorporación de nanopartículas de Fe_3O_4 en perlas que contienen polielectrolitos desencadena la ruptura en presencia de un estímulo de campo magnético oscilante.

Una perla también puede alterarse, disolverse o degradarse como resultado de la estimulación eléctrica. De manera similar a las partículas magnéticas descritas en la sección anterior, las perlas eléctricamente sensibles pueden permitir tanto la ruptura desencadenada de las perlas como otras funciones como la alineación en un campo eléctrico, la conductividad eléctrica o las reacciones redox. En un ejemplo, las perlas que contienen material eléctricamente sensible se alinean en un campo eléctrico de tal manera que pueda controlarse 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.

Para alterar las perlas también puede usarse un estímulo luminoso. Son posibles numerosos activadores luminosos y pueden incluir sistemas que usan varias moléculas, como nanopartículas y cromóforos, capaces de absorber fotones de intervalos específicos de longitudes de onda. Por ejemplo, como activadores de cápsulas pueden usarse recubrimientos de óxido metálico. La irradiación UV de cápsulas de polielectrolito recubiertas con SiO_2 puede provocar la disgregación de la pared de la perla. En otro ejemplo más, pueden incorporarse a la pared de la perla materiales fotoactivables, como grupos azobenceno. Tras la aplicación de luz UV o visible, sustancias químicas como estas experimentan una isomerización reversible de cis a trans tras la absorción de fotones. En este aspecto, la incorporación de interruptores de fotones da como resultado que la pared de una perla pueda disgregarse o volverse más porosa tras la aplicación de un activador luminoso.

Por ejemplo, en un ejemplo no limitativo de codificación con códigos de barras (por ejemplo, codificación con códigos de barras estocásticos) ilustrado en la FIG. 2, después de introducir células como células individuales en una pluralidad de micropocillos de una matriz de micropocillos en el bloque 208, pueden introducirse 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. Un código de barras puede comprender una región amina 5′ unida a una perla. El código de barras puede comprender un marcador universal, una secuencia de código de barras (por ejemplo, un marcador molecular), una región de unión a la diana o cualquier combinación de las mismas.

Los códigos de barras divulgados en la presente pueden asociarse (por ejemplo, unirse) a un soporte sólido (por ejemplo, una perla). Los códigos de barras asociados con un soporte sólido pueden comprender cada uno una secuencia de código de barras seleccionada de un grupo que comprende por lo menos 100 o 1000 secuencias de códigos de barras con secuencias únicas. En algunas realizaciones, diferentes códigos de barras asociados con un soporte sólido pueden comprender códigos de barras con secuencias diferentes. En algunas realizaciones, un porcentaje de los códigos de barras asociados a un soporte sólido comprende el mismo marcador celular. Por ejemplo, el porcentaje puede ser, o ser de aproximadamente el 60%, 70%, 80%, 85%, 90%, 95%, 97%, 99%, 100%, o un número o un intervalo entre dos cualesquiera de estos valores. Como otro ejemplo, el porcentaje puede ser por lo menos, o ser como máximo del 60%, 70%, 80%, 85%, 90%, 95%, 97%, 99%, o 100%. En algunas realizaciones, los códigos de barras asociados a un soporte sólido pueden tener el mismo marcador celular. Los códigos de barras asociados con diferentes soportes sólidos pueden tener diferentes marcadores celulares seleccionados de un grupo que comprende por lo menos 100 o 1000 marcadores celulares con secuencias únicas.

Los códigos de barras divulgados en la presente pueden asociarse a (por ejemplo, unirse a) un soporte sólido (por ejemplo, una perla). En algunas realizaciones, la codificación estocástica con códigos de barras de la pluralidad de dianas en la muestra puede realizarse con un soporte sólido que incluye una pluralidad de partículas sintéticas asociadas con la pluralidad de códigos de barras. En algunas realizaciones, el soporte sólido puede incluir una pluralidad de partículas sintéticas asociadas con la pluralidad de códigos de barras. Los marcadores espaciales de la pluralidad de códigos de barras en diferentes soportes sólidos pueden diferir en por lo menos un nucleótido. El soporte sólido puede, por ejemplo, incluir la pluralidad de códigos de barras 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, Dynabeads, perlas 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 matriz de agujas, un anticuerpo o cualquier combinación de los mismos. En algunas realizaciones, los soportes sólidos pueden flotar libremente. En algunas realizaciones, los soportes sólidos pueden estar incorporados en una matriz semisólida o sólida. Los códigos de barras pueden no estar asociados a los soportes sólidos. Los códigos de barras pueden ser nucleótidos individuales. Los códigos de barras pueden estar asociados a un sustrato.

Como se usan en la presente, los términos "atado", "unido" e "inmovilizado" se usan indistintamente y pueden referirse a medios covalentes o no covalentes para unir códigos de barras a un soporte sólido. Como soporte sólido para unir códigos de barras presintetizados o para la síntesis in situ en fase sólida de códigos de barras puede usarse

cualquiera de una variedad de soportes sólidos.

En algunas realizaciones, 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 que pueda inmovilizar un ácido nucleico (por ejemplo, covalente o no covalentemente). 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 tener una forma no esférica o irregular, como cúbica, cuboide, piramidal, cilíndrica, cónica, oblonga o en forma de disco, y similares. En algunas realizaciones, una perla puede tener una forma no esférica.

Las perlas pueden estar compuestas de una variedad de materiales que incluyen, pero no se limitan a, materiales paramagnéticos (por ejemplo, magnesio, molibdeno, litio y tántalo), materiales superparamagnéticos (por ejemplo, nanopartículas de ferrita (Fe_3O_4; magnetita)), materiales ferromagnéticos (por ejemplo hierro, níquel, cobalto, algunas aleaciones de los mismos, y algunos compuestos metálicos de tierras raras), cerámica, plástico, vidrio, poliestireno, sílice, metilestireno, polímeros acrílicos, titanio, látex, Sepharose, agarosa, hidrogel, polímero, celulosa, nylon, o cualquier combinación de los mismos.

En algunas realizaciones, la perla (por ejemplo, la perla a la que se unen los marcadores) es una perla de hidrogel. En algunas realizaciones, la perla comprende hidrogel.

Algunas realizaciones divulgadas en la presente incluyen una o más partículas (por ejemplo, perlas). Cada una de las partículas puede comprender una pluralidad de oligonucleótidos (por ejemplo, códigos de barras). Cada uno de la pluralidad de oligonucleótidos puede comprender una secuencia de códigos de barras (por ejemplo, un marcador molecular), un marcador 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 del marcador celular de cada uno de la pluralidad de oligonucleótidos puede ser la misma. Las secuencias del marcador celular de los oligonucleótidos en partículas diferentes pueden ser diferentes, de tal manera que puedan identificarse los oligonucleótidos en partículas diferentes. En diferentes implementaciones el número de secuencias de marcadores celulares diferentes puede ser diferente. En algunas realizaciones, el número de secuencias de marcadores celulares puede ser, o ser aproximadamente de 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 intervalo entre dos cualesquiera de estos valores, o más. En algunas realizaciones, el número de secuencias de marcadores celulares puede ser como mínimo o como máximo de 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 realizaciones, no más de 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 más de la pluralidad de las partículas incluyen oligonucleótidos con la misma secuencia celular. En algunas realizaciones, la pluralidad de partículas que incluyen oligonucleótidos con la misma secuencia celular puede ser como máximo del 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 realizaciones, ninguna de la pluralidad de partículas tiene la misma secuencia de marcador celular.

La pluralidad de oligonucleótidos en cada partícula puede comprender diferentes secuencias de códigos de barras (por ejemplo, marcadores moleculares). En algunas realizaciones, el número de secuencias de códigos de barras puede ser, o ser aproximadamente de 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 un intervalo entre dos cualesquiera de estos valores. En algunas realizaciones, el número de secuencias de códigos de barras puede ser como mínimo o como máximo de 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, por lo menos 100 de la pluralidad de oligonucleótidos comprenden diferentes secuencias de código de barras. Como otro ejemplo, en una única partícula, por lo menos 100, 500, 1000, 5000, 10000, 15000, 20000, 50000, un número o un intervalo entre dos cualesquiera de estos valores, o más de la pluralidad de oligonucleótidos comprenden secuencias de código de barras diferentes. Algunas realizaciones proporcionan una pluralidad de las partículas que comprenden códigos de barras. En algunas realizaciones, la proporción de una aparición (o una copia o un número) de una diana a marcar y las diferentes secuencias de código de barras puede ser de por lo menos 1:1, 12, 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 realizaciones, 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 una micropartícula.

El tamaño de las perlas puede variar. Por ejemplo, el diámetro de la perla puede variar de 0,1 micrómetros a 50 micrómetros. En algunas realizaciones, los diámetros de las perlas puede ser de, o ser de aproximadamente, 0,1, 0,5, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40 o 50 micrómetros, o un número o un intervalo entre dos cualesquiera de estos valores.

Los diámetros de la perla pueden estar relacionados con el diámetro de los pocillos del sustrato. En algunas

realizaciones, los diámetro de la perla pueden ser, o ser aproximadamente, un 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, 100%, o un número o un intervalo entre dos cualesquiera de estos valores, más largos o más cortos 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 realizaciones, los diámetros de la perla pueden ser por lo menos, o como máximo, un 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90% o 100% más largos o más cortos 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 realizaciones, los diámetro de las perlas pueden ser, o ser aproximadamente, un 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, 100%, 150%, 200%, 250%, 300%, o un número o un intervalo entre dos cualesquiera de estos valores, más largos o más cortos que el diámetro de la célula. En algunas realizaciones, los diámetros de las perlas puede ser como mínimo, o como máximo, un 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, 100%, 150%, 200%, 250% o 300% más largos o más cortos que el diámetro de la célula.

Una perla puede unirse y/o incorporarse en un sustrato. Una perla puede unirse y/o incorporarse en un gel, hidrogel, polímero y/o matriz. La posición espacial de una perla dentro de un sustrato (por ejemplo, gel, matriz, andamiaje o polímero) puede identificarse usando el marcador espacial presente en el código de barras de la perla, que puede servir como dirección de localización.

Los ejemplos de perlas pueden incluir, pero no se limitan a, perlas de estreptavidina, perlas de agarosa, perlas magnéticas, perlas Dynabeads®, microperlas MACS®, perlas conjugadas con anticuerpos (por ejemplo, microperlas antiinmunoglobulina), perlas conjugadas con proteína A, perlas conjugadas con proteína G, perlas conjugadas con proteína A/G, perlas conjugadas con proteína L, perlas conjugadas con oligo(dT), perlas de sílice, perlas similares a la sílice, microperlas antibiotina, microperlas antifluorocromo y perlas magnéticas BcMag™ terminadas en carboxilo.

Una perla puede estar asociada con (por ejemplo, impregnada de) puntos cuánticos o colorantes fluorescentes para hacerla fluorescente en un canal óptico o en múltiples canales ópticos de fluorescencia. Una perla puede asociarse con óxido de hierro u óxido de cromo para hacerla paramagnética o ferromagnética. Las perlas pueden ser identificables. Por ejemplo, pueden obtenerse imágenes de una perla usando una cámara. Una perla puede tener un código detectable asociado con la perla. Por ejemplo, una perla puede comprender un código de barras. 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.

Un soporte sólido (por ejemplo, una perla) puede visualizarse. El soporte sólido puede incluir una etiqueta de visualización (por ejemplo, un colorante fluorescente). Un soporte sólido (por ejemplo, una perla) puede grabarse con un identificador (por ejemplo, un número). El identificador puede visualizarse a través de imagenología de perlas.

Un soporte sólido puede comprender un material insoluble, semisoluble o insoluble. Un soporte sólido puede denominarse "funcionalizado" cuando incluye un conector, un andamiaje, una unidad estructural u otra fracción reactiva unida al mismo, mientras que un soporte sólido puede ser "no funcionalizado" cuando carece de dicha fracción reactiva unida al mismo. El soporte sólido puede emplearse libre en solución, como en un formato de pocillo de microtitulación; en un formato de flujo continuo, como en una columna; o en una tira reactiva.

El soporte sólido puede comprender una membrana, papel, plástico, superficie recubierta, superficie plana, vidrio, portaobjetos, viruta o cualquier combinación de los mismos. Un soporte sólido puede adoptar la forma de resinas, geles, perlas u otras configuraciones geométricas. Un soporte sólido puede comprender virutas de sílice, micropartículas, nanopartículas, placas, matrices, capilares, soportes planos como filtros de fibra de vidrio, superficies de vidrio, superficies metálicas (acero, oro, plata, aluminio, silicio y cobre), soportes de vidrio, soportes de plástico, soportes de silicio, virutas, filtros, membranas, placas de micropocillos, portaobjetos, materiales plásticos, incluyendo placas o membranas multipocillo (por ejemplo, formadas de polietileno, polipropileno, poliamida, polivinilidendifluoruro), y/i obleas, peines, alfileres o agujas (por ejemplo, matrices de alfileres adecuadas para la síntesis o el análisis combinatorios) o perlas en una matriz de fosas o pocillos de nanolitros de superficies planas como obleas (por ejemplo, obleas de silicio), obleas con fosas con o sin fondos filtrantes.

El soporte sólido puede comprender una matriz polimérica (por ejemplo, gel, hidrogel). La matriz polimérica puede ser capaz de permeabilizar el espacio intracelular (por ejemplo, alrededor de orgánulos). La matriz polimérica puede bombearse a través del sistema circulatorio.

Un soporte sólido puede ser una molécula biológica. Por ejemplo, un soporte sólido puede ser un ácido nucleico, una proteína, un anticuerpo, una histona, un compartimento celular, un lípido, un carbohidrato y similares. Los soportes sólidos que son moléculas biológicas pueden amplificarse, traducirse, transcribirse, degradarse y/o modificarse (por ejemplo, pegilarse, sumoilarse, acetilarse, metilarse). Un soporte sólido que es una molécula biológica puede proporcionar información espacial y temporal además del marcador espacial que está unida a la molécula biológica. Por ejemplo, una molécula biológica puede comprender una primera confirmación cuando está sin modificar, pero puede cambiar a una segunda confirmación cuando se modifica. Las diferentes conformaciones pueden exponer

los códigos de barras (por ejemplo, códigos de barras estocásticos) de la divulgación a las dianas. Por ejemplo, una molécula biológica puede comprender códigos de barras que son inaccesibles debido al plegamiento de la molécula biológica. Tras la modificación de la molécula biológica (por ejemplo, acetilación), la molécula biológica puede cambiar su conformación para exponer los códigos de barras. La cadencia de la modificación puede proporcionar otra dimensión temporal al método de codificación con códigos de barras de la divulgación.

En algunas realizaciones, la molécula biológica que comprende los reactivos de códigos de barras de la divulgación puede localizarse en el citoplasma de una célula. Tras la activación, la molécula biológica puede desplazarse al núcleo, donde puede tener lugar la codificación con códigos de barras. De esta manera, la modificación de la molécula biológica puede codificar información espacio-temporal adicional para las dianas identificadas por los códigos de barras.

Sustratos y matriz de micropocillos

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

Métodos de codificación con códigos de barras

La divulgación proporciona métodos para estimar el número de dianas distintas en localizaciones distintas en una muestra física (por ejemplo, tejido, órgano, tumor, célula). Los métodos pueden comprender colocar los códigos de barras (por ejemplo, códigos de barras estocásticos) en estrecha proximidad de la muestra, lisar la muestra, asociar distintas dianas con los códigos de barras, amplificar las dianas y/o contar digitalmente las dianas. El método puede comprender además analizar y/o visualizar la información obtenida de los marcadores espaciales de los códigos de barras. En algunas realizaciones, un método comprende visualizar la pluralidad de dianas en la muestra. Mapear la pluralidad de dianas 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 pueden generarse antes o después de codificar con códigos barras (por ejemplo, codificar estocásticamente con códigos de barras) la pluralidad de dianas en la muestra. Visualizar la pluralidad de dianas en la muestra puede incluir mapear la pluralidad de dianas en un mapa de la muestra. Mapear la pluralidad de dianas 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 pueden generarse antes o después de codificar con códigos de barras la pluralidad de dianas en la muestra. En algunas realizaciones, el mapa bidimensional y el mapa tridimensional pueden generarse 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.

En algunas realizaciones, codificar con códigos de barras la pluralidad de dianas comprende hibridar una pluralidad de códigos de barras con una pluralidad de dianas para crear dianas codificadas con código de barras (por ejemplo, dianas codificadas estocásticamente con códigos de barras). Codificar con códigos de barras la pluralidad de dianas puede comprender generar una biblioteca indexada de las dianas codificadas con códigos de barras. Generar una biblioteca indexada de las dianas codificadas con códigos de barras puede realizarse con un soporte sólido que comprende la pluralidad de códigos de barras (por ejemplo, códigos de barras estocásticos).

Poner en contacto una muestra y un código de barras

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 (por ejemplo, códigos de barras estocásticos). Las células pueden ponerse en contacto, por ejemplo, por flujo de gravedad en donde las células pueden asentarse y crear una monocapa. La muestra puede ser una sección delgada de tejido. La sección delgada puede colocarse sobre el sustrato. La muestra puede ser unidimensional (por ejemplo, forma una superficie plana). La muestra (por ejemplo, células) puede extenderse a través del sustrato, por ejemplo, haciendo crecer/cultivando las células en el sustrato.

Cuando los códigos de barras están muy cerca de las dianas, éstas pueden hibridar con el código de barras. Los códigos de barras pueden ponerse en contacto en una proporción no agotable, de tal manera que cada diana distinta pueda asociarse con un código de barras distinto de la divulgación. Para garantizar una asociación eficaz entre la diana y el código de barras, las dianas pueden reticularse con el código de barras.

Lisis celular

Después de la distribución de las células y los códigos de barras, las células pueden lisarse para liberar las moléculas diana. La lisis celular puede lograrse mediante cualquiera de una variedad de medios, por ejemplo, mediante medios químicos o bioquímicos, por choque osmótico o mediante lisis térmica, lisis mecánica o lisis óptica. Las células pueden lisarse 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 solvente 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, puede alterarse la tasa de difusión de las moléculas diana, por ejemplo, reduciendo la temperatura y/o aumentando la viscosidad del lisado.

En algunas realizaciones, la muestra puede lisarse usando un papel de filtro. El papel de filtro puede empaparse con un tampón de lisis en la parte superior del papel de filtro. El papel de filtro puede aplicarse a la muestra con presión, lo que puede facilitar la lisis de la muestra y la hibridación de las dianas de la muestra con el sustrato.

En algunas realizaciones, la lisis puede realizarse mediante lisis mecánica, lisis térmica, 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 puede realizarse mediante la adición de un tampón de lisis al sustrato. Un tampón de lisis puede comprender Tris HCl. Un tampón de lisis puede comprender por lo 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 de Tris HCl. Un tampón de lisis puede comprender aproximadamente 0,1 M de Tris HCl. El pH del tampón de lisis puede ser por lo menos de aproximadamente 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 o más. El pH del tampón de lisis puede ser como máximo de aproximadamente 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 o más. En algunas realizaciones, el pH del tampón de lisis es de 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 de por lo menos aproximadamente 0,1, 0,5 o 1 M o más. La concentración de sal en el tampón de lisis puede ser como máximo de aproximadamente 0,1, 0,5 o 1 M o más. En algunas realizaciones, la concentración de sal en el tampón de lisis es de aproximadamente 0,5 M. El tampón de lisis puede comprender un detergente (por ejemplo, SDS, dodecil sulfato de Li, Triton X, tween, NP-40). La concentración del detergente en el tampón de lisis puede ser de por lo menos aproximadamente un 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 de aproximadamente un 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 realizaciones, la concentración del detergente en el tampón de lisis es de aproximadamente un 1% de dodecil sulfato de Li. El tiempo usado en el método para la lisis puede depender de la cantidad de detergente usado. En algunas realizaciones, 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 de por lo 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 de aproximadamente 1, 5, 10, 15, 20, 25 o 30 mM o más. En algunas realizaciones, la concentración de agente quelante en el tampón de lisis es de 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 de por lo 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 de aproximadamente 1, 5, 10, 15 o 20 mM o más. En algunas realizaciones, la concentración de reactivo reductor en el tampón de lisis es de aproximadamente 5 mM. En algunas realizaciones, un tampón de lisis puede comprender aproximadamente 0,1M TrisHCl, aproximadamente pH 7,5, aproximadamente 0,5M LiCl, aproximadamente un 1% de dodecil sulfato de litio, aproximadamente 10 mM EDTA, y aproximadamente 5 mM DTT.

La lisis puede realizarse a una temperatura de aproximadamente 4, 10, 15, 20, 25 o 30º C. La lisis puede realizarse durante aproximadamente 1, 5, 10, 15 o 20 o más minutos. Una célula lisada puede comprender por lo menos aproximadamente 100000, 200000, 300000, 400000, 500000, 600000 o 700000 o más moléculas de ácido nucleico diana. Una célula lisada puede comprender como máximo aproximadamente 100000, 200000, 300000, 400000, 500000, 600000 o 700000 o más moléculas de ácido nucleico diana.

Unión de códigos de barras a moléculas de ácido nucleico diana

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 del soporte sólido colocalizado. La asociación puede comprender la hibridación de la región de reconocimiento de la diana de un código de barras con una porción complementaria de la molécula de ácido nucleico diana (por ejemplo, el oligo(dT) del código de barras puede interactuar con una cola de poli(A) de una diana). Las condiciones de ensayo usadas para la hibridación (por ejemplo, pH del tampón, fuerza iónica, temperatura, etc.) pueden elegirse para que promuevan la formación de híbridos específicos y estables. En algunas realizaciones, las moléculas de ácido nucleico liberadas de las células lisadas pueden asociarse con la pluralidad de sondas del sustrato (por ejemplo, hibridar con las sondas del sustrato). Cuando las sondas comprenden oligo(dT), las moléculas de ARNm pueden hibridar con las sondas y transcribirse inversamente. La porción de oligo(dT) del oligonucleótido puede actuar como cebador para la síntesis de la primera

cadena de la molécula de ADNc. Por ejemplo, en un ejemplo no limitativo de codificación con códigos de barras ilustrado en la FIG. 2, en el bloque 216, las moléculas de ARNm pueden hibridar con códigos de barras en perlas. Por ejemplo, los fragmentos de nucleótidos de cadena sencilla pueden hibridar con las regiones de unión a la diana de los códigos de barras.

La unión puede comprender además la ligación de una región de reconocimiento de la diana de un código de barras y una porción de la molécula de ácido nucleico diana. Por ejemplo, la región de unión a la diana puede comprender una secuencia de ácido nucleico que puede ser capaz de hibridación específica con un saliente de sitio de restricción (por ejemplo, un saliente de extremo pegajoso 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 de sitio de restricción. A continuación, el código de barras puede ligarse a cualquier molécula de ácido nucleico que comprenda una secuencia complementaria al saliente del sitio de restricción. Para unir los dos fragmentos puede usarse una ligasa (por ejemplo, T4 ADN ligasa).

Por ejemplo, en un ejemplo no limitativo de codificación con códigos de barras 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 diana-código de barras) pueden agruparse posteriormente, por ejemplo, en un tubo. Las dianas marcadas pueden agruparse, por ejemplo, recuperando los códigos de barras y/o las perlas a las que están unidas las moléculas de diana-código de barras.

La recuperación de colecciones basadas en soportes sólidos de moléculas de diana-código de barra unidas puede implementarse mediante el uso de perlas magnéticas y un campo magnético aplicado externamente. Una vez que se han agrupado las moléculas de diana-código de barras, todo el procesamiento posterior puede realizarse en un único recipiente de reacción. El procesamiento posterior 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. Las reacciones de procesamiento posterior pueden realizarse dentro de los micropocillos, es decir, sin agrupar primero las moléculas de ácido nucleico diana marcadas de una pluralidad de células.

Transcripción inversa

La divulgación proporciona un método para crear un conjugado diana-código de barras usando transcripción inversa (por ejemplo, en el bloque 224 de la FIG. 2). El conjugado diana-código de barras puede comprender el código de barras y una secuencia complementaria de todo o una porción del ácido nucleico diana (es decir, una molécula de ADNc codificada con código de barras, como una molécula de ADNc codificada estocásticamente con código de barras). La transcripción inversa de la molécula de ARN asociada puede producirse 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 de hexanucleótidos aleatorio o un cebador oligonucleótido específico de la diana. Los cebadores oligo(dT) pueden tener, o tener aproximadamente, 12-18 nucleótidos de longitud y se unen a la cola de poli(A) endógena 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 de oligonucleótidos específicos de la diana típicamente ceban selectivamente el ARNm de interés.

En algunas realizaciones, la transcripción inversa de la molécula de ARN marcada puede producirse mediante la adición de un cebador de transcripción inversa. En algunas realizaciones, el cebador de transcripción inversa es un cebador oligo(dT), un cebador de hexanucleótidos aleatorio o un cebador de oligonucleótidos específico de la diana. Generalmente, los cebadores oligo(dT) tienen 12-18 nucleótidos de longitud y se unen a la cola de poli(A) endógena 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 de oligonucleótidos específicos de la diana típicamente ceban selectivamente el ARNm de interés.

La transcripción inversa puede producirse repetidamente para producir múltiples moléculas de ADNc marcadas. Los métodos divulgados en la presente pueden comprender realizar por lo 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 por lo menos aproximadamente 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95 o 100 reacciones de transcripción inversa.

Amplificación

Pueden realizarse una o más reacciones de amplificación de ácidos nucleicos (por ejemplo, en el bloque 228 de la FIG. 2) para crear múltiples copias de las moléculas de ácido nucleico diana marcadas. La amplificación puede realizarse de manera multiplexada, en donde se amplifican simultáneamente múltiples secuencias de ácido nucleico diana. La reacción de amplificación puede usarse para añadir adaptadores de secuenciación a las moléculas de ácido nucleico. Las reacciones de amplificación pueden comprender amplificar por lo menos una parte de un marcador de muestra, si lo hay. Las reacciones de amplificación pueden comprender amplificar por lo menos una parte del marcador celular y/o de la secuencia del código de barras (por ejemplo, marcador molecular). Las reacciones de amplificación

pueden comprender amplificar por lo menos una porción de una etiqueta de muestra, un marcador celular, un marcador espacial, una secuencia de código de barras (por ejemplo, un marcador molecular), un ácido nucleico diana, o una combinación de los mismos. Las reacciones de amplificación pueden comprender amplificar el 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 intervalo 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 diana-código de barras que comprenden un marcador de muestra, un marcador celular, un marcador espacial y/o una secuencia de código de barras (por ejemplo, un marcador molecular).

En algunas realizaciones, la amplificación puede realizarse usando una reacción en cadena de la polimerasa (PCR). Como se usa en la presente, PCR puede referirse a una reacción para la amplificación in vitro de secuencias específicas de ADN mediante la extensión simultánea de cebadores de cadenas complementarias de ADN. Como se usa en la presente, la PCR puede abarcar formas derivadas de la reacción, incluyendo pero no limitadas a, RT-PCR, PCR en tiempo real, PCR anidada, PCR cuantitativa, PCR multiplexada, PCR digital y PCR de ensamblaje.

La amplificación de los ácidos nucleicos marcados puede comprender métodos no basados en PCR. Los 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 cadena (SDA), SDA en tiempo real, amplificación por círculo rodante o amplificación de círculo a círculo. Otros métodos de amplificación no basados en PCR incluyen ciclos múltiples de amplificación de transcripción de ARN dependiente de ADN impulsada por ARN polimerasa o síntesis y transcripción de ADN dirigida por ARN para amplificar dianas de ADN o ARN, una reacción en cadena de ligasa (LCR), y un método de Qβ replicasa (Qβ), uso de sondas palindrómicas, amplificación por desplazamiento de cadena, amplificación impulsada por oligonucleótidos usando una endonucleasa de restricción, un método de amplificación en el que un cebador 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 cadena usando una polimerasa de ácido nucleico que carece de actividad de exonucleasa 5′, amplificación por círculo rodante y amplificación por extensión de ramificaciones (RAM). En algunas realizaciones, la amplificación no produce transcritos circularizados.

En algunas realizaciones, los métodos divulgados en la presente comprenden además realizar una reacción en cadena de la polimerasa en el ácido nucleico marcado (por ejemplo, ARN marcado, ADN marcado, ADNc marcado) para producir un amplicón marcado (por ejemplo, un amplicón marcado estocásticamente). El amplicón marcado puede ser una molécula de cadena doble. La molécula de cadena doble puede comprender una molécula de cadena doble de ARN, una molécula de ADN de cadena doble o una molécula de ARN hibridada con una molécula de ADN. Una o ambas cadenas de la molécula de cadena doble pueden comprender un marcador de muestra, un marcador espacial, un marcador celular y/o una secuencia de código de barras (por ejemplo, un marcador molecular). El amplicón marcado puede ser una molécula de cadena sencilla. La molécula de cadena sencilla puede comprender ADN, ARN o una combinación de los mismos. Los ácidos nucleicos de la divulgación pueden comprender ácidos nucleicos sintéticos o alterados.

La amplificación puede comprender el uso de uno o más nucleótidos no naturales. Los nucleótidos no naturales pueden incluir nucleótidos fotolábiles o activables. Los ejemplos de nucleótidos no naturales pueden incluir, pero no se limitan a, ácido nucleico peptídico (PNA), morfolino y ácido nucleico bloqueado (LNA), así como ácido nucleico de glicol (GNA) y ácido nucleico de treosa (TNA). Los nucleótidos no naturales pueden añadirse a uno o más ciclos de una reacción de amplificación. La adición de los nucleótidos no naturales puede usarse para identificar productos como ciclos específicos o puntos temporales en la reacción de amplificación.

La realización de una o más reacciones de amplificación puede comprender el uso de uno o más cebadores. El 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. El uno o más cebadores pueden comprender por lo menos 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, o 15 o más nucleótidos. El uno o más cebadores pueden comprender menos de 12-15 nucleótidos. El uno o más cebadores pueden aparearse con por lo menos una parte de la pluralidad de dianas marcadas (por ejemplo, dianas marcadas estocásticamente). El uno o más cebadores pueden aparearse con el extremo 3′ o el extremo 5′ de la pluralidad de dianas marcadas. El uno o más cebadores pueden aparearse con una región interna de la pluralidad de dianas marcadas. La región interna puede tener por lo 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, 750, 800, 850, 900 o 1000 nucleótidos desde los extremos 3′ de la pluralidad de dianas marcadas. El uno o más cebadores puede comprender un panel fijo de cebadores. El uno o más cebadores pueden comprender por lo menos uno o más cebadores personalizados. El uno o más cebadores pueden comprender por lo menos uno o más cebadores de control. El uno o más cebadores pueden comprender por lo menos uno o más cebadores específicos de gen.

El uno o más cebadores puede incluir un cebador universal. El cebador universal puede aparearse con un sitio de unión del cebador universal. El uno o más cebadores personalizados pueden aparearse con un primer marcador de muestra, un segundo marcador de muestra, un marcador espacial, un marcador celular, una secuencia

de código de barras (por ejemplo, un marcador 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 estar diseñado para amplificar una o más dianas. Las dianas pueden comprender un subconjunto de los ácidos nucleicos totales en una o más muestras. Las dianas pueden comprender un subconjunto del total de dianas marcadas en una o más muestras. El uno o más cebadores pueden comprender por lo menos 96 o más cebadores personalizados. El uno o más cebadores pueden comprender por lo menos 960 o más cebadores personalizados. El uno o más cebadores pueden comprender por lo menos 9600 o más cebadores personalizados. El uno o más cebadores personalizados pueden aparearse con dos o más ácidos nucleicos marcados diferentes. Los dos o más ácidos nucleicos marcados diferentes pueden corresponder a uno o más genes.

En los métodos de la presente divulgación puede usarse cualquier esquema de amplificación. 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 usando 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 añade P5 y P7 e índice de muestra para convertir los productos de PCR en una biblioteca de secuenciación de Illumina. La secuenciación usando secuenciación de 150 bp x 2 puede revelar el marcador celular y la secuencia del código de barras (por ejemplo, el marcador molecular) en la lectura 1, el gen en la lectura 2 y el índice de la muestra en la lectura del índice 1.

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

Cuando las sondas son específicas de un gen, las moléculas pueden hibridar con las sondas y transcribirse inversamente y/o amplificarse. En algunas realizaciones, después de que se haya sintetizado el ácido nucleico (por ejemplo, transcribiéndolo inversamente), puede amplificarse. La amplificación puede realizarse de forma multiplex, en donde se amplifican simultáneamente múltiples secuencias de ácidos nucleicos diana. La amplificación puede añadir adaptadores de secuenciación al ácido nucleico.

En algunas realizaciones, la amplificación puede llevarse a cabo en el sustrato, por ejemplo, con amplificación en puente. Los ADNc pueden tener cola de homopolímero para generar un extremo compatible para la amplificación en puente usando sondas oligo(dT) en el sustrato. En la amplificación en puente, el cebador que es complementario al extremo 3′ del ácido nucleico plantilla 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 puede aparearse con el primer cebador y el primer cebador se alarga en la dirección directa mediante la adición de nucleótidos para formar una molécula dúplex que consiste en la molécula plantilla y una cadena de ADN recién formada que es complementaria a la plantilla. En el paso de calentamiento del siguiente ciclo, puede desnaturalizarse la molécula dúplex, liberando la molécula plantilla de la partícula y dejando la cadena de ADN complementaria unida a la partícula a través del primer cebador. En la etapa de apareamiento del paso de apareamiento y elongación que sigue, la cadena complementaria puede hibridar con el segundo cebador, que es complementario a un segmento de la cadena complementaria en una localización alejada del primer cebador. Esta hibridación puede hacer que la cadena complementaria forme un puente entre el primer y el segundo cebadores fijadas al primer cebador mediante un enlace covalente y al segundo cebador por hibridación. En la etapa de elongación, el segundo cebador puede alargarse en la dirección inversa mediante la adición de nucleótidos en la misma mezcla de reacción, convirtiendo de este modo el puente en un puente de cadena doble. A continuación comienza el siguiente ciclo, y el puente de cadena doble puede desnaturalizarse para proporcionar dos moléculas de ácido nucleico de cadena sencilla, cada una de las cuales tiene un extremo unido a la superficie de la partícula a través del primer y el segundo cebadores, respectivamente, con el otro extremo de cada uno no unido. En el paso de apareamiento y elongación de este segundo ciclo, cada cadena puede hibridar con un cebador complementario adicional, no usado previamente, en la misma partícula, para formar nuevos puentes de cadena sencilla. Los dos cebadores no usados anteriormente que están ahora hibridados se alargan para convertir los dos nuevos puentes en puentes de cadena doble.

Las reacciones de amplificación pueden comprender amplificar por lo menos el 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.

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 la amplificación exponencial de los ácidos nucleicos marcados. La amplificación de los ácidos nucleicos marcados puede comprender la amplificación lineal de los ácidos nucleicos marcados. La amplificación puede realizarse mediante reacción en cadena de la polimerasa (PCR). La PCR puede referirse a una reacción para la amplificación in vitro de secuencias específicas de ADN mediante la extensión simultánea de cebadores de cadenas complementarias de ADN. La PCR puede abarcar formas derivadas de la reacción, que incluyen pero no se limitan a, RT-PCR, PCR en tiempo real, PCR anidada, PCR cuantitativa, PCR multiplexada, PCR digital, PCR de supresión, PCR semisupresiva y PCR de ensamblaje.

En algunas realizaciones, la amplificación de los ácidos nucleicos marcados comprende métodos no basados en PCR. Ejemplos de métodos no basados en PCR incluyen, pero no se limitan a, amplificación por desplazamiento múltiple (MDA), amplificación mediada por transcripción (TMA), amplificación basada en secuencia de ácido nucleico (NASBA), amplificación por desplazamiento de cadena (SDA), SDA en tiempo real, amplificación por círculo rodante o amplificación de círculo a círculo. Otros métodos de amplificación no basados en PCR incluyen ciclos múltiples de amplificación de transcripción de ARN dependiente de ADN impulsada por ARN polimerasa o síntesis y transcripción de ADN dirigida por ARN para amplificar dianas de ADN o ARN, una reacción en cadena de ligasa (LCR), una Qβ replicasa (Qβ), uso de sondas palindrómicas, amplificación por desplazamiento de cadena, amplificación impulsada por oligonucleótidos usando una endonucleasa de restricción, un método de amplificación en el que un cebador 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 cadena usando una polimerasa de ácido nucleico que carece de actividad de exonucleasa 5′, amplificación por círculo rodante y/o amplificación por extensión ramificada (RAM).

En algunas realizaciones, los métodos divulgados en la presente comprenden además realizar una reacción en cadena de la polimerasa anidada en el amplicón amplificado (por ejemplo, diana). El amplicón puede ser una molécula de cadena doble. La molécula de cadena doble puede comprender una molécula de ARN de cadena doble, una molécula de ADN de cadena doble o una molécula de ARN hibridada con una molécula de ADN. Una o ambas cadenas de la molécula de cadena doble pueden comprender una etiqueta de muestra o un marcador de identificador molecular. Alternativamente, el amplicón puede ser una molécula de cadena sencilla. La molécula de cadena sencilla puede comprender ADN, ARN o una combinación de los mismos. Los ácidos nucleicos pueden comprender ácidos nucleicos sintéticos o alterados.

En algunas realizaciones, el método comprende amplificar repetidamente el ácido nucleico marcado para producir múltiples amplicones. Los métodos divulgados en la presente pueden comprender realizar por lo 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 por lo menos aproximadamente 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95 o 100 reacciones de amplificación.

La amplificación puede comprender además añadir uno o más ácidos nucleicos de control a una o más muestras que comprenden una pluralidad de ácidos nucleicos. La amplificación puede comprender además añadir uno o más ácidos nucleicos de control a una pluralidad de ácidos nucleicos. Los ácidos nucleicos de control pueden comprender un marcador de control.

La amplificación puede comprender el uso de uno o más nucleótidos no naturales. Los nucleótidos no naturales pueden incluir nucleótidos fotolábiles y/o activables. Ejemplos de nucleótidos no naturales incluyen, pero no se limitan a, ácido nucleico peptídico (PNA), morfolino y ácido nucleico bloqueado (LNA), así como ácido nucleico de glicol (GNA) y ácido nucleico de treosa (TNA). Los nucleótidos no naturales pueden añadirse a uno o más ciclos de una reacción de amplificación. La adición de los nucleótidos no naturales puede usarse para identificar productos como ciclos específicos o puntos temporales en la reacción de amplificación.

La 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 por lo menos aproximadamente 7-9 nucleótidos. E uno o más oligonucleótidos pueden comprender menos de 12-15 nucleótidos. El uno o más cebadores pueden aparearse con por lo menos una parte de la pluralidad de ácidos nucleicos marcados. El uno o más cebadores pueden aparearse con el extremo 3′ y/o el extremo 5′ de la pluralidad de ácidos nucleicos marcados. El uno o más cebadores pueden aparearse con una región interna de la pluralidad de ácidos nucleicos marcados. La región interna puede tener por lo 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, 750, 800, 850, 900 o 1000 nucleótidos de los extremos 3′ de la pluralidad de ácidos nucleicos marcados. El uno o más cebadores pueden comprender un panel fijo de cebadores. El uno o más cebadores pueden comprender por lo menos uno o más cebadores personalizados. El uno o más cebadores pueden comprender por lo menos uno o más cebadores de control. El uno o más cebadores puede comprender por lo menos uno o más cebadores de genes constitutivos. El uno o más

cebadores puede comprender un cebador universal. El cebador universal puede aparearse a un sitio de unión del cebador universal. El uno o más cebadores personalizados pueden aparearse a la primera etiqueta de muestra, la segunda etiqueta de muestra, al marcador identificador molecular, al ácido nucleico o a un producto de los mismos. El uno o más cebadores pueden comprender un cebador universal y un cebador personalizado. El cebador personalizado puede estar diseñado para que amplifique 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 realizaciones, los cebadores son las sondas unidas a la matriz de la divulgación.

En algunas realizaciones, la codificación con códigos de barras (por ejemplo, codificación estocástica con códigos de barras) de la pluralidad de dianas de la muestra comprende además generar una biblioteca indexada de fragmentos codificados con código de barras. Las secuencias de los códigos de barras de diferentes códigos de barras (por ejemplo, los marcadores moleculares de diferentes códigos de barras estocásticos) pueden ser diferentes entre sí. La generación de una biblioteca indexada de dianas codificadas con códigos de barras (por ejemplo, dianas codificadas estocásticamente con códigos de barras) incluye la generación de una pluralidad de polinucleótidos indexados a partir de la pluralidad de dianas de la muestra. Por ejemplo, para una biblioteca indexada de dianas codificadas con códigos de barras que comprende una primera diana indexada y una segunda diana indexada, la región del marcador del primer polinucleótido indexado puede diferir de la región del marcador del segundo polinucleótido indexado en, en aproximadamente, por lo menos, o como máximo, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50, o un número o un intervalo entre dos cualesquiera de estos valores, nucleótidos. En algunas realizaciones, la generación de una biblioteca indexada de las dianas codificadas con códigos de barras 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 de marcador; y realizar una primera síntesis de cadena usando una transcriptasa inversa para producir moléculas de ADNc marcadas de cadena sencilla que comprenden cada una una región de ADNc y una región de marcador, en donde la pluralidad de dianas incluye por lo menos dos moléculas de ARNm de secuencias diferentes y la pluralidad de oligonucleótidos incluye por lo menos dos oligonucleótidos de secuencias diferentes. Generar una biblioteca indexada de las dianas codificadas con códigos de barras puede comprender además amplificar las moléculas de ADNc marcadas de cadena sencilla para producir moléculas de ADNc marcadas de cadena doble; y la realización de PCR anidada en las moléculas de ADNc marcadas de cadena doble para producir amplicones marcados. En algunas realizaciones, el método puede incluir generar un amplicón marcado con adaptador.

La codificación con código de barras puede usar códigos de barras o etiquetas de ácidos nucleicos para marcar moléculas individuales de ácido nucleico (por ejemplo, ADN o ARN). En algunas realizaciones, implica añadir códigos de barras o etiquetas de ADN a moléculas de ADNc a medida que se generan a partir de ARNm. Puede realizarse una PCR anidada para minimizar el sesgo de amplificación de la PCR. Pueden añadirse adaptadores para la secuenciación usando, por ejemplo, secuenciación de próxima generación (NGS). Los resultados de la secuenciación pueden usarse para determinar marcadores celulares, secuencias de códigos de barras (por ejemplo, marcadores 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.

La FIG. 3 es una ilustración esquemática que muestra un proceso ejemplar no limitativo para generar una biblioteca indexada de dianas codificadas con códigos barras (por ejemplo, dianas codificadas estocásticamente con códigos barras), por ejemplo ARNm. Como se muestra en el paso 1, el proceso de transcripción inversa puede codificar cada molécula de ARNm con una secuencia de código de barras (marcador molecular) única, un marcador celular y un sitio de PCR universal. Por ejemplo, las moléculas de ARN 302 pueden transcribirse inversamente para producir moléculas de ADNc marcadas 304, incluyendo una región de ADNc 306, mediante la hibridación (por ejemplo, hibridación estocástica) de un conjunto de códigos de barras (por ejemplo, códigos de barras estocásticos) 310) a la región de cola poli(A) 308 de las moléculas de ARN 302. Cada uno de los códigos de barras 310 puede comprender una región de unión a la diana, por ejemplo una región poli(dT) 312, una secuencia de código de barras o un mar4cador molecular 314 y una región de PCR universal 316.

En algunas realizaciones, el marcador celular puede incluir de 3 a 20 nucleótidos. En algunas realizaciones, la secuencia del código de barras (por ejemplo, marcador molecular) puede incluir de 3 a 20 nucleótidos. En algunas realizaciones, cada uno de la pluralidad de los códigos de barras estocásticos comprende además uno o más de un marcador universal y un marcador celular, en donde los marcadores universales son los mismos para la pluralidad de códigos de barras estocásticos en el soporte sólido y los marcadores celulares son los mismos para la pluralidad de códigos de barras estocásticos en el soporte sólido. En algunas realizaciones, el marcador universal puede incluir de 3 a 20 nucleótidos. En algunas realizaciones, el marcador celular comprende de 3 a 20 nucleótidos.

En algunas realizaciones, la región de marcador 314 puede incluir una secuencia de código de barras o un marcador molecular 318 y un marcador celular 320. En algunas realizaciones, la región de marcador 314 puede incluir una o más de un marcador universal, un marcador de dimensión y un marcador celular. La secuencia de código de barras o marcador molecular 318 puede tener, puede tener aproximadamente, puede tener por lo menos, o puede tener 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 intervalo entre cualquiera de estos valores, nucleótidos de longitud. El marcador celular 320 puede tener, puede tener aproximadamente, puede tener por lo menos, o puede tener 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 intervalo entre cualquiera de estos valores, nucleótidos de longitud. El marcador universal puede tener, puede tener aproximadamente, puede tener por lo menos, o puede tener 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 intervalo entre cualquiera de estos valores, nucleótidos de longitud. Los marcadores universales pueden ser los mismos para la pluralidad de códigos de barras estocásticos en el soporte sólido y los marcadores celulares son los mismos para la pluralidad de códigos de barras estocásticos en el soporte sólido. El marcador de dimensión puede tener, puede tener aproximadamente, puede tener por lo menos, o puede tener 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 intervalo entre cualquiera de estos valores, nucleótidos de longitud.

En algunas realizaciones, la región de marcador 314 puede comprender, comprender aproximadamente, comprender por lo 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 un intervalo entre cualquiera de estos valores, marcadores diferentes, como una secuencia de código de barras o un marcador molecular 318 y un marcador celular 320. Cada marcador puede tener, tener aproximadamente, puede tener por lo menos, o puede tener como máximo 1, 2, 3, 4, 6, 7, 8, 9, 10, 20, 30, 40, 50, 60, 70, 90, 100, o un número o un intervalo entre cualquiera de estos valores, nucleótidos de longitud. Un conjunto de códigos de barras o códigos de barras estocásticos 310 puede contener, contener aproximadamente, contener por lo menos, o puede tener como máximo, 10, 20, 40, 50, 70, 80, 90, 10^2, 10^3, 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 intervalo entre cualquiera de estos valores, códigos de barras o códigos de barras estocásticos 310. Y el conjunto de códigos de barras o códigos de barras estocásticos 310 puede, por ejemplo, contener cada uno una región de marcador único 314. Las moléculas de ADNc marcadas 304 pueden purificarse para eliminar el exceso de códigos de barras o códigos de barras estocásticos 310. La purificación puede comprender la purificación con perlas Ampure.

Como se muestra en el paso 2, los productos del proceso de transcripción inversa del paso 1 pueden agruparse en 1 tubo y amplificarse por PCR con un 1º grupo de cebadores PCR y un 1º cebador de PCR universal. La agrupación es posible gracias a la región de marcador única 314. En particular, las moléculas de ADNc marcadas 304 pueden amplificarse para producir amplicones marcados de PCR anidada 322. La amplificación puede comprender una amplificación por PCR multiplex. La amplificación puede comprender una amplificación por PCR multiplex con 96 cebadores multiplex en un único volumen de reacción. En algunas realizaciones, la amplificación por PCR multiplex puede utilizar, utilizar aproximadamente, utilizar por lo 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 intervalo entre cualquiera de estos valores, cebadores multiplex en un único volumen de reacción. La amplificación puede comprender un 1º grupo de cebadores de PCR 324 de cebadores personalizados 326A-C dirigidos a genes específicos y un cebador universal 328. Los cebadores personalizados 326 pueden hibridar 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.

Como se muestra en el paso 3 de la FIG. 3, los productos de la amplificación por PCR del paso 2 pueden amplificarse con un grupo de cebadores de PCR anidada y un 2º cebador de PCR universal. La PCR anidada puede minimizar el sesgo de amplificación por PCR. Por ejemplo, los amplicones marcados de PCR anidada 322 pueden amplificarse adicionalmente por PCR anidada. La PCR anidada puede comprender PCR multiplex con el grupo de cebadores de PCR anidada 330 de cebadores de PCR anidada 332 a-c y un 2º cebador de PCR universal 328′ en un único volumen de reacción. El grupo de cebadores de PCR anidada 328 puede contener, contener aproximadamente, contener por lo 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 un intervalo entre cualquiera de estos valores, cebadores de PCR anidada 330 diferentes. Los cebadores de PCR anidada 332 pueden contener un adaptador 334 e hibridar con una región dentro de la porción de ADNc 306" del amplicón marcado 322. El cebador universal 328′ puede contener un adaptador 336 e hibridar con la región de PCR universal 316 del amplicón marcado 322. Por tanto, el paso 3 produce el amplicón marcado con adaptador 338. En algunas realizaciones, los cebadores de PCR anidada 332 y el 2º cebador de PCR universal 328′ pueden no contener los adaptadores 334 y 336. En su lugar, los adaptadores 334 y 336 pueden ligarse a los productos de la PCR anidada para producir el amplicón marcado con adaptador 338.

Como se muestra en el paso 4, los productos de PCR del paso 3 pueden amplificarse por PCR para secuenciación usando cebadores de amplificación de bibliotecas. En particular, los adaptadores 334 y 336 pueden usarse para realizar uno o más ensayos adicionales en el amplicón marcado con adaptador 338. Los adaptadores 334 y 336 pueden hibridar con los cebadores 340 y 342. El uno o más cebadores 340 y 342 pueden ser cebadores de amplificación de PCR. El uno o más cebadores 340 y 342 pueden ser cebadores de secuenciación. El uno o más adaptadores 334 y 336 pueden usarse para la amplificación adicional de los amplicones marcados con adaptador 338. El uno o más adaptadores 334 y 336 pueden usarse para secuenciar el amplicón marcado con adaptador 338. El cebador 342 puede contener un índice de placa 344 de tal manera que los amplicones generados usando el mismo conjunto de códigos de barras o códigos de barras estocásticos 310 puedan secuenciarse en una reacción de secuenciación usando secuenciación de próxima generación (NGS).

Agrupación de perfiles de expresión usando un dendrograma

En la presente se divulgan métodos de identificación de dianas para distinguir tipos de células. En algunas realizaciones, el método comprende: (a) recibir una estructura de datos de recuentos de dianas, en donde la estructura de datos de recuentos de dianas comprende perfiles de expresión de una pluralidad de células, y en donde los perfiles de expresión de la pluralidad de células comprenden un número de cada diana de una pluralidad de dianas para cada célula de la pluralidad de células; (b) agrupar jerárquicamente los perfiles de expresión de la pluralidad de células basándose en la estructura de datos de recuentos de dianas y las distancias entre los perfiles de expresión de la pluralidad de células para generar un dendrograma que represente los perfiles de expresión de la pluralidad de células, en donde el dendrograma comprende una pluralidad de nodos, en donde la pluralidad de nodos comprende un nodo raíz, una pluralidad de nodos hoja, y una pluralidad de nodos no raíz y no hoja, en donde cada nodo hoja de la pluralidad de nodos hoja representa un perfil de expresión de una célula diferente de la pluralidad de células, y den onde el nodo raíz representa perfiles de expresión de la pluralidad de células; (c) mientras se atraviesa a través cada nodo de la pluralidad de nodos del dendrograma desde el nodo raíz del dendrograma hasta la pluralidad de nodos hoja del dendrograma: (1) determinar si una división del nodo en nodos hijos del nodo es válida o inválida (por ejemplo, las diferencias entre los nodos hijos no son significativas); y (2) si la división del nodo en los nodos hijos del nodo es inválida, añadir el nodo a un conjunto de grupos de fusión; (d) iterativamente, para cada primer nodo en el conjunto de grupos de fusión, si una distancia entre el primer nodo en el conjunto de grupos de fusión y un segundo nodo en el conjunto de grupos de fusión que está más cerca del primer nodo está dentro de un umbral de distancia de fusión, fusionar el primer nodo con el segundo nodo para generar un nodo fusionado que comprenda perfiles de expresión representados por el primer nodo y el segundo nodo; y (e) para cada nodo en el conjunto de grupo de fusión, identificar dianas para distinguir tipos de células basándose en los perfiles de expresión de la pluralidad de dianas de las células representadas por el nodo.

La FIG. 4 es un diagrama de flujo que muestra un método ejemplar no limitativo 400 de identificación de dianas para distinguir tipos de células. El método 400 mapea una estructura de datos de recuentos moleculares (por ejemplo, una matriz de recuentos moleculares) para un conjunto de marcadores de grupos y un conjunto de genes importantes. En algunas realizaciones, la entrada puede ser una matriz de N por M de recuentos moleculares o una matriz donde la entrada *i,j*-ésima describe el número de moléculas para el gen *j* que se observaron usando las lecturas de la célula i. El algoritmo puede generar dos salidas. La primera puede ser un conjunto de N marcadores, una para cada célula (las células del mismo marcador pueden inferirse como "similares"). La segunda salida puede ser un conjunto de genes que pueden usarse para distinguir los grupos entre sí.

El método 400 genera las dos salidas usando un enfoque de división, prueba y fusión de dendrogramas. Después de preprocesar los datos y generar una estructura de datos de distancias (por ejemplo, una matriz de distancia) *D*, el algoritmo puede agrupar jerárquicamente *D* para producir un dendrograma. El algoritmo puede incluir dos fases. Durante la fase de división y prueba, el método 400 comienza desde la parte superior del dendrograma (por ejemplo, el nodo raíz 504 del árbol o dendrograma 500 de la FIG. 5). El dendrograma 500 incluye el nodo raíz 504, una pluralidad de nodos hoja 520a-520h, y una pluralidad de nodos no raíz y no hoja 508a-508b, 512a-512b, y 516a-516b. En cada nodo del dendrograma (excepto para los nodos hoja, como los nodos hoja 520a-520h), el árbol se divide en dos subárboles (por ejemplo, el nodo raíz 504 se divide en dos subárboles 508a, 508b). La división corresponde a un grupo (por ejemplo, que contiene perfiles de expresión de dos o más células) que se divide en dos subgrupos candidatos (por ejemplo, cada uno de los cuales contiene el perfil de expresión de por lo menos una célula). Puede puntuarse la calidad de la división. Si se considera que los subgrupos son suficientemente diferentes, el algoritmo continúa ejecutándose en cada subárbol. En caso contrario, el algoritmo termina para esta parte del dendrograma. Esta fase produce un conjunto de marcadores para el conjunto de datos. Durante la fase de fusión, el método 400 usa los marcadores generados durante la fase de división y prueba para determinar si debe combinarse alguno de estos grupos para formar un grupo. En algunas realizaciones, la fase de división y prueba tiende a producir grupos pequeños de unas pocas muestras cada uno. La fase de fusión puede "limpiar" los grupos más pequeños fusionándolos con grupos más grandes.

En el bloque 404, el método 400 puede incluir recibir una estructura de datos de recuentos moleculares (por ejemplo, una matriz de recuentos moleculares). La matriz puede comprender sólo entradas enteras no negativas y tiende a ser grande y dispersa. En algunas realizaciones, la entrada puede ser una matriz de N por M de recuentos moleculares o una matriz donde la entrada *i,j*-ésima describe el número de moléculas para el gen *j* que se observaron usando las lecturas de la célula *i*.

En el bloque 408, el método 400 puede incluir el preprocesamiento de la estructura de datos de recuentos moleculares para generar una estructura de datos de distancias (por ejemplo, una matriz de distancias). En algunas realizaciones, la estructura de datos de entrada se transforma logarítmicamente. Se añade el valor 1 a cada entrada antes de tomar el logaritmo natural. La distancia de correlación puede usarse para describir la disimilitud por pares *d* entre pares de células. Para las células $\boldsymbol{c}_i$ y $\boldsymbol{c}_j$, la distancia de correlación entre las dos células puede determinarse usando la Ec. [1].

$$d(\boldsymbol{c}_i, \boldsymbol{c}_j) = 1 - \frac{(\boldsymbol{c}_i - \bar{\boldsymbol{c}}_i)\cdot(\boldsymbol{c}_j - \bar{\boldsymbol{c}}_j)}{\|\boldsymbol{c}_i - \bar{\boldsymbol{c}}_i\|_2 \|\boldsymbol{c}_j - \bar{\boldsymbol{c}}_j\|_2}, \qquad \text{Ec. [1]}$$

donde $\underline{\boldsymbol{c}}_i$ denota la media de todos los elementos de $\boldsymbol{c}_i$. La salida del paso de preprocesamiento puede ser una matriz cuadrada y simétrica D de distancias con 0's a lo largo de la diagonal.

En el bloque 412, el método 400 puede incluir la agrupación jerárquica de los perfiles de expresión de las células para generar un dendrograma. Agrupar jerárquicamente los perfiles de expresión de las células para generar el dendrograma puede comprender fusionar iterativamente los dos grupos más cercanos del dendrograma. Todos los grupos pueden iniciarse como puntos individuales con distancias entre pares descritas anteriormente. El cálculo de la distancia D entre los grupos se realizó usando la vinculación completa. Para los grupos A y B, la distancia entre los dos grupos puede determinarse usando la Ec. [2]:

$$D(A, B) = \max_{\boldsymbol{a}\in A, \boldsymbol{b}\in B} d(\boldsymbol{a}, \boldsymbol{b}). \qquad \text{Ec. [2]}$$

En este bloque puede obtenerse un dendrograma completo. En algunas realizaciones, una correlación intragrupo del grupo A y una correlación intragrupo del grupo B son más altas que una correlación intergrupo del grupo A y del grupo B. Una medida o una indicación de una correlación intragrupo del grupo A y una correlación intragrupo del grupo B es más alta que una correlación intergrupo del grupo A y del grupo B. La medida de la correlación intragrupo del grupo A y la correlación intragrupo del grupo B puede basarse en por lo menos una de las siguientes: una correlación máxima intragrupo del grupo A y el grupo B, una correlación media intragrupo del grupo A y el grupo B, una correlación mediana intragrupo del grupo A y el grupo B, una correlación mínima intragrupo del grupo A y el grupo B, y cualquier combinación de las mismas. La correlación intragrupo del grupo A puede basarse en por lo menos una de: una correlación máxima intragrupo del grupo A, una correlación media intragrupo del grupo A, una correlación mediana intragrupo del grupo A, una correlación mínima intragrupo del grupo A, y cualquier combinación de las mismas. La correlación intragrupo del grupo B puede basarse en por lo menos una de: una correlación máxima intragrupo del grupo B, una correlación media intragrupo del grupo B, una correlación mediana intragrupo del grupo B, una correlación mínima intragrupo del grupo B, y cualquier combinación de las mismas. La correlación intergrupo del grupo A y el grupo B puede basarse en por lo menos una de: una correlación máxima intergrupo del grupo A y el grupo B, una correlación media intergrupo del grupo A y el grupo B, una correlación mediana intergrupo del grupo A y el grupo B, una correlación mínima intergrupo del grupo A y el grupo B, y cualquier combinación de las mismas. Por ejemplo, la correlación mediana intragrupo de los dos subgrupos puede ser mayor que la correlación mediana intergrupo.

En el bloque 416, el método 400 puede incluir dividir y comprobar el dendrograma para generar un conjunto de marcadores. La división y comprobación pueden iniciarse en la parte superior del dendrograma. Dado un subárbol de dendrograma T_0, el árbol puede dividirse exactamente en dos subárboles TL y TR. Puede realizarse una prueba estadística para determinar si las células del subárbol izquierdo TL son suficientemente diferentes de las células del subárbol derecho TR. En algunas realizaciones, la prueba estadística implica realizar una prueba t de Welch en cada gen para las dos poblaciones. Pueden producirse estadísticas t de infinito si se estima que la varianza es 0 en ambas poblaciones; estos casos pueden ignorarse. Si el valor p mínimo entre todas las pruebas es menor que un cierto umbral (corregido de manera conservadora para tener en cuenta la tasa de falsas detecciones), entonces la división puede considerarse válida y el algoritmo se ejecuta de nuevo en los dos subárboles. Si el valor p mínimo no está por debajo del umbral, el método 400 termina para el subárbol T0. Si TL contiene exactamente 1 muestra (es decir, TL es un singleton), TL puede ignorarse y el algoritmo repite el procedimiento con TR. Si TR contiene exactamente 1 muestra, el TR puede ignorarse y el algoritmo repite el procedimiento con TL. Si tanto TL como TR contienen exactamente 1 muestra cada uno, el algoritmo termina para el subárbol T0.

En el bloque 416, el método 400 puede incluir determinar los marcadores de grupos de la siguiente manera. Al principio, todos los subárboles pueden marcarse con 'r'. Cada vez que se produce una división y no se rechaza debido a problemas de valor p, todos los marcadores de las células en TL se añaden con "L" y todos los marcadores de las células en TR pueden añadirse con "R". Esto significa que, al pasar por encima de los singletones, los marcadores siguen viéndose afectados. El singleton obtiene automáticamente un marcador único no compartido con ningún otro punto de datos.

En el bloque 416, el método 400 puede incluir determinar la cohesión de cada grupo final. Si todas las muestras dentro de un grupo final están lejos unas de otras (es decir, ninguna distancia entre pares dentro del grupo está en el percentil inferior, por ejemplo, 50 de todas las distancias), entonces el grupo puede disolverse. A continuación, cada muestra puede marcarse como un singleton.

En el bloque 420, el método 400 puede incluir fusionar el conjunto de marcadores generados en el bloque 416 para generar otro conjunto de marcadores. En algunas realizaciones, la fusión puede ser un proceso de dos etapas. En la primera etapa, cada singleton puede colocarse en el mismo grupo que su vecino más cercano

determinado usando la estructura de datos de distancias (por ejemplo, matriz de distancias) del bloque de preprocesamiento 408. Si la distancia de un singleton a su vecino más cercano se encuentra dentro del 10% superior de las distancias (es decir, está lejos de todas las demás células), el singleton puede marcarse como un valor atípico y permanece en su propio grupo. Este primer paso garantiza que todos los grupos contengan por lo menos dos miembros no atípicos. En el segundo paso, una vez eliminados los valores atípicos, se calculan las distancias por pares entre los grupos usando varias pruebas estadísticas, lo que da como resultado una matriz de distancias por pares entre grupos D_c. La distancia entre dos grupos se calcula como el logaritmo negativo del valor p más pequeño obtenido en pruebas t de Welch independientes para todos los genes. A partir de la distancia total más pequeña, se fusionan los dos grupos correspondientes. Se calculan las distancias del nuevo grupo a todos los grupos existentes y se repite el proceso hasta que todas las distancias entre pares superan una cierta distancia. El paquete también ofrece un enfoque de fusión basado en la detección de comunidades mediante la ejecución del algoritmo de Lovaina en D_c.

En el bloque 424, el método 400 puede incluir seleccionar características del conjunto de marcadores determinadas en el bloque 420 para identificar características para distinguir tipos de células. En algunas realizaciones, el método 400 puede realizar dos tipos de selección de características usando los marcadores generados a partir del bloque de fusión 420. Para el primer tipo de selección de características, durante el bloque de división y prueba, cada vez que se mantiene una división, se guardan los K genes con los K valores p más pequeños. Pueden guardarse más genes de las divisiones más cercanas a la parte superior del dendrograma. Finalmente, se obtiene una lista de genes únicos de la unión de todas las divisiones. Para el segundo tipo de selección de características, para cada grupo, se realizan varias pruebas de uno contra el resto usando sólo los genes que tienen medias más altas en el grupo de interés. Puede obtenerse una tabla de genes importantes para cada grupo junto con información adicional sobre cada gen (por ejemplo, valor p, cambio de veces, nivel de expresión medio dentro del grupo).

El método 400 puede incluir la realización de un análisis exploratorio. En algunas realizaciones, el método 400 puede utilizar varias funciones para visualizar ciertos pasos en las etapas de división y fusión. Por ejemplo, estas funciones ilustran las células implicadas en la división (o fusión), las células que terminan en cada subárbol (o el grupo combinado), y los genes que dictaron esta división (o fusión). Como otro ejemplo, el método 400 puede realizar comparaciones por pares entre todos los grupos (por ejemplo, para determinar qué genes distinguen cada par de grupos) y funciones para dibujar el dendrograma. El método 400 puede basarse en las distribuciones de distancias por pares dentro de los grupos. En algunas realizaciones, el método 400 puede incluir realizar barridos de parámetros.

Agrupación de perfiles de expresión

En la presente se divulgan métodos para identificar dianas para distinguir tipos de células. En algunas realizaciones, el método comprende: (a) recibir perfiles de expresión de una pluralidad de células, en donde los perfiles de expresión comprenden un número de cada diana de una pluralidad de dianas para cada célula de la pluralidad de células; (b) agrupar los perfiles de expresión de la pluralidad de células para generar una pluralidad de grupos de perfiles de expresión basados en las distancias entre los perfiles de expresión de la pluralidad de células, en donde cada grupo tiene una o más asociaciones con uno o ambos de (1) un grupo padre y (2) dos o más grupos hijo, en donde el grupo padre representa perfiles de expresión de una o más células de la pluralidad de células representadas por el grupo, y en donde el grupo representa perfiles de expresión representados por los dos o más grupos hijo; (c) para cada grupo con dos o más grupos hijo, si las asociaciones entre el grupo con los dos o más grupos hijos no son válidas (por ejemplo, las diferencias entre los dos o más grupos hijos no son significativas) añadir el grupo a un conjunto de grupos de fusión; (d) iterativamente, para cada primer grupo en el conjunto de grupos de fusión, si una distancia entre el primer grupo en el conjunto de grupos de fusión y un segundo grupo en el conjunto de grupos de fusión que es el más cercano al primer grupo está dentro de un umbral de distancia de fusión, fusionar el primer grupo y el segundo grupo para generar un grupo fusionado, en donde el grupo fusionado comprende perfiles de expresión del primer grupo y del segundo grupo; y (e) para cada grupo del conjunto de grupos de fusión, identificar dianas para distinguir tipos de células basándose en los perfiles de expresión de la pluralidad de dianas de las células representadas por el grupo.

La FIG. 6 es un diagrama de flujo que muestra un método ejemplar no limitativo para identificar dianas para distinguir tipos de células agrupando perfiles de expresión de células. En el bloque 604, el método 600 recibe perfiles de expresión de una pluralidad de células. Cada perfil de expresión puede comprender un número de cada diana de una pluralidad de dianas para una célula diferente de la pluralidad de células. En algunas realizaciones, la recepción de perfiles de expresión de la pluralidad de células comprende recibir una estructura de datos de recuentos de dianas (por ejemplo, una matriz de recuentos de dianas). Cada fila de la matriz de recuentos de dianas puede comprender un perfil de expresión de una célula de una pluralidad de células.

En diferentes implementaciones el número de perfiles de expresión recibidos puede ser diferente. En algunas realizaciones, el número de perfiles de expresión recibidos puede ser, o ser aproximadamente, de 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 un intervalo entre dos cualesquiera de estos valores. En algunas realizaciones, el número de perfiles de expresión recibidos puede ser como mínimo, 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, o 10000.

En algunas realizaciones, el método 600 comprende: antes de recibir los perfiles de expresión de la pluralidad de células, en el bloque 604: codificar estocásticamente con códigos de barras la pluralidad de dianas en la pluralidad de células usando una pluralidad de códigos de barras estocásticos para crear una pluralidad de dianas codificadas estocásticamente con códigos de barras, en donde cada una de la pluralidad de códigos de barras estocásticos comprende un marcador celular y un marcador molecular, en donde las dianas codificadas estocásticamente con códigos de barras creadas a partir de dianas de diferentes células tienen diferentes marcadores celulares, y en donde las dianas codificadas estocásticamente con códigos de barras creadas a partir de dianas de una célula de la pluralidad de células tienen diferentes marcadores moleculares; obtener datos de secuenciación de la pluralidad de dianas codificadas estocásticamente con códigos de barras; y para cada una de la pluralidad de células (1) contar el número de marcadores moleculares con secuencias distintas asociadas con cada diana de la pluralidad de dianas en los datos de secuenciación para la célula; y (2) estimar el número de cada diana de la pluralidad de dianas para la célula basándose en el número de marcadores moleculares con secuencias distintas asociadas con la diana en los datos de secuenciación contados en (1). En algunas realizaciones, el perfil de expresión de la célula de la pluralidad de células comprende el número de cada diana de la pluralidad de dianas para la célula estimado en (2).

En el bloque 608, el método 600 puede incluir agrupar los perfiles de expresión de la pluralidad de células para generar una pluralidad de grupos de perfiles de expresión. El método 600 puede generar las grupos de perfiles de expresión basándose en las distancias entre los perfiles de expresión de la pluralidad de células. En diferentes implementaciones el número de perfiles de expresión representados por cada grupo puede ser diferente. En algunas realizaciones, el número de perfiles de expresión representados por cada grupo puede ser, o ser aproximadamente, de 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, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10000, o un número o un intervalo entre dos cualesquiera de estos valores. En algunas realizaciones, el número de perfiles de expresión representados por cada grupo puede ser por lo menos, o 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, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, o 10000. En algunas realizaciones, los perfiles de expresión representados por cada grupo pueden ser, o ser aproximadamente, del 1%, 2%, 3%, 4%, 5%, 6%, 7%, 8%, 9%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, 100%, o un número o un intervalo entre dos cualesquiera de estos valores, del número de perfiles de expresión recibidos en el bloque 604. En algunas realizaciones, los perfiles de expresión representados por cada grupo pueden ser por lo menos, o como máximo, del 1%, 2%, 3%, 4%, 5%, 6%, 7%, 8%, 9%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, o 100%, del número de perfiles de expresión recibidos en el bloque 604.

Cada grupo puede tener una asociación con uno o ambos de (1) un grupo padre y (2) dos o más grupos hijos (como 3, 4, 5, 6, 7, 8, 9, 10, o más grupos hijos). El grupo padre representa perfiles de expresión de una o más células de la pluralidad de células representadas por el grupo. El grupo representa perfiles de expresión representados por sus dos o más grupos hijos.

En algunas realizaciones, los perfiles de expresión pueden agruparse como se describe con referencia a la FIG. 4, como el bloque 412 de la FIG. 4. Por ejemplo, el método 600 puede incluir agrupar jerárquicamente los perfiles de expresión de la pluralidad de células para generar un dendrograma que representa los perfiles de expresión de la pluralidad de células basándose en las distancias entre los perfiles de expresión de la pluralidad de células. El dendrograma puede comprender una pluralidad de grupos. La pluralidad de grupos puede comprender un grupo de raíces, una pluralidad de grupos de hojas y una pluralidad de grupos que no son raíces ni hojas. El número de grupos hoja puede ser, por ejemplo, el mismo que el número de perfiles de expresión n. El número de grupos no raíz y no hoja puede ser, por ejemplo, n-2.

Cada uno de la pluralidad de grupos de hoja y la pluralidad de grupos no raíz y no hoja puede tener una asociación con un grupo padre. Cada uno de los grupos raíz y la pluralidad de grupos no raíz y no hoja puede tener asociaciones con un grupo hijo izquierdo y un grupo hijo derecho y representa perfiles de expresión representados por el grupo hijo izquierdo y el grupo hijo derecho del grupo. El grupo raíz puede representar los perfiles de expresión de la pluralidad de células. En algunas implementaciones, un grupo hoja puede representar un perfil de expresión de una célula. Un grupo no raíz y no hoja puede representar perfiles de expresión de células representadas por los grupos hijos de los grupos no raíz y no hoja. El grupo raíz puede representar perfiles de expresión de sus grupos hijos.

En algunas realizaciones, la agrupación de los perfiles de expresión de la pluralidad de células sobre la base de las distancias entre los perfiles de expresión de la pluralidad de células en el bloque 608 comprende: asignar cada perfil de expresión de la pluralidad de células a un grupo de hoja diferente en la pluralidad de grupos; y combinar iterativamente un primer grupo y un segundo grupo de la pluralidad de grupos para generar un grupo padre del primer grupo y del segundo grupo si el segundo grupo es el grupo más cercano de la pluralidad de grupos al primer grupo. La distancia entre el primer grupo y el segundo grupo puede ser la distancia máxima entre cualquier célula con un perfil de expresión representado por el primer grupo y cualquier célula con un perfil de expresión representado por el segundo grupo.

En algunas realizaciones, una correlación intragrupo del primer grupo y una correlación intragrupo del

segundo grupo son mayores que una correlación intergrupo del primer grupo y del segundo grupo. Una medida o una indicación de una correlación intragrupo del primer grupo y una correlación intragrupo del segundo grupo es mayor que una correlación intergrupo del primer grupo y el segundo grupo. La medida de la correlación intragrupo del primer grupo y la correlación intragrupo del segundo grupo puede basarse en por lo menos una de: una correlación máxima intragrupo del primer grupo y el segundo grupo, una correlación media intragrupo del primer grupo y el segundo grupo, una correlación mediana intragrupo del primer grupo y el segundo grupo, una correlación mínima intragrupo del primer grupo y el segundo grupo, y cualquier combinación de las mismas. La correlación intragrupo del primer grupo puede basarse en por lo menos una de: una correlación máxima intragrupo del primer grupo, una correlación media intragrupo del primer grupo, una correlación mediana intragrupo del primer grupo, una correlación mínima intragrupo del primer grupo y cualquier combinación de las mismas. La correlación intragrupo del segundo grupo puede basarse en por lo menos una de: una correlación máxima intragrupo del segundo grupo, una correlación media intragrupo del segundo grupo, una correlación mediana intragrupo del segundo grupo, una correlación mínima intragrupo del segundo grupo y cualquier combinación de las mismas. La correlación intergrupo del primer grupo y el segundo grupo puede basarse en por lo menos una de las siguientes: una correlación máxima intergrupo del primer grupo y el segundo grupo, una correlación media intergrupo del primer grupo y el segundo grupo, una correlación mediana intergrupo del primer grupo y el segundo grupo, una correlación mínima intergrupo del primer grupo y el segundo grupo, y cualquier combinación de las mismas.

En algunas realizaciones, el método 600 puede incluir, antes de agrupar los perfiles de expresión de la pluralidad de células para generar la pluralidad de grupos de perfiles de expresión sobre la base de las distancias entre los perfiles de expresión de la pluralidad de células en el bloque 608: determinar una estructura de datos de distancias (por ejemplo, una matriz de distancias) de los perfiles de expresión de la pluralidad de células. La matriz puede tener una dimensión de n x n, en donde n denota el número de perfiles de expresión recibidos en el bloque 604. Cada elemento diagonal de la matriz de distancias puede tener una dimensión de n x n. Cada elemento diagonal de la matriz de distancias tiene un valor de cero. La agrupación de los perfiles de expresión de la pluralidad de células para generar la pluralidad de grupos de perfiles de expresión sobre la base de las distancias entre los perfiles de expresión de la pluralidad de células en el bloque 608 puede comprender: agrupar los perfiles de expresión de la pluralidad de células para generar la pluralidad de grupos de perfiles de expresión sobre la base de la estructura de datos de distancia. Las distancias entre los perfiles de expresión de la pluralidad de células pueden ser distancias de correlación por pares entre los perfiles de expresión de la pluralidad de células.

En algunas realizaciones, el método 600 puede incluir, antes de determinar la estructura de datos de distancias en (i), transformar logarítmicamente la estructura de datos de recuentos de dianas en una estructura de datos de recuentos de dianas transformada logarítmicamente (por ejemplo, una matriz de recuentos de dianas transformada logarítmicamente). Determinar la estructura de datos de distancias de los elementos de la estructura de datos de recuentos de dianas comprende determinar la estructura de datos de distancias de la estructura de datos de recuentos de dianas transformada logarítmicamente. La agrupación de los perfiles de expresión de la pluralidad de células para generar la pluralidad de grupos de perfiles de expresión sobre la base de las distancias entre los perfiles de expresión de la pluralidad de células en el bloque 608 puede comprender agrupar los perfiles de expresión de la pluralidad de células sobre la base de la estructura de datos de recuentos de dianas transformada logarítmicamente y la estructura de datos de distancias para generar la pluralidad de grupos. La transformación logarítmica de la estructura de datos de recuentos de dianas en la estructura de datos de recuentos de dianas transformada logarítmicamente puede comprender aumentar el valor de cada elemento de la estructura de datos de recuentos de dianas en un incremento. El incremento puede ser, por ejemplo, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, o más.

En el bloque 612, el método 600 puede incluir añadir cada grupo de los grupos de los perfiles de expresión con dos o más grupos hijos a un conjunto de grupos de fusión si las asociaciones entre el grupo y sus grupos hijos son inválidas (por ejemplo, las diferencias entre los grupos hijos no son significativas). En algunas realizaciones, si en el bloque 608 se han agrupado los perfiles de expresión para generar un dendrograma, el método 600 puede añadir cada grupo con dos o más grupos hijos a un conjunto de grupos de fusión dividiendo y probando el dendrograma para generar un conjunto de marcadores como se describe con referencia a la FIG. 4, como el bloque 416 de la FIG. 4.

En algunas realizaciones, para cada grupo con dos o más grupos hijos, si las asociaciones entre el grupo con los dos o más grupos hijos son inválidas, el método 600 puede añadir el grupo a un conjunto de grupos de fusión mediante: mientras se atraviesa cada grupo del dendrograma desde el grupo raíz del dendrograma hasta la pluralidad de grupos hoja del dendrograma: (1) determinar si las asociaciones del grupo con los grupos hijos del grupo son válidas o inválidas; y (2) si las asociaciones son inválidas, añadir el grupo a un conjunto de grupos de fusión.

En el bloque 616, el método 600 puede incluir fusionar cada grupo en el conjunto de grupos de fusión con su grupo más cercano en el conjunto de grupos de fusión si una distancia entre los dos grupos está dentro de un umbral de distancia de fusión. El grupo fusionado comprende perfiles de expresión del primer grupo y del segundo grupo. El método 600 puede fusionar cada grupo en el conjunto de grupos de fusión con su grupo más cercano como se describe con referencia a la FIG. 4, como el bloque 420 de la FIG. 4.

En algunas realizaciones, el método 600 puede incluir, en cada grupo cuando se atraviesa la pluralidad de

grupos del dendrograma: si las asociaciones son válidas, continuar atravesando desde el grupo hasta el grupo hijo izquierdo y el grupo hijo derecho del grupo; y si las asociaciones son inválidas, interrumpir el recorrido desde el grupo hasta el grupo hijo izquierdo y el grupo hijo derecho del grupo. Determinar si las asociaciones del grupo con los grupos hijos del grupo son válidas o inválidas puede comprender: determinar que las asociaciones son válidas si la distancia entre el grupo hijo izquierdo y el grupo hijo derecho está por encima de un umbral de asociación, y si no lo está son inválidas.

En algunas realizaciones, por lo menos una de una correlación intranodo del primer nodo y una correlación intranodo del segundo nodo puede ser mayor que una correlación internodo del primer nodo y el segundo nodo. Una medida o una indicación de una correlación intranodo del primer nodo y una correlación intranodo del segundo nodo puede ser mayor que una correlación internodo del primer nodo y el segundo nodo. La medida de la correlación intranodo del primer nodo y la correlación intranodo del segundo nodo puede basarse en por lo menos una de: una correlación máxima intranodo del primer nodo y el segundo nodo, una correlación media intranodo del primer nodo y el segundo nodo, una correlación mediana intranodo del primer nodo y el segundo nodo, una correlación mínima intranodo del primer nodo y el segundo nodo, y cualquier combinación de las mismas. La correlación intranodo del primer nodo puede basarse en por lo menos una de: una correlación máxima intranodo del primer nodo, una correlación media intranodo del primer nodo, una correlación mediana intranodo del primer nodo, una correlación mínima intranodo del primer nodo, y cualquier combinación de las mismas. La correlación intranodo del segundo nodo puede basarse en por lo menos una de: una correlación máxima intranodo del segundo nodo, una correlación media intranodo del segundo nodo, una correlación mediana intranodo del segundo nodo, una correlación mínima intranodo del segundo nodo, y cualquier combinación de las mismas. La correlación internodo del primer nodo y el segundo nodo puede basarse en por lo menos una de: una correlación máxima internodo del primer nodo y el segundo nodo, una correlación media internodo del primer nodo y el segundo nodo, una correlación mediana internodo del primer nodo y el segundo nodo, una correlación mínima internodo del primer nodo y el segundo nodo, y cualquier combinación de las mismas.

En algunas realizaciones, la distancia entre el grupo hijo izquierdo y el grupo hijo derecho puede determinarse basándose en una prueba estadística realizada en cada diana de la pluralidad de dianas entre los perfiles de expresión representados por el grupo hijo izquierdo y el grupo hijo derecho. La prueba estadística puede ser, por ejemplo, una prueba t de Welch. La distancia entre el grupo hijo izquierdo y el grupo hijo derecho puede determinarse basándose en el valor p máximo de la prueba estadística realizada en cada diana de la pluralidad de dianas entre el perfil de expresión representado por el grupo hijo izquierdo y cada perfil de expresión representado por el grupo hijo derecho.

En algunas realizaciones, el método 600 comprende, antes de fusionar el primer grupo con el segundo grupo para generar el grupo fusionado en el bloque 616: fusionar cada tercer grupo en el conjunto de grupos de fusión que representa un perfil de expresión de una única célula con un cuarto grupo en el conjunto de grupos de fusión si una distancia entre el tercer grupo y el cuarto grupo está dentro de un umbral de distancia de grupo. El método puede comprender clasificar la pluralidad de células basándose en los grupos del conjunto de grupos de fusión que representan perfiles de expresión de las células. El método puede comprender diseñar un ensayo de transcriptoma completo sobre la base de las dianas para distinguir los tipos de células identificados o diseñar un ensayo de transcriptoma dirigido basado en las dianas para distinguir los tipos de células identificados.

En algunas realizaciones, el método 600 comprende, en cada grupo cuando se atraviesa la pluralidad de grupos del dendrograma: (3) añadir el grupo al conjunto de grupos de fusión si el grupo representa un perfil de expresión de una única célula. El método puede comprender, en cada grupo al atravesar la pluralidad de grupos del dendrograma: asignar un marcador de grupo al grupo. En algunas realizaciones, si el grupo representa un perfil de expresión de una única célula, el marcador de grupo del grupo comprende una designación de una única célula, de lo contrario, si el grupo es el grupo hijo izquierdo del grupo padre, el marcador de grupo del grupo comprende el marcador de grupo del grupo padre y una designación izquierda, y de lo contrario, el marcador de grupo del grupo comprende el marcador de grupo del grupo padre y una designación derecha.

En el bloque 620, el método 600 puede incluir identificar dianas para distinguir tipos de células basándose en perfiles de expresión de la pluralidad de dianas de células representadas por cada grupo en el conjunto de grupos de fusión. El método 600 puede identificar las dianas para distinguir tipos de células como se describe con referencia a la FIG. 4, como el bloque 424 de la FIG. 4. En algunas realizaciones, para cada grupo en el conjunto de grupos de fusión, la identificación de las dianas para distinguir los tipos de células basándose en los perfiles de expresión de la pluralidad de dianas de las células representadas por el grupo comprende: determinar una diferencia, entre los perfiles de expresión representados por el grupo y los perfiles de expresión representados por otro grupo en el conjunto de grupos de fusión, en números de marcadores moleculares con secuencias distintas asociadas con las dianas para distinguir los tipos de células es mayor que un umbral de significancia.

Secuenciación

En algunas realizaciones, estimar el número de dianas codificadas con códigos de barras diferentes (por ejemplo, dianas codificadas estocásticamente con códigos de barras) puede comprender determinar las secuencias

de las dianas marcadas, el marcador espacial, el marcador molecular, el marcador de la 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 una diana codificada con código de barras (por ejemplo, una diana codificada estocásticamente con código de barras) o cualquier producto de la misma puede comprender la realización de una reacción de secuenciación para determinar la secuencia de por lo menos una parte de un marcador de muestra, un marcador espacial, un marcador celular, un marcador molecular, por lo menos una parte de la diana marcada (por ejemplo, una diana marcada estocásticamente), un complemento de la misma, un complemento inverso de la misma, o cualquier combinación de los mismos.

La determinación de la secuencia de una diana codificada con código de barras o una diana codificada estocásticamente con código de barras (por ejemplo, ácido nucleico amplificado, ácido nucleico marcado, copia de ADNc de un ácido nucleico marcado, etc.) puede realizarse usando una variedad de métodos de secuenciación que incluyen, pero no se limitan a, secuenciación por hibridación (SBH) o la secuenciación por ligadura (SBL), secuenciación por adición cuantitativa incremental de nucleótidos fluorescentes (QIFNAS), ligadura y escisión por pasos, transferencia de energía por resonancia de fluorescencia (FRET), balizas moleculares, digestión con sonda informadora TaqMan, pirosecuenciación, secuenciación in situ fluorescente (FISSEQ), perlas FISSEQ, secuenciación Wobble, secuenciación multiplex, secuenciación de colonias polimerizadas (POLONY); secuenciación por círculo rodante en nanorejilla (ROLONY), ensayos de ligadura de oligo específicos de alelos (por ejemplo, ensayo de ligadura oligo (OLA), OLA de molécula de plantilla única usando una sonda lineal ligada y una lectura de amplificación por círculo rodante (RCA), sondas de candado ligadas u OLA de molécula de plantilla única usando una sonda de candado circular ligada y una lectura de amplificación de círculo rodante (RCA), y similares.

En algunas realizaciones, la determinación de la secuencia de la diana codificada con código barras (por ejemplo, la diana codificada estocásticamente con código barras) o de cualquier producto de la misma comprende secuenciación de extremos emparejados, secuenciación de nanoporos, secuenciación de alto rendimiento, secuenciación de escopeta, secuenciación de colorante-terminador, secuenciación de ADN de cebadores múltiples, paseos de cebadores, secuenciación dideoxídica de Sanger, secuenciación de Maxim-Gilbert, pirosecuenciación, secuenciación de molécula única real o cualquier combinación de las mismas. Alternativamente, la secuencia de la diana codificada con código de barras o cualquier producto de la misma puede determinarse mediante microscopía electrónica o una matriz de transistores de efecto campo químico-sensibles (chemFET).

Pueden utilizarse métodos de secuenciación de alto rendimiento, como la secuenciación de matriz cíclica usando plataformas como Roche 454, Illumina Solexa, ABI-SOLiD, ION Torrent, Complete Genomics, Pacific Bioscience, Helicos o la plataforma Polonator. En algunas realizaciones, la secuenciación puede comprender la secuenciación MiSeq. En algunas realizaciones, la secuenciación puede incluir la secuenciación HiSeq.

Las dianas marcadas (por ejemplo, las dianas marcadas estocásticamente) pueden comprender ácidos nucleicos que representen 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, puede secuenciarse 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 usando una región complementaria diana que comprende una pluralidad de multímeros capturando los genes que contienen una secuencia complementaria de la muestra. En algunas realizaciones, las dianas codificadas con códigos de barras comprenden ácidos nucleicos que representan de aproximadamente el 0,01% de los transcritos del transcriptoma de un organismo hasta aproximadamente el 100% de los transcritos del transcriptoma de un organismo. Por ejemplo, puede secuenciarse desde aproximadamente el 0,501% de los transcritos del transcriptoma de un organismo hasta aproximadamente el 100% de los transcritos del transcriptoma de un organismo usando una región complementaria diana que comprende una cola poli(T) capturando los ARNm de la muestra.

La determinación de las secuencias de los marcadores espaciales y los marcadores moleculares de la pluralidad de códigos de barras (por ejemplo, códigos de barras estocásticos) puede incluir la secuenciación del 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 intervalo entre dos cualesquiera de estos valores, de la pluralidad de códigos de barras. La determinación de las secuencias de los marcadores de la pluralidad de códigos de barras, por ejemplo los marcadores de muestra, los marcadores espaciales y los marcadores moleculares, puede incluir la secuenciación de 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^{15}, 10^{19}, 10^{20}, o un número o un intervalo entre dos cualesquiera de estos valores, de la pluralidad de códigos de barras. La secuenciación de una parte o de la totalidad de la pluralidad de códigos de barras puede incluir la generación de secuencias con longitudes de lectura de, de aproximadamente, de por lo menos, o de 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 un intervalo entre dos cualesquiera de estos valores, de nucleótidos o bases.

La secuenciación puede comprender la secuenciación de por lo menos o por lo menos aproximadamente 10, 20, 30, 40, 50, 60, 70, 80, 90, 100 o más nucleótidos o pares de bases de las dianas codificadas con códigos de barras.

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 amplificación por reacción en cadena de la polimerasa (PCR) de la pluralidad de dianas codificadas con códigos de barras. La secuenciación puede comprender secuenciar por lo menos o por lo menos aproximadamente 200, 300, 400, 500, 600, 700, 800, 900, 1.000 o más nucleótidos o pares de bases de las dianas codificadas con códigos de barras. La secuenciación puede comprender secuenciar por lo menos 1500, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, o 10000 o más nucleótidos o pares de bases de las dianas codificadas con códigos de barras.

La secuenciación puede comprender por lo menos aproximadamente 200, 300, 400, 500, 600, 700, 800, 900, 1.000 o más lecturas de secuenciación por serie. En algunas realizaciones, la secuenciación comprende secuenciar por lo menos 1500, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000 o 10000 o más lecturas de secuenciación por serie. La secuenciación puede comprender menos de o igual a aproximadamente 1.600.000.000 lecturas de secuenciación por serie. La secuenciación puede comprender menos de o igual a aproximadamente 200.000.000 de lecturas de secuenciación por serie.

Muestras

En algunas realizaciones, la pluralidad de dianas puede estar comprendida en una o más muestras. Una muestra puede comprender una o más células, o ácidos nucleicos de una o más células. Una muestra puede ser una única célula o ácidos nucleicos de una única célula. La una o más células pueden ser de uno o más tipos de células. Por lo menos uno del uno o más tipos de células puede ser célula cerebral, célula cardiaca, célula cancerosa, célula tumoral circulante, célula de órgano, célula epitelial, célula metastásica, célula benigna, célula primaria, célula circulatoria, o cualquier combinación de las mismas.

Una muestra para uso en el método de la divulgación puede comprender una o más células. Una muestra puede referirse a una o más células. En algunas realizaciones, la pluralidad de células puede incluir uno o más tipos de células. Por lo menos uno del uno o más tipos de células puede ser célula cerebral, célula cardiaca, célula cancerosa, célula tumoral circulante, célula de órgano, célula epitelial, célula metastásica, célula benigna, célula primaria, célula circulatoria, o cualquier combinación de las mismas. En algunas realizaciones, 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 cerebral, melanoma y cánceres de piel no melanoma, y similares. En algunas realizaciones, las células se derivan de un cáncer pero se recogen de un fluido corporal (por ejemplo, células tumorales circulantes). Ejemplos no limitativos 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 puede obtenerse de un 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.

En algunas realizaciones, las células son células que han sido infectadas con virus y contienen oligonucleótidos víricos. En algunas realizaciones, la infección vírica puede estar provocada por un virus como los virus de ADN de cadena sencilla (cadena + o "de sentido") (por ejemplo, parvovirus), o virus de ARN de cadena doble (por ejemplo, reovirus). En algunas realizaciones, las células son bacterias. Estas pueden incluir o bacterias grampositivas o gramnegativas. En algunas realizaciones, las células son hongos. En algunas realizaciones, las células son protozoos u otros parásitos.

Como se usa en la presente, el término "célula" puede referirse a una o más células. En algunas realizaciones, 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 realizaciones, 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 realizaciones, las células son otros tipos de células animales o vegetales. En otras realizaciones, las células pueden ser células procariotas o eucariotas.

En algunas realizaciones, las células se clasifican antes de asociar una célula a una perla. Por ejemplo, las células pueden clasificarse mediante clasificación celular activada por fluorescencia o clasificación celular activada por magnetismo, o de manera más general mediante citometría de flujo. Las células pueden filtrarse por tamaño. En algunas realizaciones, una fracción retenida contiene las células que se van a asociar a la perla. En algunas realizaciones, el flujo contiene las células que se van a asociar a la perla.

Una muestra puede referirse a una pluralidad de células. La muestra puede referirse a una monocapa de células. La muestra puede referirse a una sección delgada (por ejemplo, una sección delgada de tejido). La muestra puede referirse a una colección sólida o semisólida de células que pueden colocarse en una dimensión en una matriz.

Entorno de ejecución

La presente divulgación proporciona sistemas informáticos que están programados para implementar métodos de la divulgación (por ejemplo, el método 400 o el método 600). La FIG. 7 muestra un sistema informático 700 que está programado o configurado de otro modo para implementar cualquiera de los métodos divulgados en la presente. El sistema informático 700 puede ser un dispositivo electrónico de un usuario o un sistema informático que está localizado remotamente con respecto al dispositivo electrónico. El dispositivo electrónico puede ser un dispositivo electrónico móvil.

El sistema informático 700 incluye una unidad central de procesamiento (CPU, también "procesador" y "procesador informático" en la presente) 705, que puede ser un procesador de núcleo único o de núcleo múltiple, o una pluralidad de procesadores para procesamiento paralelo. El sistema informático 700 también incluye memoria o ubicación de memoria 710 (por ejemplo, memoria de acceso aleatorio, memoria de sólo lectura, memoria flash), unidad de almacenamiento electrónico 715 (por ejemplo, disco duro), interfaz de comunicaciones 720 (por ejemplo, adaptador de red) para comunicarse con uno o más sistemas diferentes, y dispositivos periféricos 725, como caché, otra memoria, almacenamiento de datos y/o adaptadores de pantalla electrónica. La memoria 710, la unidad de almacenamiento 715, la interfaz 720 y los dispositivos periféricos 725 están en comunicación con la CPU 705 a través de un bus de comunicación (líneas continuas), como una placa base. La unidad de almacenamiento 715 puede ser una unidad de almacenamiento de datos (o repositorio de datos) para almacenar datos. El sistema informático 700 puede estar acoplado operativamente a una red informática ("red") 730 con la ayuda de la interfaz de comunicaciones 720. La red 730 puede ser Internet, una Internet y/o extranet, o una intranet y/o extranet que está en comunicación con Internet. En algunos casos, la red 730 es una red de telecomunicaciones y/o de datos. La red 730 puede incluir uno o más servidores informáticos, que pueden permitir la computación distribuida, como la computación en nube. La red 730, en algunos casos con la ayuda del sistema informático 700, puede implementar una red de pares, que puede permitir que los dispositivos acoplados al sistema informático 700 se comporten como un cliente o un servidor.

La CPU 705 puede ejecutar una secuencia de instrucciones legibles por máquina, que pueden estar incorporadas en un programa o software. Las instrucciones pueden almacenarse en una ubicación de memoria, como la memoria 710. Las instrucciones pueden dirigirse a la CPU 705, que posteriormente puede programar o configurar de otro modo la CPU 705 para que implemente los métodos de la presente divulgación. Los ejemplos de operaciones realizadas por la CPU 705 pueden incluir extraer, decodificar, ejecutar y reescribir. La CPU 705 puede ser parte de un circuito, como un circuito integrado. En el circuito pueden incluirse uno o más de otros componentes del sistema 700. En algunos casos, el circuito es un circuito integrado de aplicación específica (ASIC).

La unidad de almacenamiento 715 puede almacenar archivos, como controladores, bibliotecas y programas guardados. La unidad de almacenamiento 715 puede almacenar datos de usuario, por ejemplo, preferencias de usuario y programas de usuario. El sistema informático 700 en algunos casos puede incluir una o más unidades de almacenamiento de datos adicionales que son externas al sistema informático 700, como localizadas en un servidor remoto que está en comunicación con el sistema informático 700 a través de una intranet o Internet.

El sistema informático 700 puede comunicarse con uno o más sistemas informáticos remotos a través de la red 730. Por ejemplo, el sistema informático 700 puede comunicarse con un sistema informático remoto de un usuario (por ejemplo, un microbiólogo). Los ejemplos de sistemas informáticos remotos incluyen ordenadores personales (por ejemplo, PC portátiles), pizarras o tabletas (por ejemplo, Apple® iPad, Samsung® Galaxy Tab), teléfonos, teléfonos inteligentes (por ejemplo, Apple® iPhone, dispositivos con Android, Blackberry®) o asistentes digitales personales. El usuario puede acceder al sistema informático 700 a través de la red 730.

El sistema informático 700 puede incluir o estar en comunicación con una pantalla electrónica 735 que comprende una interfaz de usuario (UI) 740 para proporcionar, por ejemplo, una salida indicativa de la coocurrencia de cadenas o interacciones de una pluralidad de taxones de microorganismos, representados por cadenas. Ejemplos de UI incluyen, sin limitación, una interfaz gráfica de usuario (GUI) y una interfaz de usuario basada en web.

Los métodos descritos en la presente pueden implementarse por medio de código ejecutable por máquina (por ejemplo, procesador de ordenador) almacenado en una ubicación de almacenamiento electrónico del sistema informático 700 como, por ejemplo, en la memoria 710 o la unidad de almacenamiento electrónico 715. El código ejecutable por máquina o legible por máquina puede proporcionarse en forma de software. Durante su uso, el código puede ser ejecutado por el procesador 705. En algunos casos, el código puede recuperarse de la unidad de almacenamiento 715 y almacenarse en la memoria 710 para acceso inmediato por el procesador 705. En algunas situaciones, puede excluirse la unidad de almacenamiento electrónico 715, y las instrucciones ejecutables por máquina se almacenan en la memoria 710.

El código puede precompilarse y configurarse para su uso con una máquina que tenga un procesador adaptado para ejecutar el código, o puede compilarse durante el tiempo de ejecución. El código puede suministrarse en un lenguaje de programación que puede seleccionarse para permitir que el código se ejecute de manera precompilada o tal como está compilado.

Los aspectos de los sistemas y métodos que se proporcionan en la presente, como el sistema informático 700, pueden incorporarse en programación. Varios aspectos de la tecnología pueden considerarse "productos" o "artículos de fabricación", típicamente en forma de código ejecutable por máquina (o procesador) y/o datos asociados que se transportan o incorporan en un tipo de medio legible por máquina. El código ejecutable por máquina puede almacenarse en una unidad de almacenamiento electrónico, como una memoria (por ejemplo, memoria de sólo lectura, memoria de acceso aleatorio, memoria flash) o un disco duro. Los medios de tipo "almacenamiento" pueden incluir cualquiera o toda la memoria tangible de los ordenadores, procesadores o similares, o módulos asociados a los mismos, como las varias memorias semiconductoras, unidades de cinta, unidades de disco y similares, que pueden proporcionar almacenamiento no transitorio en cualquier momento para la programación del software. La totalidad o parte del software puede comunicarse en ocasiones a través de Internet o de otras varias redes de telecomunicaciones. Tales comunicaciones, por ejemplo, pueden permitir la carga del software desde un ordenador o procesador a otro, por ejemplo, desde un servidor de gestión o un ordenador central a la plataforma informática de un servidor de aplicaciones. Por tanto, otro tipo de medios que pueden llevar los elementos de software incluyen las ondas ópticas, eléctricas y electromagnéticas, como las que se usan a través de interfaces físicas entre dispositivos locales, a través de redes fijas cableadas y ópticas y a través de varios enlaces aéreos. Los elementos físicos que transportan dichas ondas, como los enlaces por cable o inalámbricos, los enlaces ópticos o similares, también pueden considerarse medios portadores de software. Como se usan en la presente, a menos que se limiten a medios de "almacenamiento" tangibles no transitorios, términos como "medio legible" por ordenador o máquina se refieren a cualquier medio que participe en el suministro de instrucciones a un procesador para su ejecución.

Por lo tanto, un medio legible por máquina, como un código ejecutable por ordenador, puede adoptar muchas formas incluyendo, entre otras, un medio de almacenamiento tangible, un medio de onda portadora o un medio de transmisión física. Los medios de almacenamiento no volátiles incluyen, por ejemplo, discos ópticos o magnéticos, como cualquiera de los dispositivos de almacenamiento de cualquier ordenador u ordenadores o similares, como los que pueden usarse para implementar las bases de datos, etc. que se muestran en los dibujos. Los medios de almacenamiento volátiles incluyen la memoria dinámica, como la memoria principal de dicha plataforma informática. Los medios de transmisión tangibles incluyen cables coaxiales, cables de cobre y fibra óptica, incluyendo los cables que componen un bus dentro de un sistema informático. Los medios de transmisión de ondas portadoras pueden adoptar la forma de señales eléctricas o electromagnéticas, u ondas acústicas o luminosas como las generadas durante las comunicaciones de datos por radiofrecuencia (RF) e infrarrojos (IR). Las formas comunes de medios legibles por ordenador incluyen, por lo tanto, por ejemplo: un disquete, un disco flexible, un disco duro, una cinta magnética, cualquier otro medio magnético, un CD-ROM, DVD o DVD-ROM, cualquier otro medio óptico, tarjetas perforadas, cinta de papel, cualquier otro medio de almacenamiento físico con patrones de agujeros, una RAM, una ROM, una PROM y EPROM, una FLASH-EPROM, cualquier otro chip o cartucho de memoria, una onda portadora que transporte datos o instrucciones, cables o enlaces que transporten dicha onda portadora, o cualquier otro medio a partir del cual un ordenador pueda leer código de programación y/o datos. Muchas de estas formas de medios legibles por ordenador pueden estar implicados en el transporte de una o más secuencias de una o más instrucciones a un procesador para su ejecución.

En algunas realizaciones, algunas o todas las funcionalidades de análisis del sistema informático 700 pueden empaquetarse en un único paquete de software. En algunas realizaciones, el conjunto completo de capacidades de análisis de datos puede comprender una suite de paquetes de software. En algunas realizaciones, el software de análisis de datos puede ser un paquete independiente que se pone a disposición de los usuarios independientemente de un sistema de instrumento de ensayo. En algunas realizaciones, el software puede estar basado en la web, y puede permitir a los usuarios compartir datos. En algunas realizaciones, puede usarse software disponible comercialmente para realizar la totalidad o una parte del análisis de datos, por ejemplo, puede usarse el software Seven Bridges (https://www.sbgenomics.com/) para compilar tablas del número de copias de uno o más genes que se producen en cada célula para toda la colección de células.

Los métodos y sistemas de la presente divulgación pueden implementarse mediante uno o más algoritmos o métodos. Un método puede implementarse mediante software tras ser ejecutado por la unidad central de procesamiento 705. Las aplicaciones ejemplares de algoritmos o métodos implementados mediante software incluyen métodos bioinformáticos para el procesamiento de lecturas de secuencias (por ejemplo, fusión, filtrado, recorte, agrupación), alineación y llamada, y procesamiento de datos de cadenas y datos de densidad óptica (por ejemplo, determinaciones de número más probable y abundancia cultivable).

En una realización ejemplar, el sistema informático 700 puede realizar análisis de datos en los conjuntos de datos de secuencias generados al realizar ensayos de códigos de barras estocásticos de una única célula. Los ejemplos de funcionalidad de análisis de datos incluyen, pero no se limitan a, (i) algoritmos para decodificar/demultiplexar el marcador de muestra, el marcador celular, el marcador espacial y el marcador molecular, y los datos de secuencia de dianas proporcionados por la secuenciación de la biblioteca de códigos de barras estocástica 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, basándose en los datos, y creando tablas de resumen, (iii) análisis estadístico de los datos de secuencia, por ejemplo, para la agrupación de células por datos de expresión génica, o para predecir intervalos de confianza para determinaciones del número de moléculas de transcripción por

gen y por célula, etc., (iv) algoritmos para identificar subpoblaciones de células raras, por ejemplo, usando análisis de componentes principales, agrupación jerárquica, agrupación de medias k, mapas autoorganizativos, redes neuronales, etc., (v) capacidades de alineación de secuencias para alinear datos de secuencias de genes con secuencias de referencia conocidas y detectar mutaciones, marcadores polimórficos y variantes de corte y empalme, y (vi) agrupación automatizada de marcadores moleculares para compensar errores de amplificación o secuenciación. En algunas realizaciones, el sistema informático 700 puede mostrar 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 se producen en cada célula de una colección de células. En algunas realizaciones, el sistema informático 700 puede ejecutar algoritmos para extraer significado biológico de los resultados de la secuenciación, por ejemplo, correlacionando el número de copias de uno o más genes que se producen en cada célula de una colección de células con un tipo de célula, un tipo de célula rara, o una célula derivada de un sujeto que tiene una enfermedad o afección específica. En algunas realizaciones, el sistema informático 700 puede ejecutar algoritmos para comparar poblaciones de células a través de diferentes muestras biológicas.

EJEMPLOS

Ejemplo 1

Agrupación por división recursiva de dendrogramas y comprobación seguido de fusión

Este ejemplo describe un método de agrupación por división recursiva (por ejemplo, división recursiva de dendrogramas) y comprobación seguido de fusión.

Notas

En el método ilustrado en este ejemplo, durante el paso de división del dendrograma, las divisiones se consideran (por ejemplo, por defecto) biológicamente relevantes si el algoritmo puede encontrar por lo menos un gen que haya alcanzado un valor p lo suficientemente bajo (o un -log10(valor p) lo suficientemente alto). En otras palabras, en algunas realizaciones el único hiperparámetro a ajustar es el parámetro de umbral de puntuación. Un umbral de puntuación más alto (por ejemplo, 100) corresponde a un valor p más bajo (10e-100), lo que significa que debe encontrarse un gen más significativo para que la división se considere válida. Los umbrales de puntuación más altos dan como resultado un número de grupos más pequeño.

Si se generan demasiados grupos después del paso de división, entonces el usuario puede intentar aumentar el umbral de puntuación. Si se generan demasiado pocos grupos en el paso de división, entonces el usuario puede intentar disminuir el umbral de puntuación. Pueden probarse múltiples umbrales de puntuación en la misma matriz de distancias. Precalculando la matriz de distancias, puede ahorrarse mucho tiempo de cálculo.

Si el barrido a través de diferentes umbrales de puntuación sigue generando resultados sin sentido, entonces el problema puede residir en el dendrograma generado en primer lugar (es decir, la matriz de distancias). Como se muestra en la célula [3], el primer paso del algoritmo requiere pasar de la matriz de recuentos de moléculas a una matriz de distancias (el paso de preprocesamiento). Puede ser deseable probar un tipo de preprocesamiento diferente. Un usuario puede quizás probar otra métrica de distancia, intentar no tomar el registro, o prefiltrar células y/o genes que puedan generar una métrica de distancia más precisa para su aplicación.

Si el paso de división genera muchos grupos pequeños que parecen irrelevantes, puede disminuirse el parámetro de percentil de disolución. Este parámetro decide si se mantiene o no un grupo final en función de cuántas de sus distancias por pares se encuentran dentro del percentil de disolución inferior de las distancias por pares totales. Ejecutar el algoritmo con un percentil de disolución de 20, por ejemplo, sólo mantendrá un grupo si por lo menos una distancia por pares se encuentra en el 20% inferior de las distancias totales.

Para determinar por qué se separa un grupo en dos grupos, identificar esos dos grupos y realizar un análisis de prueba t por pares. Esto puede hacerse en la célula [13] del Ejemplo 2 para cada par de grupos. Esta función mostrará los marcadores que distinguieron los dos grupos. Comprobar también la función Explorar de cómo se decidieron las divisiones para ver el paso exacto del algoritmo de división que dio como resultado la división.

Para determinar si se están fusionando grupos incorrectos, disminuir el parámetro de umbral de puntuación en el paso de fusión. Cuanto mayor sea el umbral de puntuación aquí, más probable será que se fusionen dos grupos diferentes. Comprobar también la función Explorar de cómo se decidieron las fusiones para ver el paso exacto del algoritmo de fusión que dio como resultado la fusión.

Para identificar más valores atípicos, probar a reducir el valor atípico_umbral_percentil_parámetro en el paso de fusión.

Dependencias

Los módulos tenían las siguientes dependencias:-numpy (1.10.4)-scipy (0.17.0)-matplotlib (1.5.1)-sklearn (0.17.1)-networkx (1.11) - community - rpy2 (2.8.2)

networkx, community y rpy2 no son necesarios por defecto. networkx y community se usan para la detección de comunidades. networkx también se usa para la igualación de la ponderación maximizadao (como métrica de lo cerca que están dos conjuntos de marcadores). rpy2 se usa para ejecutar sigclust, una prueba estadística para saber si dos poblaciones deberían ser realmente una población. Para ejecutar sigclust, es posible que el usuario necesite tener instalado R junto con el paquete sigclust.

En [1]: # Cargar módulos y librerías relevantes

```
%load_ext autoreload
%autoreload 2
%matplotlib inline
from dendrosplit import split,merge import pickle
import numpy as np
import matplotlib.pyplot as plt np.set_printoptions(precision=2,suppress=True)
```

Funcionamiento de los canales

La entrada a los canales es una matriz de N por M de recuentos moleculares (números naturales) denominada 'X'. genes' es una lista de longitud M de nombres de genes. 'x1' y 'x2' representan la incorporación bidimensional de los datos usando cualquier método elegido por el usuario. 'x1' y 'x2' se usan únicamente para visualizar los resultados de los canales junto con los pasos intermedios. El algoritmo requiere que se eliminen todas las columnas de 'X' que sumen 0, y esta celda de código se encarga de ello.

En [2]: # Cargar datos

```
dataset = 'Resolve4'
pickledir = '/Users/user1/Desktop/datasets/'
x1,x2 = pickle.load(file(pickledir+dataset+'tsne.pickle'))
# Remove columns of X that sum to 0
X,genes = split.filter_genes(X,genes)
Mantener 19307 genes para tener > 0 recuentos en todas las celdas
```

En primer lugar, puede generarse una matriz de distancias a partir de la matriz de recuentos. La celda siguiente lo consigue calculando las distancias de correlación por pares entre muestras transformadas logarítmicamente (log(X+1)). La parte de división del algoritmo sólo requiere la matriz de recuentos, aunque el usuario también puede introducir una matriz de distancias, como se muestra a continuación. Esta parte del algoritmo devolvía un conjunto de marcadores (cadenas) de longitud N para las muestras junto con el 'historial', una estructura de datos que rastreaba toda la información intermedia generada por el algoritmo. El 'historial' era útil para funciones posteriores usadas para diseccionar cómo generaba el algoritmo tales marcadors (y qué características eran las más importantes para generar tales marcadores). Los marcadores eran cadenas que indicaban dónde se encontraba un grupo de acuerdo con el dendrograma generado usando la matriz de distancias. Por ejemplo, "rLLR" significa que este punto pertenece al subárbol derecho del subárbol izquierdo del subárbol izquierdo de la raíz.

En [3]: # Obtener el primer conjunto de marcadores. Es muy recomendable calcular la matriz de distancias fuera del algoritmo

```
D = split.log_correlation (X)
ys,shistory = split.dendrosplit((D,X), preprocessing='precomputed', score_threshold=10,
verbose=True, disband_percentile=50)
    Resultado de división potencial: 883 y 3
    dendrosplit/feature selection.py:106: RuntimeWarning: divide by zero encountered in log10
    gene_scores = np.nan_to_num(-np.log10(p[keep_inds]))
    Puntuación de división 1.8E+308
    Resultado de división potencial: 1 y 882
    Resultado de división potencial: 484 y 398
    /Users/user1/anaconda2/lib/python2.7/site-packages/scipy/stats/
    _distn_infrastructure.py:1748: Runtime
    cond1 = (scale > 0) & (x > self.a) & (x < self.b)
/Users/user1/anaconda2/lib/python2.7/site-packages/scipy/stats/
_distn_infrastructure.py:1748: Runtime
cond1 = (scale > 0) & (x > self.a) & (x < self.b)
```

/Users/user1/anaconda2/lib/python2.7/site-packages/scipy/stats/
_distn_infrastructure.py:1749: Runtime
cond2 = cond0 & (x <= self.a)
Puntuación de división 182.26
Resultado de división potencial: 481 y 3 Puntuación de división 1.8E+308
Resultado de división potencial: 1 y 480
Resultado de división potencial: 1 y 479
Resultado de división potencial: 195 y 284
Puntuación de división 125.49
Resultado de división potencial: 177 y 18
Puntuación de división 15.35
Resultado de división potencial: 1 y 176
Resultado de división potencial: 1 y 175
Resultado de división potencial: 1 y 174
Resultado de división potencial: 12 y 162
Puntuación de división 18.88
Resultado de división potencial: 1 y 11
Resultado de división potencial: 1 y 10
Resultado de división potencial: 2 y 8
Puntuación de división 6.11
Resultado de división potencial: 1 y 161
Resultado de división potencial: 1 y 160
Resultado de división potencial: 28 y 132
Puntuación de división 12.32
Resultado de división potencial: 25 y 3
Puntuación de división 13.94
Resultado de división potencial: 1 y 24
Resultado de división potencial: 11 y 13
Puntuación de división 4.77
Resultado de división potencial: 1 y 2
resultado de división: 1 y 1
Disolución (puntos en el grupo demasiado alejados entre sí) Resultado de división potencial: 122 y 10
Puntuación de división 18.52
Resultado de división potencial: 13 y 109
Puntuación de división 24.92
Resultado de división potencial: 6 y 7
Puntuación de división 3.77
Resultado de división potencial: 105 y 4
Puntuación de división 31.72
Resultado de división potencial: 15 y 90
Puntuación de división 11.31
Resultado de división potencial: 3 y 12
Puntuación de división 6.55
Resultado de división potencial: 17 y 73
Puntuación de división 8.91
Resultado de división potencial: 2 y 2
Puntuación de división 1.58
Disolución (puntos en el grupo demasiado alejados entre sí) Resultado de división potencial: 2 y 8
Puntuación de división 5.79
Resultado de división potencial: 1 y 17
Resultado de división potencial: 1 y 16
Resultado de división potencial: 1 y 15
Resultado de división potencial: 4 y 11
Puntuación de división 4.57
Resultado de división potencial: 1 y 283
Resultado de división potencial: 1 y 282
Resultado de división potencial: 1 y 281
Resultado de división potencial: 271 y 10
Puntuación de división 38.04
Resultado de división potencial: 2 y 269
Puntuación de división 233.23
Resultado de división potencial: 1 y 1
Disolución (puntos en el grupo demasiado alejados entre sí) Resultado de división potencial: 1 y 268
Resultado de división potencial: 265 y 3
Puntuación de división 80.24

Resultado de división potencial: 4 y 261
Puntuación de división 100.26
Resultado de división potencial: 1 y 3
Resultado de división potencial: 1 y 2
Resultado de división potencial: 1 y 1
Disolución (puntos en el grupo demasiado alejados entre sí) Resultado de división potencial: 192 y 69
Puntuación de división 9.66
Resultado de división potencial: 1 y 2
Resultado de división potencial: 1 y 1
Disolución (puntos en el grupo demasiado alejados entre sí) Resultado de división potencial: 2 y 8
Puntuación de división 5.12
Resultado de división potencial: 1 y 2
Resultado de división potencial: 1 y 1
Disolución (puntos en el grupo demasiado alejados entre sí) Resultado de división potencial: 1 y 397
Resultado de división potencial: 1 y 396
Resultado de división potencial: 1 y 395
Resultado de división potencial: 392 y 3
Puntuación de división 228.58
Resultado de división potencial: 1 y 391
Resultado de división potencial: 1 y 390
Resultado de división potencial: 1 y 389
Resultado de división potencial: 1 y 388
Resultado de división potencial: 1 y 387
Resultado de división potencial: 1 y 386
Resultado de división potencial: 32 y 354
Puntuación de división 33.24
Resultado de división potencial: 1 y 31
Resultado de división potencial: 1 y 30
Resultado de división potencial: 21 y 9
Puntuación de división 7.20
Resultado de división potencial: 1 y 353
Resultado de división potencial: 1 y 352
Resultado de división potencial: 1 y 351
Resultado de división potencial: 19 y 332
Puntuación de división 32.86
Resultado de división potencial: 1 y 18
Resultado de división potencial: 3 y 15
Puntuación de división 8.90
Resultado de división potencial: 6 y 326
Puntuación de división 83.57
Resultado de división potencial: 1 y 5
Resultado de división potencial: 1 y 4
Resultado de división potencial: 2 y 2
Puntuación de división 1.32
Disolución (puntos en el grupo demasiado alejados entre sí) Resultado de división potencial: 3 y 323
Puntuación de división 148.25
Resultado de división potencial: 1 y 2
Resultado de división potencial: 1 y 1
Disolución (puntos en el grupo demasiado alejados entre sí) Resultado de división potencial: 314 y 9
Puntuación de división 71.43
Resultado de división potencial: 221 y 93
Puntuación de división 48.70
Resultado de división potencial: 1 y 220
Resultado de división potencial: 1 y 219
Resultado de división potencial: 1 y 218
Resultado de división potencial: 1 y 217
Resultado de división potencial: 215 y 2
Puntuación de división 133.42
Resultado de división potencial: 166 y 49
Puntuación de división 7.64
Resultado de división potencial: 1 y 1
Resultado de división potencial: 40 y 53
Puntuación de división 9.31
Resultado de división potencial: 4 y 5
Puntuación de división 3.20

Disolución (puntos en el grupo demasiado alejados entre sí) Resultado de división potencial: 1 y 2
Resultado de división potencial: 1 y 1
Disolución (puntos en el grupo demasiado alejados entre sí) Resultado de división potencial: 1 y 2
Resultado de división potencial: 1 y 1
Disolución (puntos en el grupo demasiado alejados entre sí)
Nº de veces que se llamó a la función de puntuación: 40
El tiempo de cálculo total fue de 9.532 s

El paso de fusión implicó realizar la comparación por pares todos los grupos generados por el procedimiento de división anterior. Los grupos que no eran suficientemente diferentes se fusionaban empezando por los dos grupos más similares. Al igual que la división, el paso de fusión devuelve tanto los marcadores (longitud-N) como un historial de pasos intermedios. Los marcadores son números enteros. Los valores atípicos se marcan como '-1'. A continuación se presenta un enfoque alternativo al paso de fusión basado en la detección de comunidades.

En [4]: # Fusionar marcadores de grupo

```
ym,mhistory=merge.dendromerge((D,X),ys,score_threshold=10,
preprocessing,'precomputed',verbose=True,outlier_threshold_percentile=90)
```

0 de 886 muestras son singletones

El umbral de valores atípicos es 0,51

vecino más cercano de 821: 72 en grupo 76 (D = 0.375)
vecino más cercano de 661: 29 en grupo 76 (D = 0.379)
vecino más cercano de 729: 281 en grupo 76 (D = 0.381)
vecino más cercano de 559: 79 en grupo 76 (D = 0.381)
vecino más cercano de 690: 171 en grupo 76 (D = 0.381)
vecino más cercano de 564: 79 en grupo 76 (D = 0.381)
vecino más cercano de 776: 474 en grupo 38 (D = 0.387)
vecino más cercano de 860: 340 en grupo 38 (D = 0.390)
vecino más cercano de 816: 379 en grupo 78 (D = 0.390)
vecino más cercano de 787: 63 en grupo 38 (D = 0.391)
vecino más cercano de 737: 72 en grupo 76 (D = 0.392)
vecino más cercano de 874: 220 en grupo 76 (D = 0.392)
vecino más cercano de 743: 72 en grupo 76 (D = 0.394)
vecino más cercano de 877: 174 en grupo 76 (D = 0.394)
vecino más cercano de 753: 190 en grupo 76 (D = 0.397)
vecino más cercano de 774: 158 en grupo 38 (D = 0.398)
vecino más cercano de 565: 190 en grupo 76 (D = 0.399)
vecino más cercano de 785: 79 en grupo 76 (D = 0.401)
vecino más cercano de 706: 101 en grupo 18 (D = 0.403)
vecino más cercano de 829: 213 en grupo 38 (D = 0.404)
vecino más cercano de 701: 179 en grupo 76 (D = 0.404)
vecino más cercano de 770: 453 en grupo 38 (D = 0.404)
vecino más cercano de 630: 79 en grupo 76 (D = 0.406)
vecino más cercano de 866: 87 en grupo 38 (D = 0.407)
vecino más cercano de 795: 159 en grupo 76 (D = 0.407)
vecino más cercano de 865: 179 en grupo 76 (D = 0.407)
vecino más cercano de 869: 101 en grupo 18 (D = 0.409)
vecino más cercano de 830: 165 en grupo 38 (D = 0.412)
vecino más cercano de 851: 29 en grupo 76 (D = 0.412)
vecino más cercano de 782: 50 en grupo 76 (D = 0.412)
vecino más cercano de 627: 72 en grupo 76 (D = 0.412)
vecino más cercano de 848: 83 en grupo 76 (D = 0.413)
vecino más cercano de 883: 687 en grupo 12 (D = 0.413)
vecino más cercano de 793: 107 en grupo 76 (D = 0.414)
vecino más cercano de 631: 101 en grupo 18 (D = 0.416)
vecino más cercano de 720: 101 en grupo 18 (D = 0.418)
vecino más cercano de 885: 101 en grupo 18 (D = 0.418)
vecino más cercano de 813: 101 en grupo 18 (D = 0.419)
vecino más cercano de 788: 278 en grupo 38 (D = 0.420)
vecino más cercano de 748: 101 en grupo 18 (D = 0.422)
vecino más cercano de 762: 158 en grupo 38 (D = 0.423)
vecino más cercano de 804: 177 en grupo 18 (D = 0.425)

vecino más cercano de 854: 101 en grupo 18 (D = 0.426)
vecino más cercano de 605: 159 en grupo 76 (D = 0.437)
vecino más cercano de 849: 101 en grupo 18 (D = 0.437)
vecino más cercano de 835: 101 en grupo 18 (D = 0.438)
vecino más cercano de 790: 32 en grupo 76 (D = 0.442)
vecino más cercano de 744: 188 en grupo 38 (D = 0.448)
vecino más cercano de 822: 282 en grupo 38 (D = 0.449)
vecino más cercano de 723: 170 en grupo 76 (D = 0.456)
vecino más cercano de 884: 101 en grupo 18 (D = 0.459)
vecino más cercano de 563: 34 en grupo 76 (D = 0.463)
vecino más cercano de 867: 160 en grupo 18 (D = 0.463)
vecino más cercano de 771: 34 en grupo 76 (D = 0.473)
vecino más cercano de 826: 165 en grupo 38 (D = 0.475)
vecino más cercano de 777: 174 en grupo 76 (D = 0.478)
vecino más cercano de 759: 101 en grupo 18 (D = 0.483)
vecino más cercano de 855: 101 en grupo 18 (D = 0.485)
vecino más cercano de 702: 160 en grupo 18 (D = 0.492)
vecino más cercano de 750: 230 en grupo 76 (D = 0.495)
vecino más cercano de 704: 216 en grupo 78 (D = 0.497)
vecino más cercano de 711: 55 en grupo 76 (D = 0.502)
vecino más cercano de 708: 537 en grupo 78 (D = 0.510)
vecino más cercano de 791: 115 en grupo 76 (D = 0.534)
vecino más cercano de 722: 15 en grupo 76 (D = 0.547)
vecino más cercano de 700: 107 en grupo 76 (D = 0.549)
vecino más cercano de 846: 72 en grupo 76 (D = 0.552)
vecino más cercano de 876: 85 en grupo 76 (D = 0.560)
vecino más cercano de 868: 740 en grupo 78 (D = 0.562)
vecino más cercano de 569: 68 en grupo 76 (D = 0.572)
vecino más cercano de 817: 56 en grupo 76 (D = 0.582)
vecino más cercano de 798: 310 en grupo 38 (D = 0.585)
vecino más cercano de 717: 216 en grupo 78 (D = 0.597)
vecino más cercano de 879: 209 en grupo 76 (D = 0.612)
vecino más cercano de 727: 96 en grupo 76 (D = 0.616)
vecino más cercano de 828: 142 en grupo 38 (D = 0.618)
vecino más cercano de 840: 632 en grupo 78 (D = 0.640)
vecino más cercano de 747: 202 en grupo 76 (D = 0.698)
vecino más cercano de 842: 797 en grupo 38 (D = 0.703)
vecino más cercano de 442: 336 en grupo 78 (D = 0.735)

Número total de valores atípicos: 18

Singletones asignados (0,052 s)
Dc generado (13,181 s)
Antes de la fusión: 14 grupos
Marcadores de fusión 0 (N = 10) y 6 (N = 15) con distancia 3.60
Antes de la fusión: 13 grupos
Marcadores de fusión 2 (N = 15) y 4 (N = 10) con distancia 4.31 Antes de la fusión: 12 grupos
Marcadores de fusión 1 (N = 13) y 11 (N = 25) con distancia 4.37 Antes de la fusión: 11 grupos
Marcadores de fusión 0 (N = 25) y 10 (N = 38) con distancia 5.23 Antes de la fusión: 10 grupos
Marcadores de fusión 3 (N = 30) y 7 (N = 95) con distancia 6.04 Antes de la fusión: 9 grupos
Marcadores de fusión 2 (N = 10) y 5 (N = 2) con distancia 6.81 Antes de la fusión: 8 grupos
Marcadores de fusión 4 (N = 25) y 5 (N = 63) con distancia 7.19 Antes de la fusión: 7 grupos
Marcadores de fusión 2 (N = 18) y 5 (N = 12) con distancia 7.23 Antes de la fusión: 6 grupos
Marcadores de fusión 3 (N = 125) y 5 (N = 30) con distancia 9.76 la fusión de grupos llevó 25.977 s

La fusión basada en la detección de comunidades usó los módulos de python networkx y community. La estructura de datos del historial que se devuelve aquí solo contiene los marcadores de entrada y los marcadores posteriores al procesamiento de singletones.

En [5]: ym_community=merge.dendromerge((D,X),ys, preprocessing='precomputed', verbose=True,outlier_threshold_percentile=90, perform_community_detection=True)

80 de las 886 muestras son singletones. El umbral de valores atípicos es 0,51

vecino más cercano de 821: 72 en grupo 76 (D = 0.375)

vecino más cercano de 661: 29 en grupo 76 (D = 0.379)
vecino más cercano de 729: 281 en grupo 76 (D = 0.381)
vecino más cercano de 559: 79 en grupo 76 (D = 0.381)
vecino más cercano de 690: 171 en grupo 76 (D = 0.381)
vecino más cercano de 564: 79 en grupo 76 (D = 0.381)
vecino más cercano de 776: 474 en grupo 38 (D = 0.387)
vecino más cercano de 860: 340 en grupo 38 (D = 0.390)
vecino más cercano de 816: 379 en grupo 78 (D = 0.390)
vecino más cercano de 787: 63 en grupo 38 (D = 0.391)
vecino más cercano de 737: 72 en grupo 76 (D = 0.392)
vecino más cercano de 874: 220 en grupo 76 (D = 0.392)
vecino más cercano de 743: 72 en grupo 76 (D = 0.394)
vecino más cercano de 877: 174 en grupo 76 (D = 0.394)
vecino más cercano de 753: 190 en grupo 76 (D = 0.397)
vecino más cercano de 774: 158 en grupo 38 (D = 0.398)
vecino más cercano de 565: 190 en grupo 76 (D = 0.399)
vecino más cercano de 785: 79 en grupo 76 (D = 0.401)
vecino más cercano de 706: 101 en grupo 18 (D = 0.403)
vecino más cercano de 829: 213 en grupo 38 (D = 0.404)
vecino más cercano de 701: 179 en grupo 76 (D = 0.404)
vecino más cercano de 770: 453 en grupo 38 (D = 0.404)
vecino más cercano de 630: 79 en grupo 76 (D = 0.406)
vecino más cercano de 866: 87 en grupo 38 (D = 0.407)
vecino más cercano de 795: 159 en grupo 76 (D = 0.407)
vecino más cercano de 865: 179 en grupo 76 (D = 0.407)
vecino más cercano de 869: 101 en grupo 18 (D = 0.409)
vecino más cercano de 830: 165 en grupo 38 (D = 0.412)
vecino más cercano de 851: 29 en grupo 76 (D = 0.412)
vecino más cercano de 782: 50 en grupo 76 (D = 0.412)
vecino más cercano de 627: 72 en grupo 76 (D = 0.412)
vecino más cercano de 848: 83 en grupo 76 (D = 0.413)
vecino más cercano de 883: 687 en grupo 12 (D = 0.413)
vecino más cercano de 793: 107 en grupo 76 (D = 0.414)
vecino más cercano de 631: 101 en grupo 18 (D = 0.416)
vecino más cercano de 720: 101 en grupo 18 (D = 0.418)
vecino más cercano de 885: 101 en grupo 18 (D = 0.418)
vecino más cercano de 813: 101 en grupo 18 (D = 0.419)
vecino más cercano de 788: 278 en grupo 38 (D = 0.420)
vecino más cercano de 748: 101 en grupo 18 (D = 0.422)
vecino más cercano de 762: 158 en grupo 38 (D = 0.423)
vecino más cercano de 804: 177 en grupo 18 (D = 0.425)
vecino más cercano de 854: 101 en grupo 18 (D = 0.426)
vecino más cercano de 605: 159 en grupo 76 (D = 0.437)
vecino más cercano de 849: 101 en grupo 18 (D = 0.437)
vecino más cercano de 835: 101 en grupo 18 (D = 0.438)
vecino más cercano de 790: 32 en grupo 76 (D = 0.442)
vecino más cercano de 744: 188 en grupo 38 (D = 0.448)
vecino más cercano de 822: 282 en grupo 38 (D = 0.449)
vecino más cercano de 723: 170 en grupo 76 (D = 0.456)
vecino más cercano de 884: 101 en grupo 18 (D = 0.459)
vecino más cercano de 563: 34 en grupo 76 (D = 0.463)
vecino más cercano de 867: 160 en grupo 18 (D = 0.463)
vecino más cercano de 771: 34 en grupo 76 (D = 0.473)
vecino más cercano de 826: 165 en grupo 38 (D = 0.475)
vecino más cercano de 777: 174 en grupo 76 (D = 0.478)
vecino más cercano de 759: 101 en grupo 18 (D = 0.483)
vecino más cercano de 855: 101 en grupo 18 (D = 0.485)
vecino más cercano de 702: 160 en grupo 18 (D = 0.492)
vecino más cercano de 750: 230 en grupo 76 (D = 0.495)
vecino más cercano de 704: 216 en grupo 78 (D = 0.497)
vecino más cercano de 711: 55 en grupo 76 (D = 0.502)
vecino más cercano de 708: 537 en grupo 78 (D = 0.510)
vecino más cercano de 791: 115 en grupo 76 (D = 0.534)
vecino más cercano de 722: 15 en grupo 76 (D = 0.547)
vecino más cercano de 700: 107 en grupo 76 (D = 0.549)

vecino más cercano de 846: 72 en grupo 76 (D = 0.552)
vecino más cercano de 876: 85 en grupo 76 (D = 0.560)
vecino más cercano de 868: 740 en grupo 78 (D = 0.562)
vecino más cercano de 569: 68 en grupo 76 (D = 0.572)
vecino más cercano de 817: 56 en grupo 76 (D = 0.582)
vecino más cercano de 798: 310 en grupo 38 (D = 0.585)
vecino más cercano de 717: 216 en grupo 78 (D = 0.597)
vecino más cercano de 879: 209 en grupo 76 (D = 0.612)
vecino más cercano de 727: 96 en grupo 76 (D = 0.616)
vecino más cercano de 828: 142 en grupo 38 (D = 0.618)
vecino más cercano de 840: 632 en grupo 78 (D = 0.640)
vecino más cercano de 747: 202 en grupo 76 (D = 0.698)
vecino más cercano de 842: 797 en grupo 38 (D = 0.703)
vecino más cercano de 442: 336 en grupo 78 (D = 0.735)

Número total de valores atípicos: 18

Singletones asignados (0,054 s)
Dc generado (12,773 s)
Gráfico construido con 14 nodos y 24 aristas (12,774 s)
La fusión de grupos llevó 12,775 s

En conjunto, estos datos demuestran la agrupación por división recursiva y comprobación seguido de fusión.

Ejemplo 2

Visualización de los resultados de la agrupación por división recursiva de dendrogramas y comprobación seguido de fusión

Este ejemplo describe la visualización de los resultados de la agrupación por división recursiva de dendrogramas y comprobación seguido de la fusión ilustrada en el Ejemplo 1.

Se examinaron los marcadores de los grupos generados tras los pasos de división y fusión.

En [6]: plt.scatter(x1,x2,edgecolors='none')

_=plt.axis('off')
plt.title('Pre-clustering')
 # Clustering results using pre-merged labels (label singletons)
plt.figure()
split.plot_labels_legend(x1,x2,split.str_labels_to_ints(ys))
plt.title('After splitting step')
 # Clustering results using post-merged labels
plt.figure()
split.plot_labels_legend(x1,x2,ym)
plt.title('After merging step')
 # Clustering results using post-merged labels
plt.figure()
split.plot_labels_legend(x1,x2,ym_community)
plt.title('After merging step using community detection')
Out[6]: <matplotlib.text.Text at 0x112674510>

La FIG. 8, paneles (a)-(d) muestran gráficos ejemplares no limitativos de perfiles de expresión en un espacio bidimensional después de dividir y fusionar los perfiles de expresión de células individuales.

Análisis de cómo se decidieron las escisiones

Puede usarse la función "print_history" para analizar cómo el paso de división del método maneja el conjunto de datos usando la función "print_history". La línea i describe la i-ésima división válida. Una división se considera válida si ambos grupos generados están por encima del "min_clust_size" y el valor p más bajo generado a partir de la división está por debajo del umbral.

En [7]: split.print_history(genes,shistory)

Predivisión: 886 L: 883 R: 3 Puntuación: 1.8E+308 Gen principal: RPL31 Gen principal Puntuación: 1.8E+308

Predivisión: 882 L: 484 R: 398 Puntuación: 182.26 Gen principal: FTL Puntuación gen principal: 182.26
Predivisión: 484 L: 481 R: 3 Puntuación: 1.8E+308 Gen principal: RPL23 Puntuación gen principal: 1.8E+308
Predivisión: 479 L: 195 R: 284 Puntuación: 125.49 Gen principal: IGHM Puntuación gen principal: 125.49
Predivisión: 195 L: 177 R: 18 Puntuación: 15.35 Gen principal: RRP7A Puntuación gen principal: 15.35
Predivisión: 174 L: 12 R: 162 Puntuación: 18.88 Gen principal: ANXA11 Puntuación gen principal: 18.88
Predivisión: 160 L: 28 R: 132 Puntuación: 12.32 Gen principal: TTF1 Puntuación gen principal: 12.32
Predivisión: 28 L: 25 R: 3 Puntuación: 13.94 Gen principal: SRPK1 Puntuación gen principal: 13.94
Predivisión: 132 L: 122 R: 10 Puntuación: 18.52 Gen principal: TOP2A Puntuación gen principal: 18.52
Predivisión: 122 L: 13 R: 109 Puntuación: 24.92 Gen principal: CACYBP Puntuación gen principal: 24.92
Predivisión: 109 L: 105 R: 4 Puntuación: 31.72 Gen principal: RPSA Puntuación gen principal: 31.72
Predivisión: 105 L: 15 R: 90 Puntuación: 11.31 Gen principal: PSMD14 Puntuación gen principal: 11.31
Predivisión: 281 L: 271 R: 10 Puntuación: 38.04 Gen principal: RNASEH2B Puntuación gen principal: 38.04
Predivisión: 271 L: 2 R: 269 Puntuación: 233.23 Gen principal: GAS8 Puntuación gen principal: 233.23
Predivisión: 268 L: 265 R: 3 Puntuación: 80.24 Gen principal: CNPY3 Puntuación gen principal: 80.24
Predivisión: 265 L: 4 R: 261 Puntuación: 100.26 Gen principal: MZB1 Puntuación gen principal: 100.26
Predivisión: 395 L: 392 R: 3 Puntuación: 228.58 Gen principal: CREB3L1 Puntuación gen principal: 228.58
Predivisión: 386 L: 32 R: 354 Puntuación: 33.24 Gen principal: VMP1 Puntuación gen principal: 33.24
Predivisión: 351 L: 19 R: 332 Puntuación: 32.86 Gen principal: EIF2B1 Puntuación gen principal: 32.86
Predivisión: 332 L: 6 R: 326 Puntuación: 83.57 Gen principal: NUDT5 Puntuación gen principal: 83.57
Predivisión: 326 L: 3 R: 323 Puntuación: 148.25 Gen principal: TMSB4X Puntuación gen principal: 148.25
Predivisión: 323 L: 314 R: 9 Puntuación: 71.43 Gen principal: C12orf57 Puntuación gen principal: 71.43
Predivisión: 314 L: 221 R: 93 Puntuación: 48.70 Gen principal: RPL27A Puntuación gen principal: 48.70
Predivisión: 217 L: 215 R: 2 Puntuación: 133.42 Gen principal: JUN Puntuación gen principal: 133.42

Se visualizaron los puntos implicados en cada división. Cada fila tiene 2 cifras. La fila i describe la i-ésima división guardada. La FIG. 9, paneles (a)-(x) son gráficos ejemplares no limitativos de perfiles de expresión en un espacio bidimensional que muestran cómo se decidieron las divisiones. En cada panel, la cifra izquierda indica cómo se llevó a cabo la división. Los puntos azules no participaron en absoluto en la división. Los puntos rojos y verdes solían estar en el mismo grupo y luego se separaron. El título de la cifra de la izquierda indica el número de división y los 3 genes que obtuvieron el mayor estadístico t (tras adoptar un valor absoluto). El número asociado a cada gen es el -log10 del valor p correspondiente. El '0' o '1' entre paréntesis junto a cada gen indica el grupo que tuvo la mayor expresión media de ese gen. La cifra derecha muestra la expresión logarítmica del gen que obtuvo el mayor estadístico t.

En [8]: split.visualize_history(np.log(1+X),x1,x2,genes,shistory)/Users/user1/anaconda2/lib/python2.7/site-packages/matplotlib/pyplot.py:516: RuntimeWarning: More max open warning, RuntimeWarning)

La función "analyze_split" puede usarse para examinar más de cerca los genes que determinaron por qué se mantuvo una división particular. Usar la palabra clave "show_background" para mostrar también las células no implicadas en la división. Usar "dust" para ver sólo los genes más altamente expresados en un grupo particular. Puede usarse "num_genes" para mostrar un número personalizado de genes.

En [9]: # Observar la división 5

```
split_num = 5
cluster_of_interest = None
show_background = False
split.analyze_split(X,x1,x2,genes,shistory, split_num, num_genes=12,
show_background=show_background, clust=cluster_of_ interest)
```

La FIG. 10 muestra un gráfico ejemplar no limitativo de los perfiles de expresión en un espacio bidimensional después del quinto ciclo de división. La FIG. 11, paneles (a)-(l) son trazados ejemplares no limitativos de perfiles de expresión en un espacio bidimensional que muestran por qué se mantuvieron ciertas divisiones en el dendrograma para el quinto ciclo de división mostrado en la FIG. 10.

Análisis de cómo se decidieron las fusiones

La función para analizar cómo se decidieron las divisiones puede usarse para analizar cómo se realizó la fusión.

En [10]: split.print_history(genes,mhistoria)

split.visualize_history(np.log(1+X),x1,x2,genes,mhistory)

80 de las 886 muestras son singletones

Singleton(es) 442, 569, 700, 708, 717, 722, 727, 747, 791, 798, 817, 828, 840, 842, 846, 868, 876, 879 m
Singleton(es) 15 fusionado con grupo 12 (N = 24) para formar grupo 1 (N = 25)
Singleton(es) 3, 4, 6, 7, 9, 10, 11, 13, 14, 19, 20, 21, 22, 24, 25, 26 fusionado con grupo 18 (N = 90)
Singleton(es) 5, 29, 30, 31, 32, 33, 34, 35, 36, 37, 39, 40, 41 fusionado con grupo 38 (N = 261) para formar
Singleton(es) 28, 53, 55, 56, 58, 59, 60, 61, 63, 64, 65, 67, 68, 69, 70, 71, 72, 73, 74, 75, 79, 80, 81
Singleton(es) 54, 66 fusionado con grupo 78 (N = 93) para formar grupo 13 (N = 95)
Post-fusión: 25 L: 10 R: 15 Puntuación: 3.60 Gen principal: ENOSF1 Puntuación gen principal: 3.60
Post-fusión: 25 L: 15 R: 10 Puntuación: 4.31 Gen principal: MAGED1 Puntuación gen principal: 4.31
Post-fusión: 38 L: 13 R: 25 Puntuación: 4.37 Gen principal: PRPF40A Puntuación gen principal: 4.37
Post-fusión: 63 L: 25 R: 38 Puntuación: 5.23 Gen principal: ALDOC Puntuación gen principal: 5.23
Post-fusión: 125 L: 30 R: 95 Puntuación: 6.04 Gen principal: PARP1 Puntuación gen principal: 6.04
Post-fusión: 12 L: 10 R: 2 Puntuación: 6.81 Gen principal: IGLC3 Puntuación gen principal: 6.81
Post-fusión: 88 L: 25 R: 63 Puntuación: 7.19 Gen principal: HMGB2 Puntuación gen principal: 7.19
Post-fusión: 30 L: 18 R: 12 Puntuación: 7.23 Gen principal: VIM Puntuación gen principal: 7.23
Post-fusión: 155 L: 125 R: 30 Puntuación: 9.76 Gen principal: HMGN5 Puntuación gen principal: 9.76

La FIG. 12, paneles (a)-(i) son gráficos ejemplares no limitativos de perfiles de expresión en un espacio bidimensional que muestran cómo se decidieron las fusiones.

En [11]: # Observar la fusión 2

```
merge_num = 2
cluster_of_interest = None
show_background = False
split.analyze_split(X,x1,x2,genes,mhistory, merge_num, num_genes=4,
show_background=show_background, clust=cluster_of_ interest)
```

La FIG. 13 muestra un gráfico ejemplar no limitativo de perfiles de expresión en un espacio bidimensional después del segundo ciclo de fusiones. La FIG. 14, paneles (a)-(d) son gráficos ejemplares no limitativos de perfiles de expresión en un espacio bidimensional que muestran cómo se decidió el segundo ciclo de fusiones mostrado en la FIG. 13.

Expresión diferencial

El módulo permite dos tipos de análisis de expresión diferencial simple. El primero hace una comparación uno-v-resto para cada grupo, visualizando los genes más importantes para cada grupo de acuerdo con una prueba t para cada gen. El segundo hace una comparación por pares para cada dos pares de grupos.

En [12]:

```
split.save_more_highly_expressed_genes_in_one_clust(X,genes,ym,x1,x2,num _genes=3,
show_plots=True)
```

La FIG. 15, paneles (a)-(f) son gráficos que muestran un tipo ejemplar no limitativo de análisis de expresión diferencial.

En [13]: pairwise_cluster_comparison(X,genes,ym,x1=x1,x2=x2, num_genes=3, show_plots=True,verbose=F

```
dendrosplit/utils.py:39: FutureWarning: elementwise comparison failed; returning scalar
instead, but in plt.plot(x1[y==i],x2[y==i],'.',c=RGBs[j],label=str(i)+'
('+str(np.sum(y==i))+')')
dendrosplit/feature selection.py:221: RuntimeWarning: divide by zero encountered in double scalars fold = g mean
j/g mean
```

La FIG. 16, paneles (a)-(o) son gráficos que muestran otro tipo ejemplar no limitativo de análisis de expresión diferencial.

Distribución de distancias

El módulo también permite al usuario visualizar la distribución de distancias dentro de cada grupo. Para un grupo dado, esta función traza la proporción de distancias por pares (entre puntos del grupo) para cada intervalo de percentiles del conjunto global de distancias por pares (para todos los puntos de todos los grupos). Por ejemplo, 0,3 en 1 indica que el 30% de las distancias por pares se sitúan entre los percentiles 5º y 10º de las distancias totales.

Usar esta función para hacerse una idea de la cohesión de los grupos de acuerdo con la matriz de distancias original. Intuitivamente, un buen grupo debería tener puntos cercanos entre sí. Por ejemplo, un grupo sin distancias en los 10 intervalos inferiores (es decir, el percentil 50 inferior) se consideraría malo. Observar que, como era de esperar, éste es el caso del grupo "-1" siguiente, que contiene los valores atípicos.

En [14]: merge.visualize_within_cluster_distance_distributions(D, ym, show_D_dist=True)
La FIG. 17, paneles (a)-(g) son gráficos ejemplares no limitativos que visualizan las distancias entre grupos.

Dendrograma

El módulo también permite a los usuarios generar un dendrograma y obtener el orden de las células de acuerdo con el dendrograma. El dendrograma puede ser difícil de ver en un cuaderno iPython. En algunas realizaciones, el dendrograma puede guardarse, como se muestra a continuación. Los usuarios pueden introducir los marcadores de los grupos (palabra clave "marcadores"). Si se desea, la función puede colorear los nombres de todas las muestras dentro de un grupo del mismo color.

En [15]: cell_order = split.print_dendro(D,return_cell_order=True, labels=ym, save _name= ' /Users/user 1/Desktop/dendrogram ')
dendrosplit/split.py:233: FutureWarning: la comparación con 'Ninguno' dará como resultado una composición de objetos por elementos si marcadores != Ninguno:

La FIG. 18 muestra un dendrograma ejemplar no limitativo.

En conjunto, estos datos demuestran las varias herramientas de la divulgación para visualizar los varios pasos y resultados de la división y fusión recursivas seguido de la fusión.

Ejemplo 3

Barrido de parámetros para la agrupación por división recursiva de dendrogramas y comprobación seguido de fusión

Este ejemplo describe el barrido de parámetros para optimizar los parámetros para la división recursiva y la comprobación seguido de la fusión.

Durante el paso de división del método, pueden ajustarse dos hiperparámetros: el umbral de puntuación y el percentil de disolución. Pueden analizarse qué grupos diferentes pueden generarse con diferentes hiperparámetros. Pueden obtenerse rápidamente varios resultados de agrupación (postdivisión prefusión) explotando el hecho de que los grupos generados con un umbral de puntuación más pequeño (un umbral más pequeño da como resultado más grupos) dividen los grupos generados con un umbral de puntuación más grande. En primer lugar, se ejecuta el paso de división con un umbral muy bajo. En segundo lugar, se usa la función de obtener grupos del historial().

A continuación se muestra un ejemplo de barrido a través de varios valores de umbral de puntuación. Lo mismo puede hacerse con valores de percentil de disolución.

En [16]: ys,shistory=split.dendrosplit((D,X), preprocessing='precomputed',

```
score_threshold=2, verbose=False, disband_percentile=50)
ys_sweep = []
thresholds = range(5,100,5)
for threshold in thresholds:
ys_sweep.append(split.get_clusters_from_history(D,shistory,threshold,50))
plt.figure()
split.plot_labels_legend(x1,x2,split.str_labels_to_ints(ys_sweep[-1]))
plt.title('Clustering result using a threshold of %.3f'%(threshold))
```

La FIG. 19, paneles (a)-(s) son gráficos ejemplares no limitativos que muestran el barrido de parámetros.

Al barrer los parámetros, puede investigarse cómo cambia el número de grupos en función del umbral. Esto puede dar a los usuarios ideas sobre cómo seleccionar un umbral óptimo para una aplicación particular.

En [17]: def count_nonsingleton_clusters(y):

```
return sum([1 for i in np.unique(y) if np.sum(y==i) != 1])
plt.plot(thresholds,[count_nonsingleton _clusters(i) for i in ys_sweep])
plt.grid()
plt.xlabel('thresholds(-log10(p-value))')
```

plt.ylabel('number of nonsingletone clusters')
Out[17]: <matplotlib.text.Text at 0x117fb3290>

La FIG. 20 es un gráfico ejemplar no limitativo que muestra cómo puede usarse el barrido de parámetros para identificar un umbral. Como se identificó un gran número de grupos de perfiles de expresión con un umbral de solo 5, se identificaron menos grupos de perfiles de expresión con umbrales mayores (por ejemplo, 40 mostrados en la FIG. 19, panel (h)).

En conjunto, estos datos demuestran la optimización de hiperparámetros mediante barrido de parámetros para la división recursiva y la comprobación seguido de fusión.

Ejemplo 4

Agrupación por división recursiva de dendrogramas y comprobación seguido de fusión

Este ejemplo describe un método de agrupación por división recursiva (por ejemplo, división recursiva de dendrogramas) y comprobación seguido de fusión. En cada grupo o nodo del dendrograma (excepto en los nodos hoja), la correlación mediana intragrupo de los dos subgrupos fue mayor que la correlación mediana intergrupo en este ejemplo.

Durante la fase de división y comprobación de los perfiles de expresión de 357 células, partiendo de la parte superior del dendrograma, el árbol se dividió en dos subárboles candidatos. La división correspondía a la división de un grupo en dos subgrupos candidatos, con la condición de que la correlación mediana intragrupo de los dos subgrupos fuera superior a la correlación mediana intergrupo. Se puntúa la calidad de la división. Si se consideraba que los subgrupos eran suficientemente diferentes, la fase continuaba para cada subárbol. En caso contrario, el método terminaba para esta parte del dendrograma. Esta fase produjo un conjunto de marcadores para el conjunto de datos.

La FIG. 21, paneles (a)-(j) son gráficos ejemplares no limitativos que muestran los resultados de la primera división. Durante la primera división, se determinó que veinte genes (mostrados en la Tabla 1) se expresaban de manera diferente en las 357 células.

Tabla 1. Se determinó que 20 genes se expresaban de manera diferente en las 357 células durante la primera división.

División	Gen	Valor p	Grupo más grande
1	IGLC3|ENST00000390325.2|Referencia_final	201.35	0
1	JCHAIN|NM_144646.3|Referencia_final	105.57	0
1	ADA|NM_000022.3|Referencia_final	89.27	1
1	TCL1A|NM_021966.2|Referencia_final	81.19	0
1	CD74|NM_004355.3|Referencia_final	62.65	0
1	CD3D|NM_000732.4|Referencia_final	50.32	1
1	POU2AF1|NM_006235.2|Referencia_final	39.94	0
1	CD52|NM_001803.2|Referencia_final	39.1	0
1	QPCT|NM_012413.3|Referencia_final	38.87	0
1	HLA-DRA|NM_019111.4|Referencia_final	26.64	0
1	CD22|NM_001771.3|Referencia_final	25.96	0
1	IRF8|NM_002163.2|Referencia_final	21.25	0
1	MS4A1|NM_021950.3|PoloA_1	19.99	0
1	CD37|NM_001774.2|Referencia_final	18.49	0
1	LEF1|NM_016269.4|Referencia_final	17.63	1
1	MME|NM_000902.3|Referencia_final	15.59	0
1	BCL6|NM_001706.4|Referencia_final	13.39	0
1	CD27|NM_001242.4|Referencia_final	11.02	0
1	IL32|NM_004221.4|Referencia_final	10.86	1
1	CD38|NM_001775.3|Referencia_final	10.65	0

La FIG. 22 es un gráfico de incorporación de vecinos estocástica distribuidos en t (t-SNE) ejemplar no

limitativo de que ilustra el resultado de la división de los perfiles de expresión de 357 células, mostrando que las 357 células se clasificaron en dos grupos con un umbral de 10. La FIGS. 23 muestra un dendrograma ejemplar no limitativo que muestra perfiles de expresión clasificados en dos grupos basados en las características mostradas en la Tabla 2 (el grupo 0 de la Tabla 1 corresponde al grupo 1 de la Tabla 2, y el grupo 1 de la Tabla 2 corresponde al grupo 2 de la Tabla 2). La FIG. 24 es un gráfico ejemplar no limitativo que muestra el barrido de parámetros. Como se identificaron dos grupos de perfiles de expresión con un umbral de sólo 10, se identificaron los mismos dos grupos de perfiles de expresión con umbrales mayores (Comparar la FIG. 24 con la FIG. 20).

Tabla 2. Características de los grupos por pares de los dos grupos ordenadas por valor p.

Comparación	Gen	Valor p	Grupo más grande	Cambio de veces de la expresión para el grupo más grande
Grupo1 frente a Grupo2	IGLC3|ENST00000390325.2| Referencia_final	201.35	1	183.947
Grupo1 frente a Grupo2	JCHAIN|NM_144646.3| Referencia_final	105.572	1	50.085
Grupo1 frente a Grupo2	ADA|NM_000022.3| Referencia_final	89.274	2	11.82
Grupo1 frente a Grupo2	TCL1A|NM_021966.2| Referencia_final	81.191	1	134.689
Grupo1 frente a Grupo2	CD741NM_004355.3| Referencia_final	62.653	1	173.268
Grupo1 frente a Grupo2	CD3D|NM_000732.4| Referencia_final	50.32	2	109.703
Grupo1 frente a Grupo2	POU2AF1|NM_006235.2| Referencia_final	39.943	1	19.778
Grupo1 frente a Grupo2	CD52|NM_001803.2| Referencia_final	39.105	1	50.988
Grupo1 frente a Grupo2	QPCT|NM_012413.3| Referencia_final	38.87	1	INF
Grupo1 frente a Grupo2	BLA-DRA|NM_019111.4| Referencia_final	26.642	1	71.744
Grupo1 frente a Grupo2	CD22|NM_001771.3| Referencia_final	25.96	1	64.976
Grupo1 frente a Grupo2	IRF8|NM_002163.2| Referencia_final	21.245	1	INF
Grupo1 frente a Grupo2	MS4A1|NM_021950.3| PoliA_1	19.993	1	INF
Grupo1 frente a Grupo2	CD37|NM_001774.2| Referencia_final	18.492	1	INF
Grupo1 frente a Grupo2	LEF1|NM-016269.4| Referencia_final	17.63	2	4.807
Grupo1 frente a Grupo2	MME|NM_000902.3| Referencia_final	15.585	1	37.451
Grupo1 frente a Grupo2	BCL6|NM_001706.4| Referencia_final	13.393	1	INF
Grupo1 frente a Grupo2	CD27|NM_001242.4 | Referencia_final	11.018	1	INF

(continuación)

Comparación	Gen	Valor p	Grupo más grande	Cambio de veces de la expresión para el grupo más grande
Grupo1 frente a Grupo2	IL32|NM_004221.4| Referencia_final	10.862	2	INF
Grupo1 frente a Grupo2	CD38|NM_001775.3| Referencia_final	10.649	1	INF

En conjunto, estos datos demuestran la agrupación por división recursiva y la comprobación seguido de fusión. En este ejemplo, en cada grupo o nodo del dendrograma (excepto en los nodos hoja), la correlación mediana intragrupo de los dos subgrupos fue más alta que la correlación mediana intergrupo.

En por lo menos algunas de las realizaciones descritas anteriormente, uno o más elementos usados en una realización pueden usarse indistintamente en otra realización, a menos que dicha sustitución no sea técnicamente factible.

REIVINDICACIONES

1. Un método implementado por ordenador para identificar dianas para distinguir tipos de células, que comprende:

(a) recibir perfiles de expresión de una pluralidad de células, en donde los perfiles de expresión comprenden un número de cada diana de una pluralidad de dianas para cada célula de la pluralidad de células;
(b) agrupar los perfiles de expresión de la pluralidad de células para generar una pluralidad de grupos de perfiles de expresión sobre la base de las distancias entre los perfiles de expresión de la pluralidad de células, en donde cada grupo tiene una o más asociaciones con uno o ambos de (1) un grupo padre y (2) dos o más grupos hijos, en donde el grupo padre representa perfiles de expresión de una o más células de la pluralidad de células representadas por el grupo, y en donde el grupo representa perfiles de expresión representados por los dos o más grupos hijos; en donde la agrupación de los perfiles de expresión de la pluralidad de células comprende:

agrupar jerárquicamente los perfiles de expresión de la pluralidad de células para generar un dendrograma que represente los perfiles de expresión de la pluralidad de células basándose en las distancias entre los perfiles de expresión de la pluralidad de células, en donde el dendrograma comprende la pluralidad de grupos, en donde la pluralidad de grupos comprende un grupo raíz, una pluralidad de grupos hoja, y una pluralidad de grupos no raíz y no hoja,
en donde cada uno de la pluralidad de grupos hoja y la pluralidad de grupos no raíz y no hoja tiene una asociación con un grupo padre,
en donde cada uno del grupos raíz y la pluralidad de grupos no raíz y no hoja, tiene asociaciones con un grupo hijo izquierdo y un grupo hijo derecho y representa perfiles de expresión representados por el grupo hijo izquierdo y el grupo hijo derecho del grupo, y
en donde el grupo raíz representa los perfiles de expresión de la pluralidad de células;

(c) mientras se atraviesa cada grupo del dendrograma desde el grupo raíz del dendrograma hasta la pluralidad de grupos hoja del dendrograma para cada grupo con dos o más grupos hijos,

determinar si las asociaciones del grupo con los grupos hijos del grupo son válidas o inválidas, en donde se determina que las asociaciones entre el grupo con los dos o más grupos hijos son válidas cuando la distancia entre el grupo hijo izquierdo y el grupo hijo derecho está por encima de un umbral de asociación, y en caso contrario son inválidas,
en donde la distancia entre el grupo hijo izquierdo y el grupo hijo derecho está representada por un valor p derivado de pruebas estadísticas independientes realizadas en cada diana de la pluralidad de dianas entre los perfiles de expresión representados por el grupo hijo izquierdo y el grupo hijo derecho; y

si las asociaciones entre el grupo con los dos o más grupos hijos son inválidas, añadir el grupo a un conjunto de grupos de fusión;
(d) iterativamente, para cada primer grupo en el conjunto de grupos de fusión, si una distancia entre el primer grupo en el conjunto de grupos de fusión y un segundo grupo en el conjunto de grupos de fusión que es el más cercano al primer grupo está dentro de un umbral de distancia de fusión, fusionar el primer grupo y el segundo grupo para generar un grupo fusionado, en donde el grupo fusionado comprende perfiles de expresión del primer grupo y del segundo grupo; y
(e) para cada grupo en el conjunto de grupos de fusión, identificar dianas para distinguir tipos de células basándose en los perfiles de expresión de la pluralidad de dianas de las células representadas por el grupo, en donde la identificación de dianas comprende: determinar una diferencia, entre los perfiles de expresión representados por el grupo y los perfiles de expresión representados por otro grupo del conjunto de grupos de fusión, en números de marcadores moleculares con secuencias distintas asociadas a las dianas para distinguir los tipos de células que es mayor que un umbral de significancia.

2. El método de la reivindicación 1, que comprende, en cada grupo al atravesar la pluralidad de grupos del dendrograma: (3) añadir el grupo al conjunto de grupos de fusión si el grupo representa un perfil de expresión de una única célula.

3. El método de la reivindicación 2, que comprende, en cada grupo al atravesar la pluralidad de grupos del dendrograma: asignar un marcador de grupo al grupo, en donde, si el grupo representa un perfil de expresión de una única célula, el marcador de grupo del grupo comprende una designación de única célula,

de lo contrario, si el grupo es el grupo hijo izquierdo del grupo padre, el marcador de grupo del grupo comprende el marcador de grupo del grupo padre y una designación izquierda, y
de lo contrario, el marcador de grupo del grupo comprende el marcador de grupo del grupo padre y una designación derecha.

4. El método de la reivindicación 1 o 2, que comprende, antes de agrupar los perfiles de expresión de la pluralidad de células para generar la pluralidad de grupos de perfiles de expresión sobre la base de las distancias entre los perfiles

de expresión de la pluralidad de células en (b):

(i) determinar una estructura de datos de distancias de los perfiles de expresión de la pluralidad de células, en donde la estructura de datos de distancias comprende una matriz de distancias de los perfiles de expresión de la pluralidad de células,

en donde la agrupación de los perfiles de expresión de la pluralidad de células para generar la pluralidad de grupos de perfiles de expresión basándose en las distancias entre los perfiles de expresión de la pluralidad de células en (b) comprende: agrupar los perfiles de expresión de la pluralidad de células para generar la pluralidad de grupos de perfiles de expresión sobre la base de la matriz de distancias.

5. El método de la reivindicación 4 que comprende, antes de determinar la estructura de datos de distancias en (i), transformar logarítmicamente la estructura de datos de recuentos de dianas en una estructura de datos de recuentos de dianas transformada logarítmicamente,

en donde la determinación de la estructura de datos de distancias de los elementos de la estructura de datos de recuentos de dianas comprende determinar la estructura de datos de distancias de la estructura de datos de recuentos de dianas transformada logarítmicamente, y
en donde la agrupación de los perfiles de expresión de la pluralidad de células para generar la pluralidad de grupos de perfiles de expresión sobre la base de las distancias entre los perfiles de expresión de la pluralidad de células en (b) comprende: agrupar los perfiles de expresión de la pluralidad de células sobre la base de la estructura de datos de recuentos de dianas transformada logarítmicamente y la estructura de datos de distancias para generar la pluralidad de grupos.

6. El método de cualquiera de las reivindicaciones 1-5, en donde la agrupación de los perfiles de expresión de la pluralidad de células sobre la base de las distancias entre los perfiles de expresión de la pluralidad de células en (b) comprende:

asignar cada perfil de expresión de la pluralidad de células a un grupo de hojas diferente en la pluralidad de grupos; y
combinar iterativamente un primer grupo y un segundo grupo de la pluralidad de grupos para generar un grupo padre del primer grupo y del segundo grupo si el segundo grupo es el grupo de la pluralidad de grupos más cercano al primer grupo.

7. El método de la reivindicación 6, en donde una indicación de una correlación intragrupo del primer grupo y una correlación intragrupo del segundo grupo es mayor que una correlación intergrupo del primer grupo y el segundo grupo.

8. El método de la reivindicación 7, en donde la indicación de la correlación intragrupo del primer grupo y la correlación intragrupo del segundo grupo se basa en por lo menos uno de:

una correlación máxima intragrupo del primer grupo y del segundo grupo,
una correlación media intragrupo del primer grupo y del segundo grupo,
una correlación mediana intragrupo del primer grupo y del segundo grupo,
una correlación mínima intragrupo del primer grupo y del segundo grupo, y
cualquier combinación de los mismos;
en donde la correlación intragrupo del primer grupo se basa en por lo menos uno de:

una correlación máxima intragrupo del primer grupo,
una correlación media intragrupo del primer grupo,
una correlación mediana intragrupo del primer grupo,
una correlación mínima intragrupo del primer grupo, y
cualquier combinación de las mismas;

en donde la correlación intragrupo del segundo grupo se basa en por lo menos una de:

una correlación máxima intragrupo del segundo grupo,
una correlación media intragrupo del segundo grupo,
una correlación mediana intragrupo del segundo grupo,
una correlación mínima intragrupo del segundo grupo, y
cualquier combinación de las mismas; y

en donde la correlación intergrupo del primer grupo y el segundo grupo se basa en por lo menos una de:

una correlación máxima intergrupo del primer grupo y del segundo grupo,

una correlación media intergrupo del primer grupo y del segundo grupo,
una correlación mediana intergrupo del primer grupo y del segundo grupo,
una correlación mínima intergrupo del primer grupo y del segundo grupo, y
cualquier combinación de las mismas.

9. Un sistema informático para determinar el número de dianas que comprende:

un procesador de hardware; y
una memoria no transitoria que tiene instrucciones almacenadas en la misma, que cuando son ejecutadas por el procesador de hardware hacen que el procesador realice el método de cualquiera de las reivindicaciones 1-8.

10. Un medio legible por ordenador que comprende código para realizar el método de cualquiera de las reivindicaciones 1-8.

11. El método de cualquiera de las reivindicaciones 1-8, que comprende:

diseñar un ensayo del transcriptoma completo basado en las dianas para distinguir los tipos de células identificados; o
diseñar un ensayo del transcriptoma dirigido basado en las dianas para distinguir los tipos de células identificados.

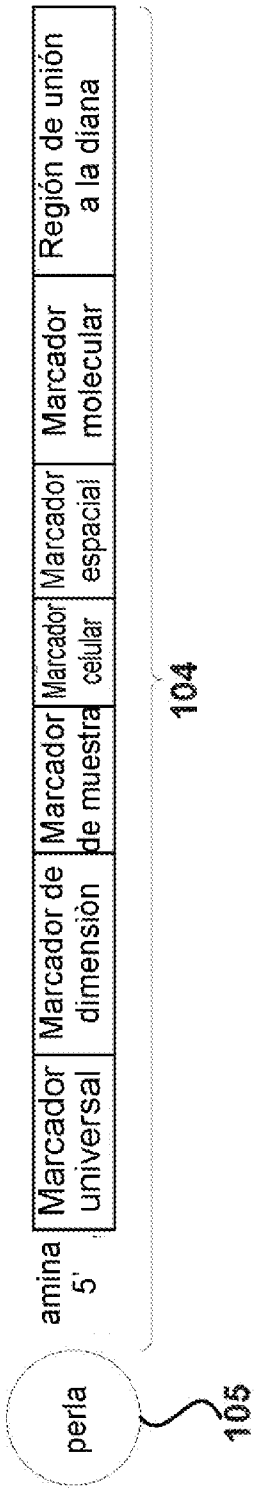

perla
105
amina 5'
Marcador universal
Marcador de dimensión
Marcador de muestra
Marcador celular
Marcador espacial
Marcador molecular
Región de unión a la diana
104

FIG. 1

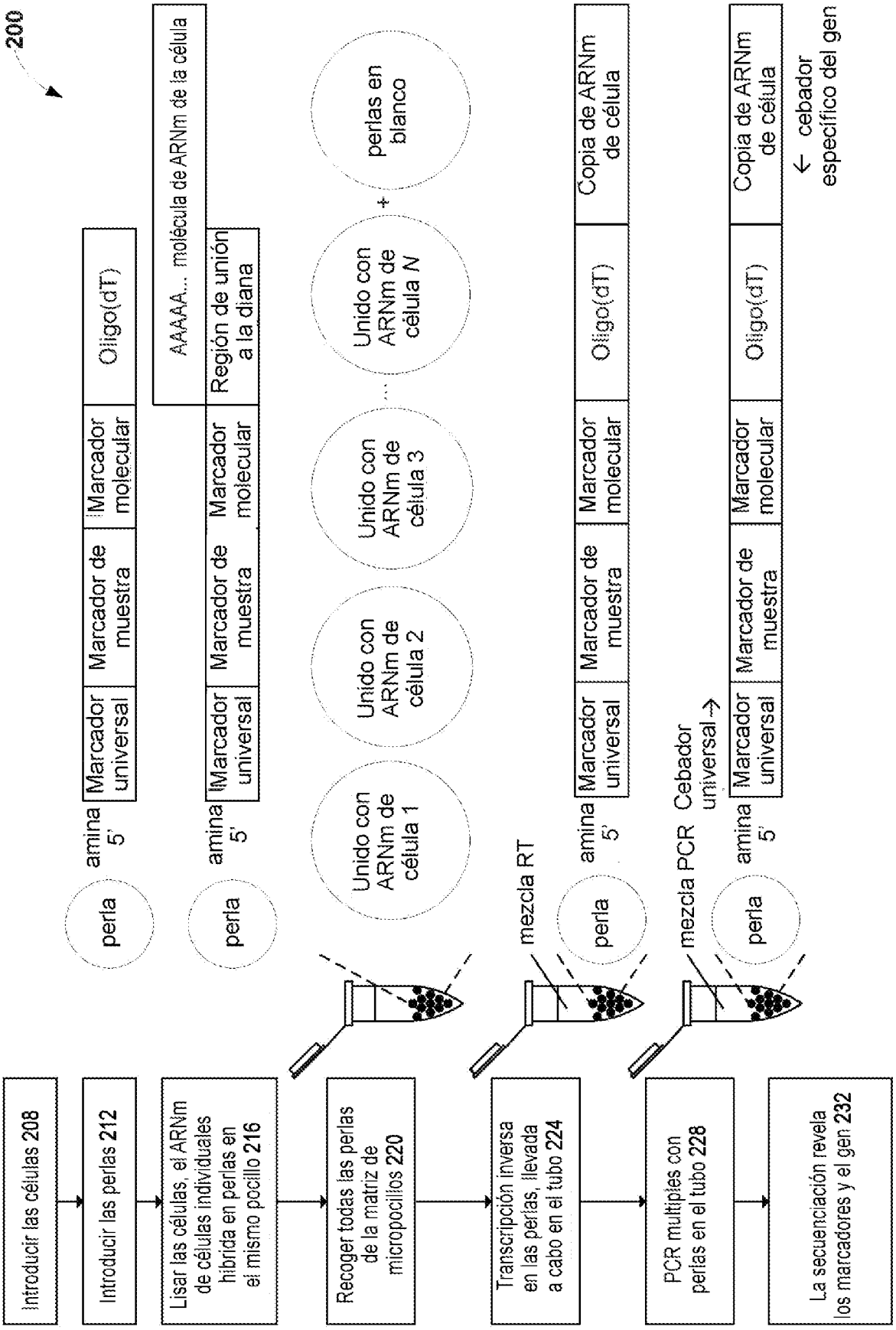

200
Introducir las células 208
Introducir las perlas 212
Lisar las células, el ARNm de células individuales hibrida en perlas en el mismo pocillo 216
Recoger todas las perlas de la matriz de micropocillos 220
Transcripción inversa en las perlas, llevada a cabo en el tubo 224
PCR multiples con perlas en el tubo 228
La secuenciación revela los marcadores y el gen 232
perla
amina 5'
Marcador universal
Marcador de muestra
Marcador molecular
Oligo(dT)
AAAAA... molécula de ARNm de la célula
Región de unión a la diana
Unido con ARNm de célula 1
Unido con ARNm de célula 2
Unido con ARNm de célula 3
Unido con ARNm de célula N
perlas en blanco
mezcla RT
mezcla PCR
Cebador universal →
Copia de ARNm de célula
← cebador específico del gen

FIG. 2

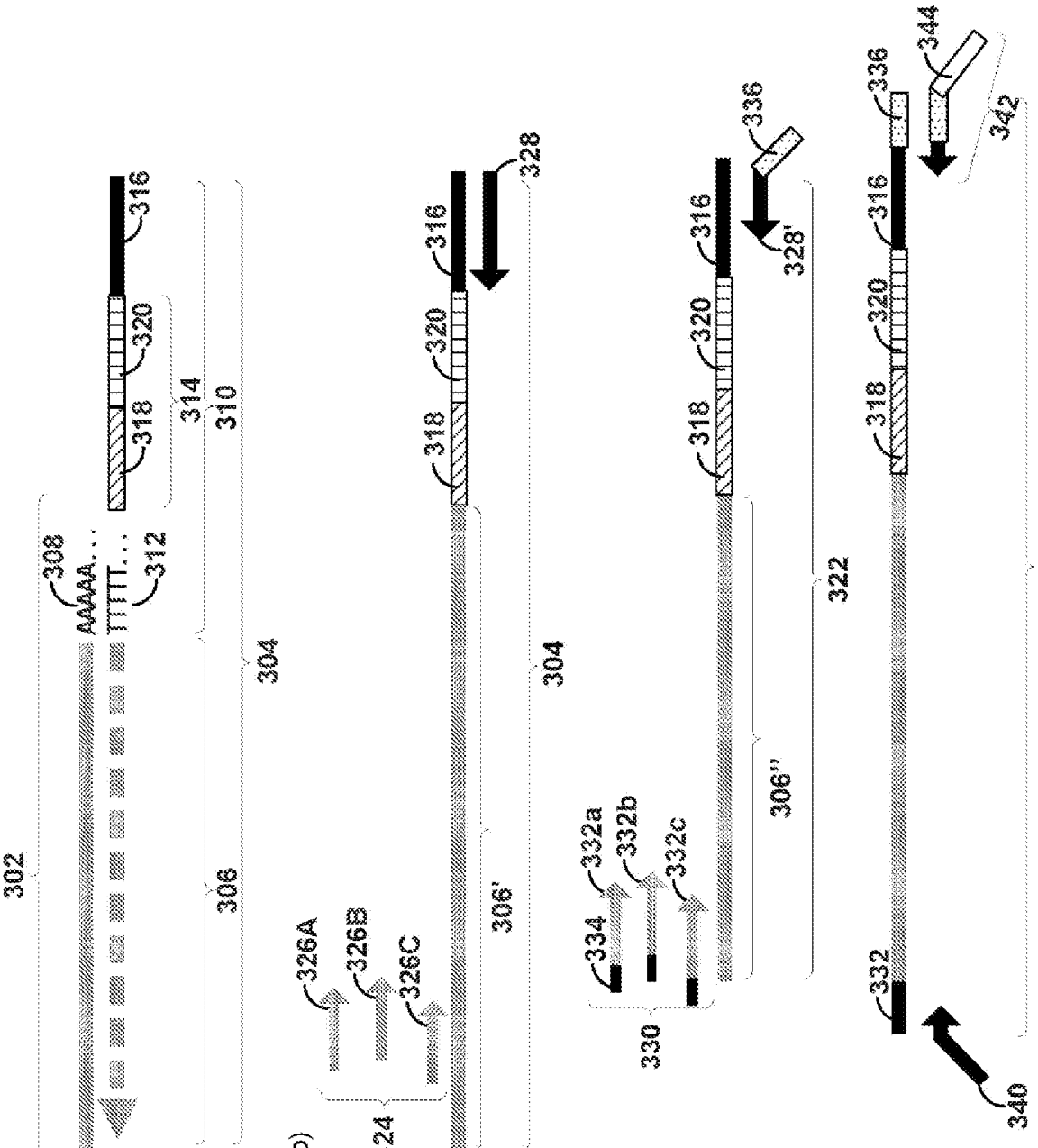

1. RT (formato de 96 pocillos)
302
308
AAAA...
TTTT...
312
316
320
318
314
310
304
306
2. 1º PCR (agrupado en 1 tubo)
326A
326B
326C
324
306'
328
3. PCR anidada
334
332a
332b
332c
330
306''
322
328'
4. PCR de índice (único tubo)
332
340
338
342
344
336

FIG. 3

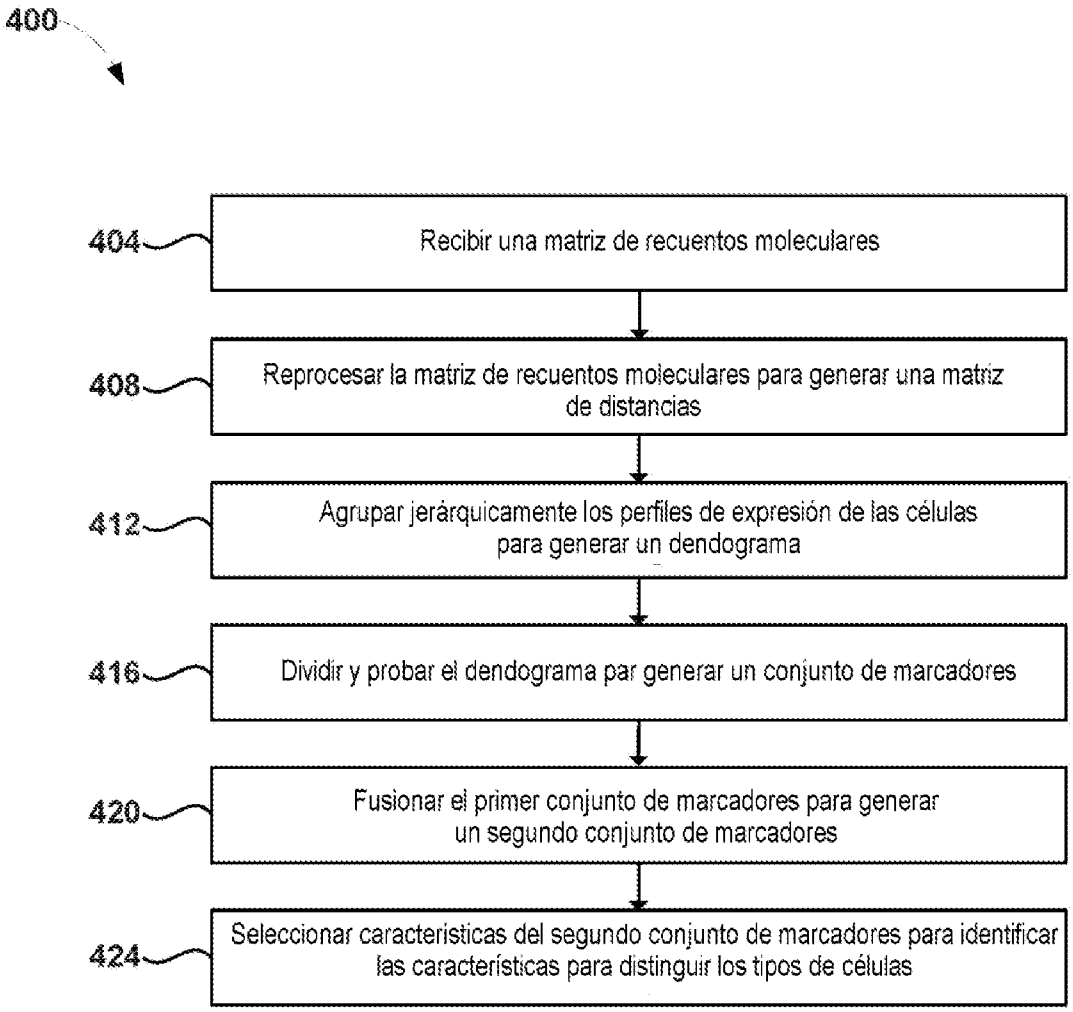

400
404
Recibir una matriz de recuentos moleculares
408
Reprocesar la matriz de recuentos moleculares para generar una matriz de distancias
412
Agrupar jerárquicamente los perfiles de expresión de las células para generar un dendograma
416
Dividir y probar el dendograma par generar un conjunto de marcadores
420
Fusionar el primer conjunto de marcadores para generar un segundo conjunto de marcadores
424
Seleccionar características del segundo conjunto de marcadores para identificar las características para distinguir los tipos de células

FIG. 4

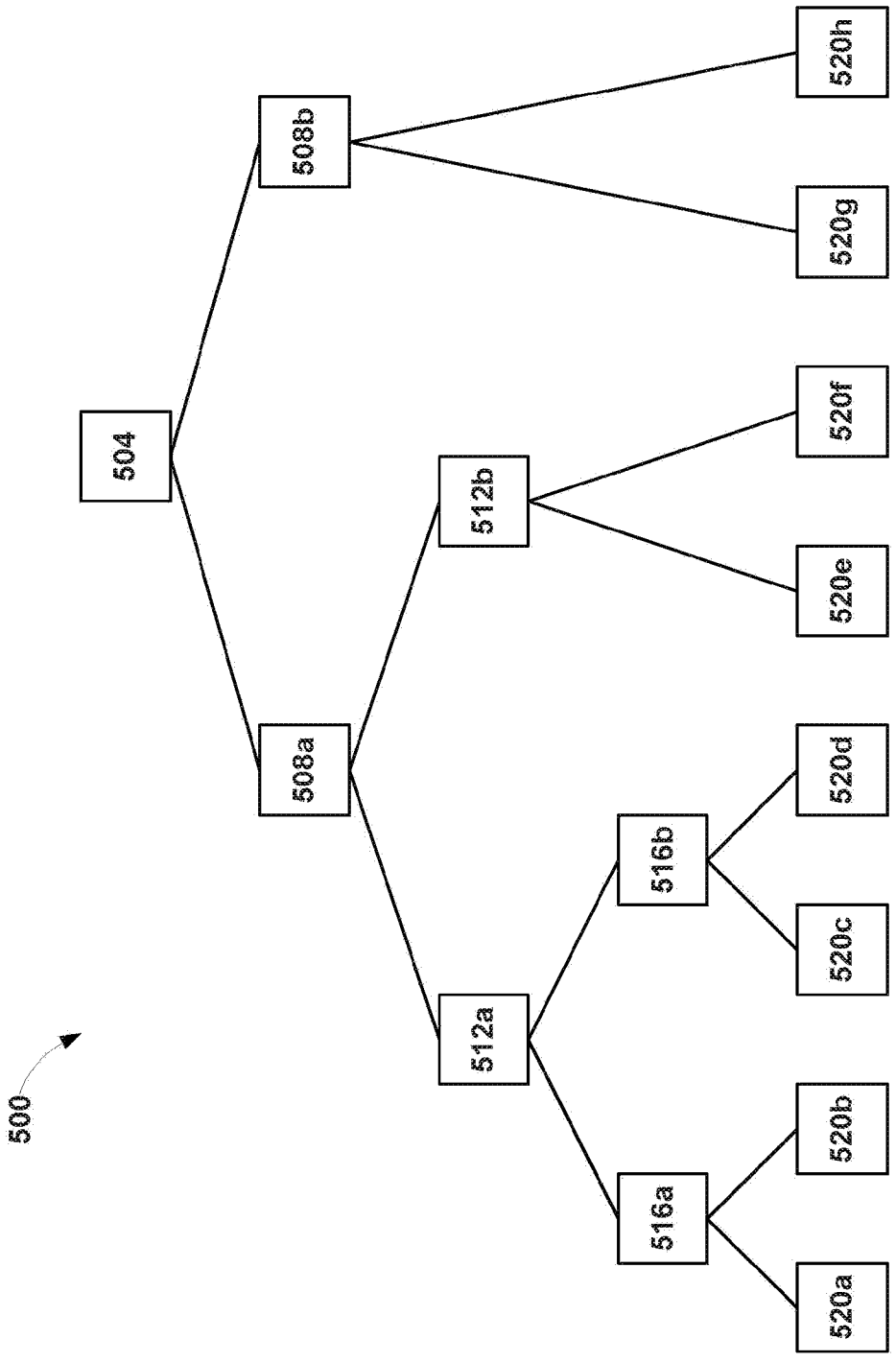

500
504
508b
508a
512b
512a
516b
516a
520h
520g
520f
520e
520d
520c
520b
520a

FIG. 5

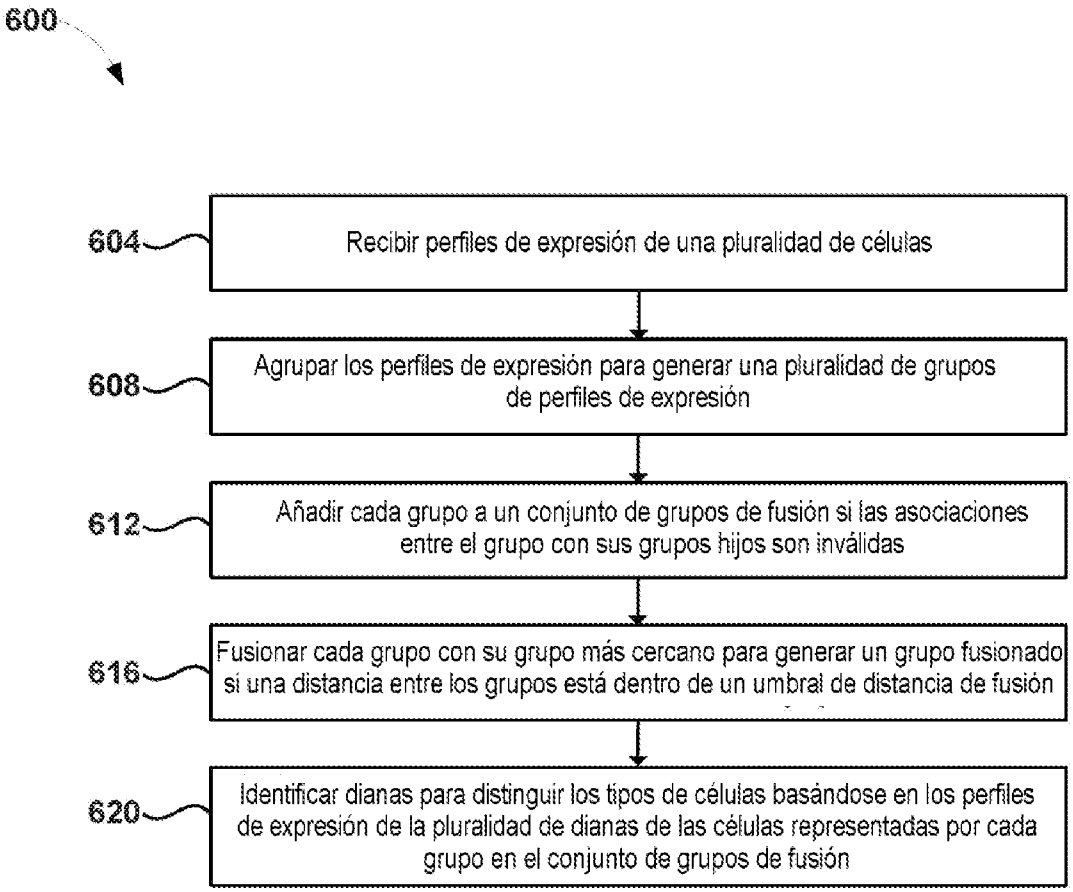

600
604
Recibir perfiles de expresión de una pluralidad de células
608
Agrupar los perfiles de expresión para generar una pluralidad de grupos de perfiles de expresión
612
Añadir cada grupo a un conjunto de grupos de fusión si las asociaciones entre el grupo con sus grupos hijos son inválidas
616
Fusionar cada grupo con su grupo más cercano para generar un grupo fusionado si una distancia entre los grupos está dentro de un umbral de distancia de fusión
620
Identificar dianas para distinguir los tipos de células basándose en los perfiles de expresión de la pluralidad de dianas de las células representadas por cada grupo en el conjunto de grupos de fusión

FIG. 6

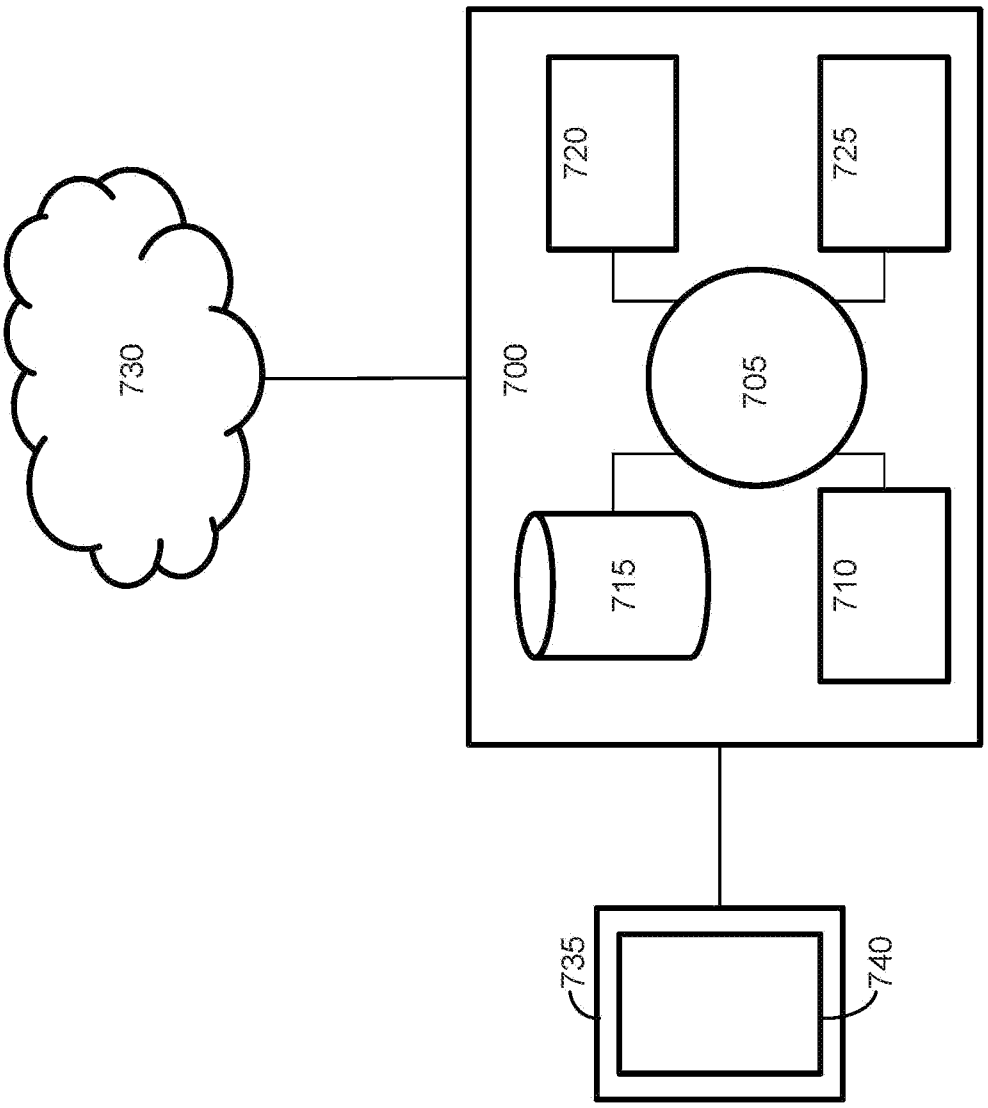


FIG. 7

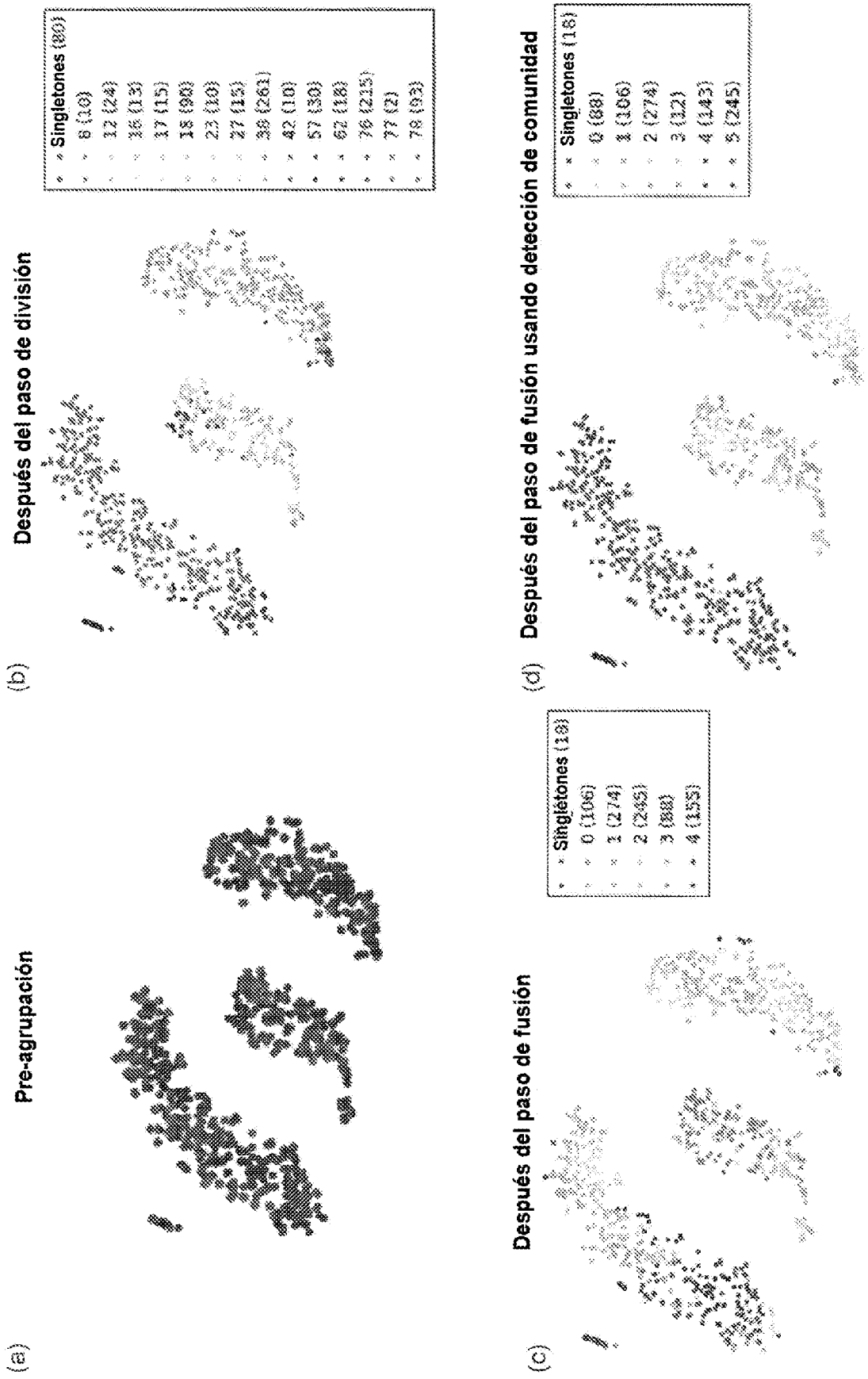

(a)
Pre-agrupación
(b)
Después del paso de división
Singletones (20)
8 (10)
12 (24)
16 (13)
17 (15)
18 (90)
23 (10)
27 (15)
39 (261)
42 (10)
57 (30)
62 (18)
76 (215)
77 (2)
78 (93)
(c)
Después del paso de fusión
Singletones (18)
0 (106)
1 (274)
2 (245)
3 (88)
4 (155)
(d)
Después del paso de fusión usando detección de comunidad
Singletones (18)
0 (88)
1 (106)
2 (274)
3 (32)
4 (143)
5 (245)

FIG. 8

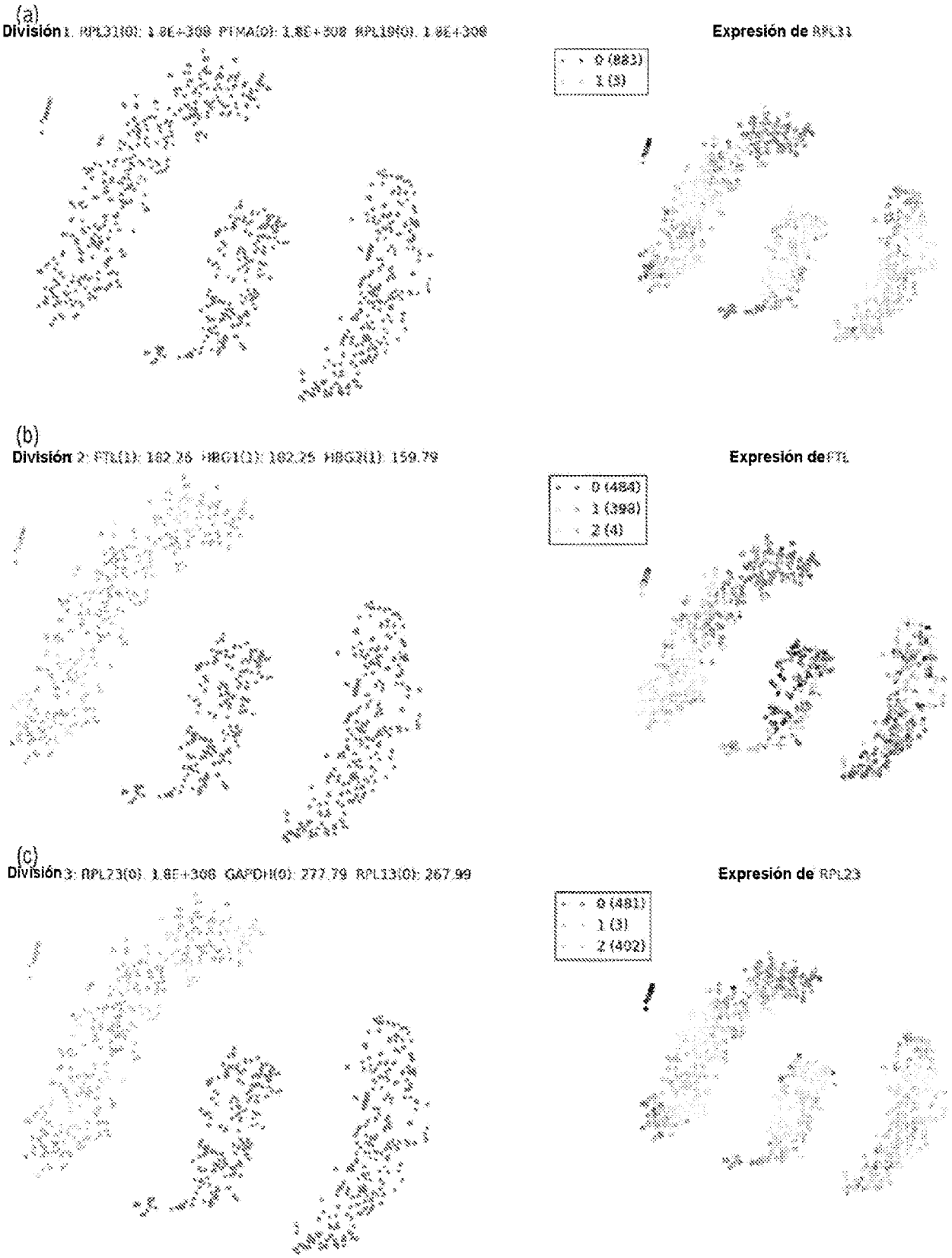

(a)
División 1: RPL31(0): 1.8E+308 PTMA(0): 1.8E+308 RPL19(0): 1.8E+308
Expresión de RPL31
0 (883)
1 (3)
(b)
División 2: FTL(1): 182.26 HBG1(1): 182.25 HBG2(1): 159.79
Expresión de FTL
0 (484)
1 (398)
2 (4)
(c)
División 3: RPL23(0): 1.8E+308 GAPDH(0): 277.79 RPL13(0): 267.99
Expresión de RPL23
0 (481)
1 (3)
2 (402)

FIG. 9

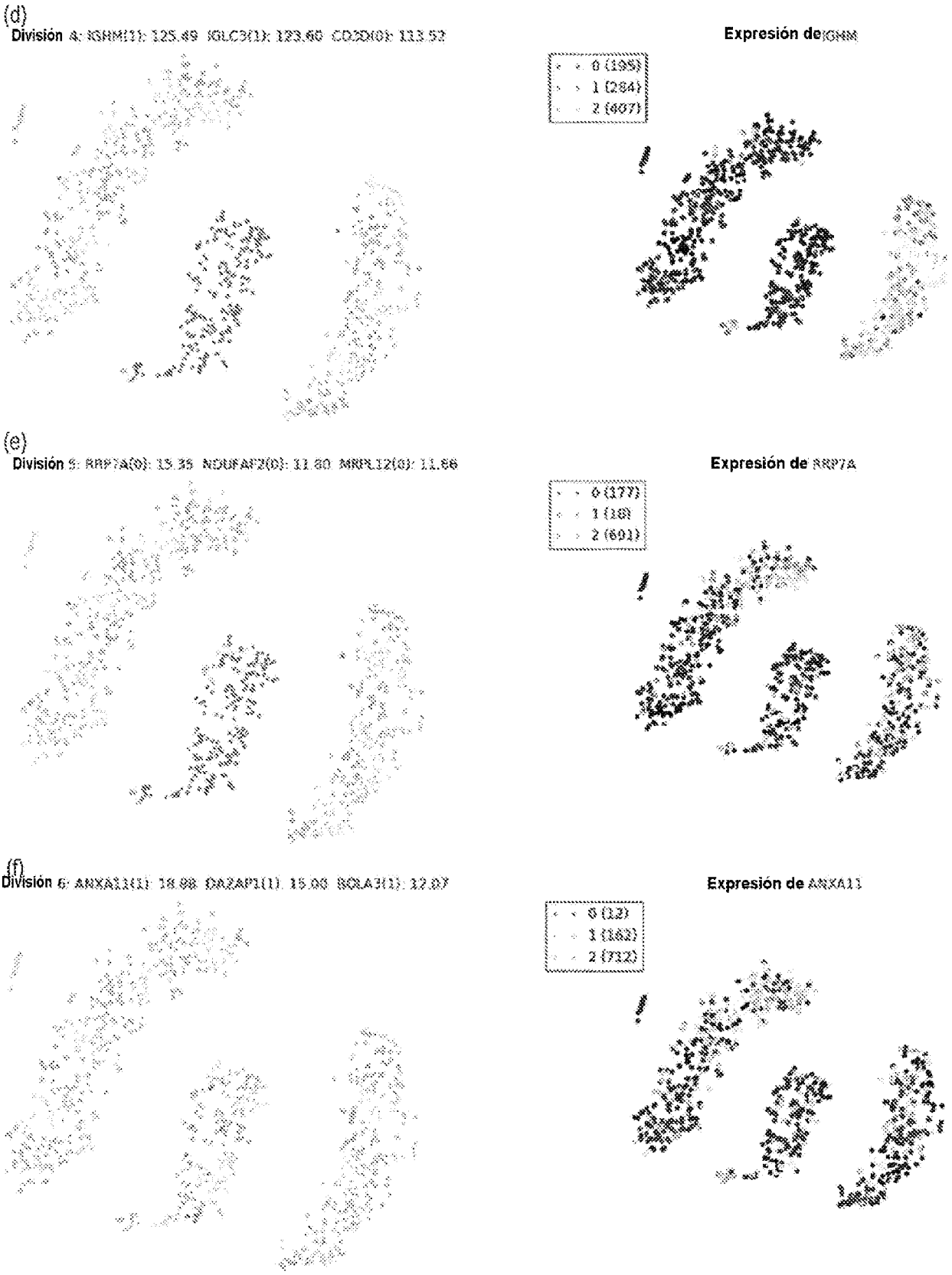


FIG. 9 (Continuación)

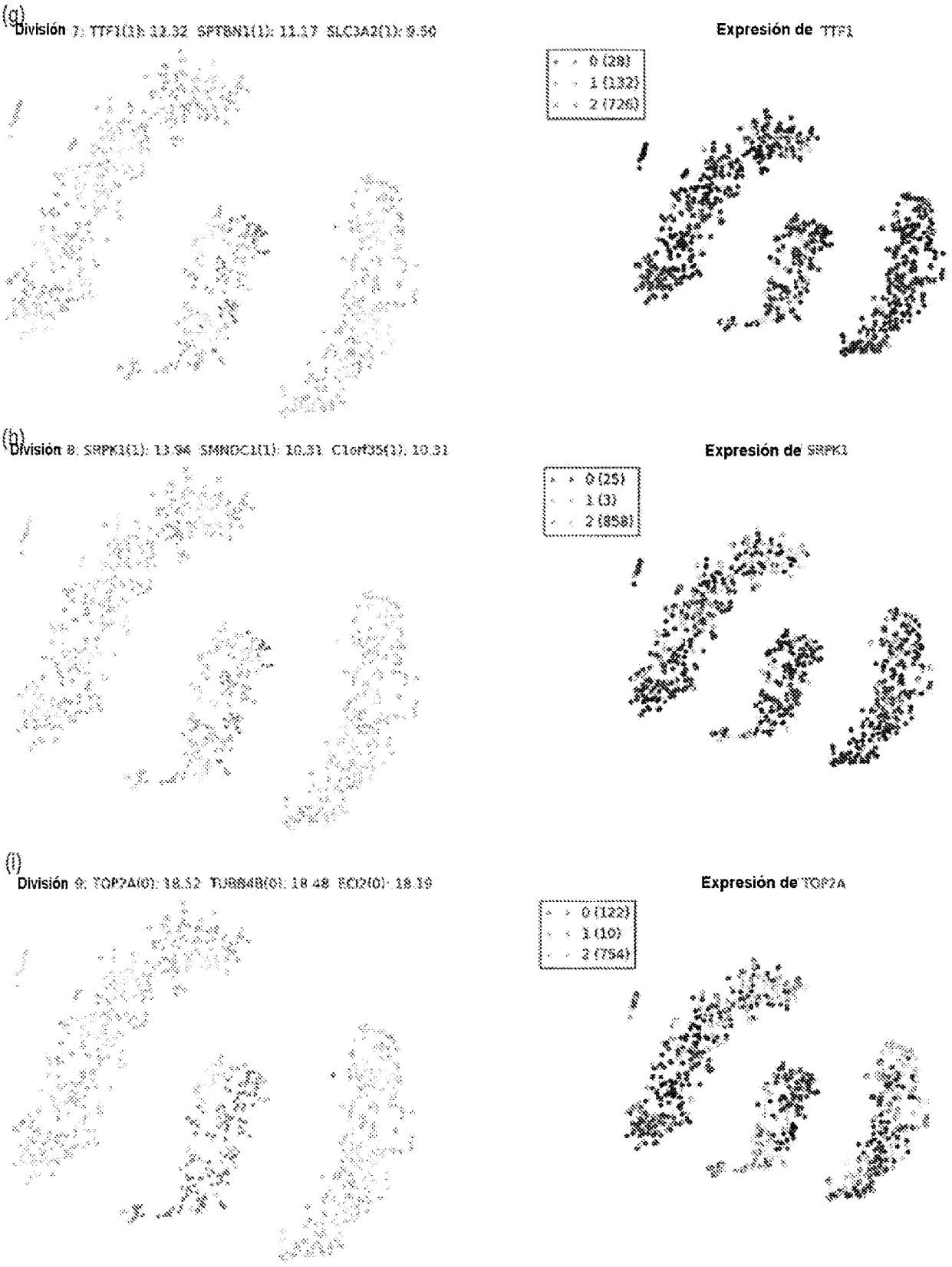


FIG. 9 (Continuación)

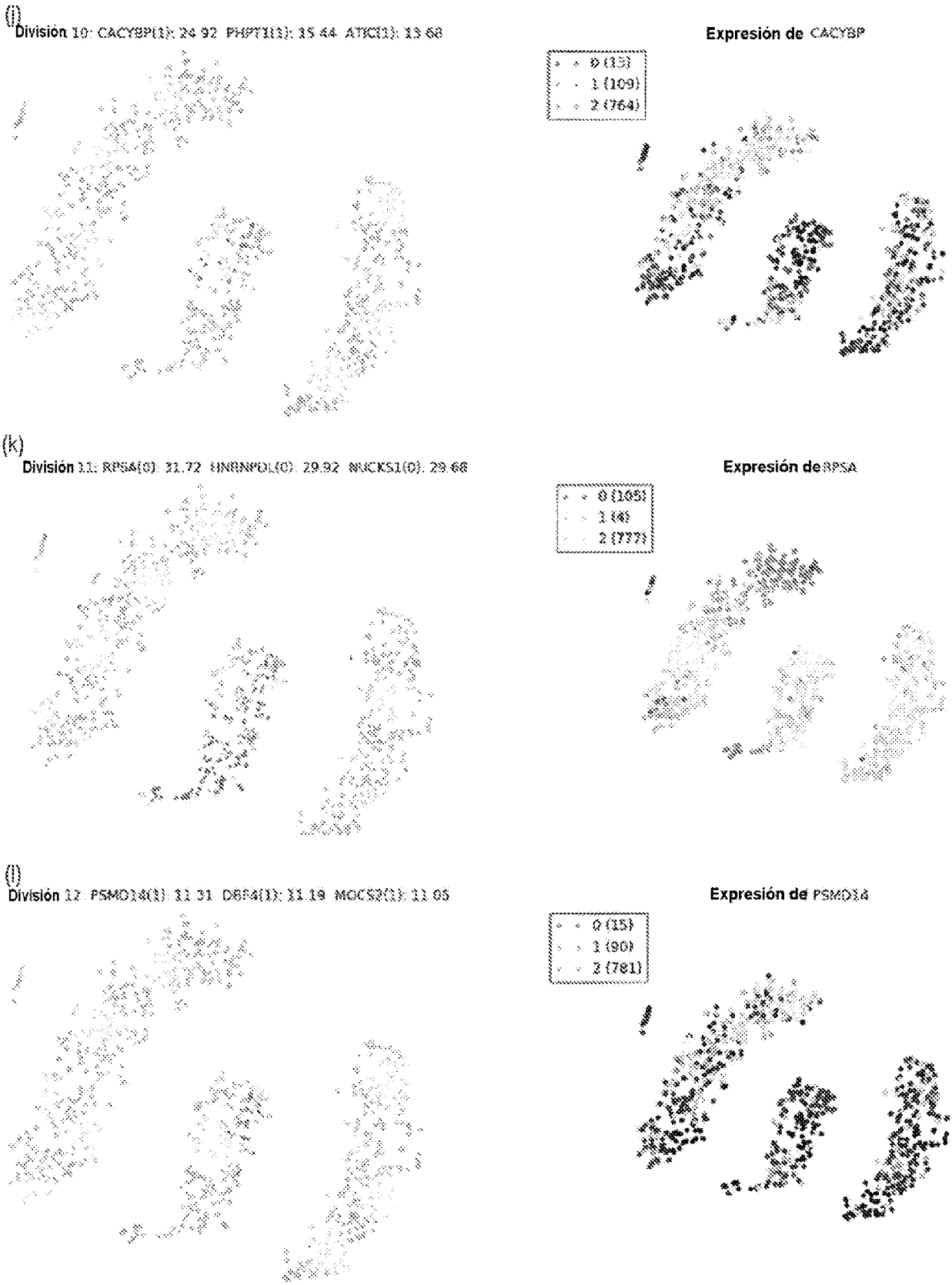


FIG. 9 (Continuación)

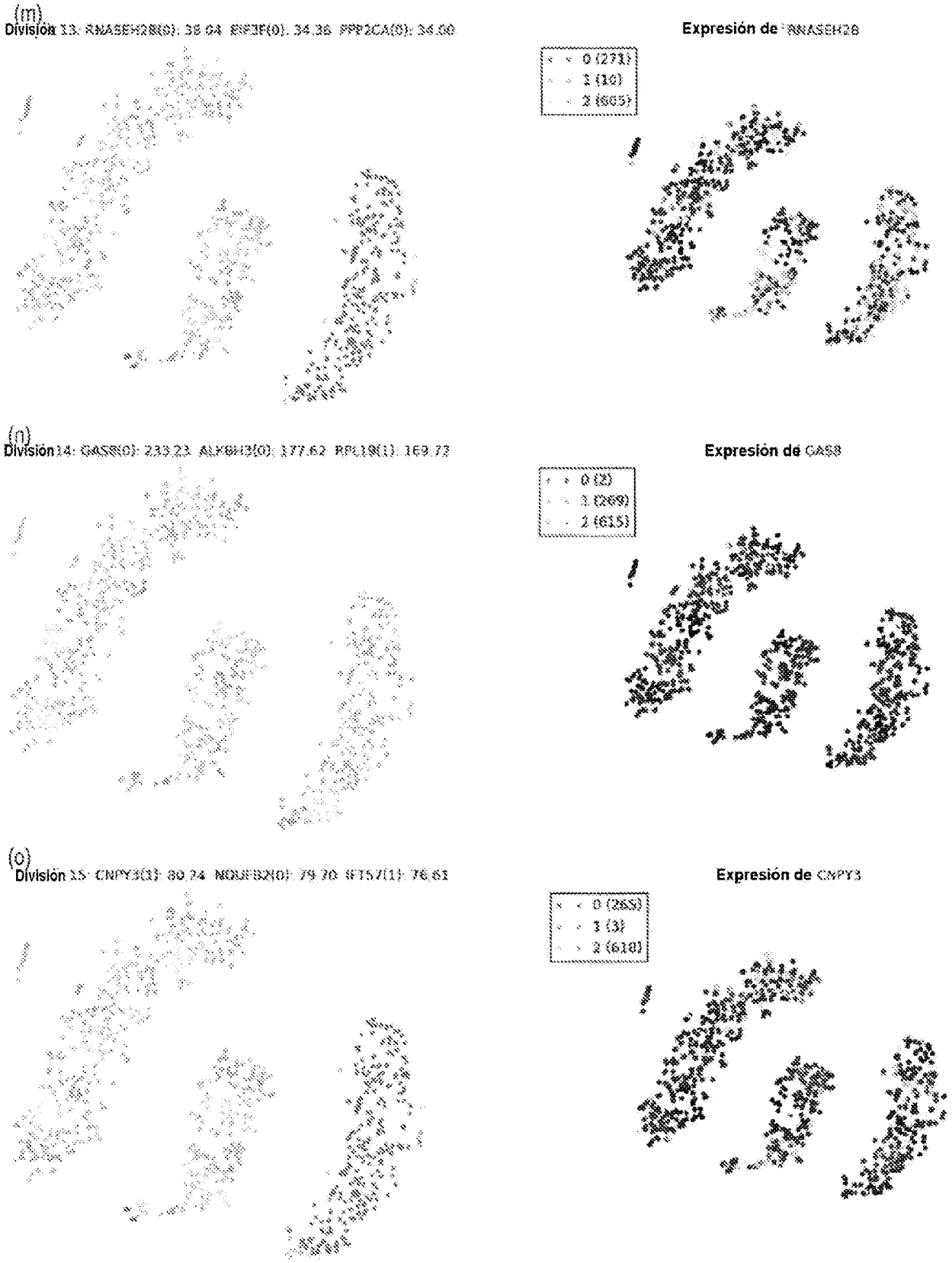


FIG. 9 (Continuación)

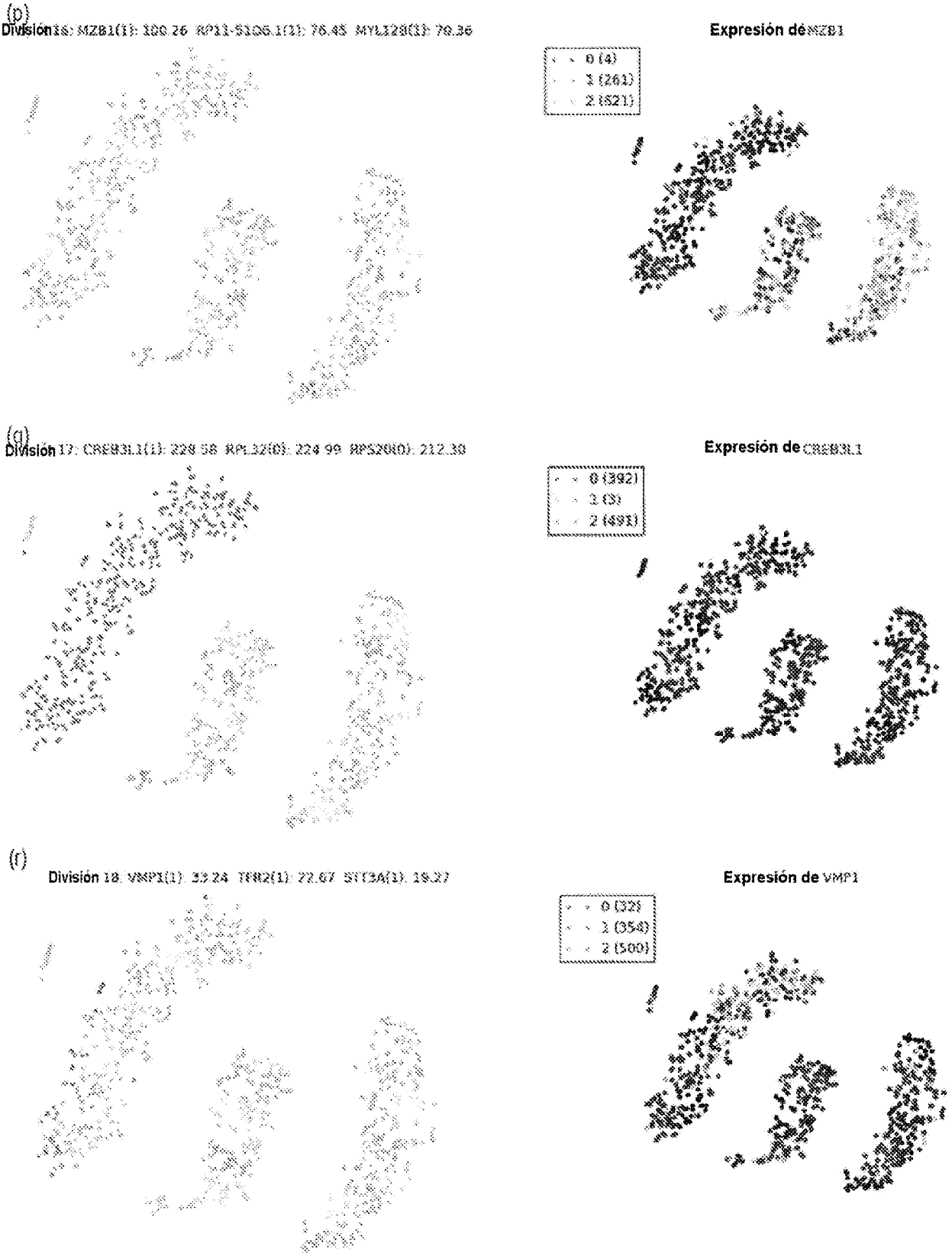


FIG. 9 (Continuación)

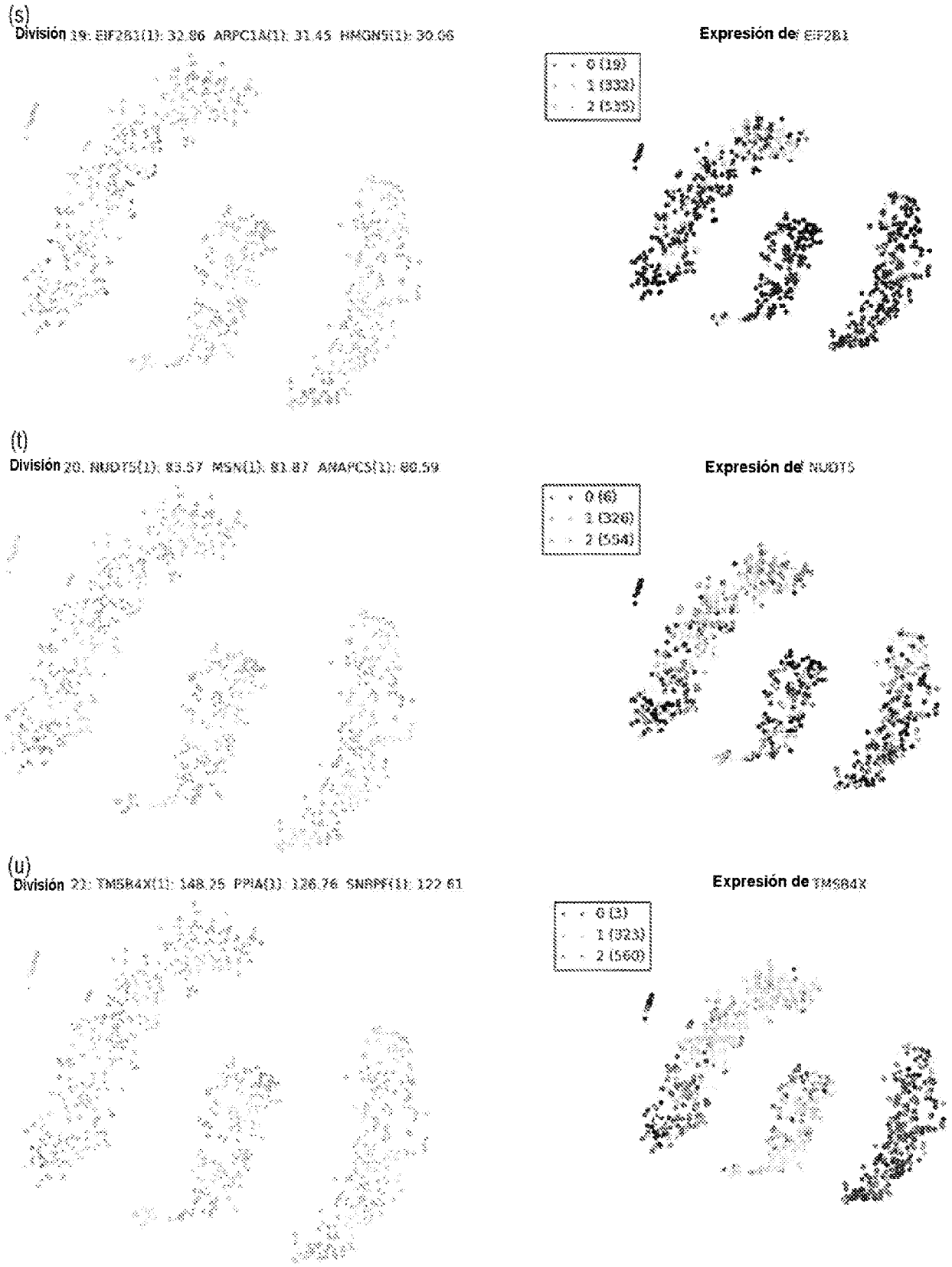


FIG. 9 (Continuación)

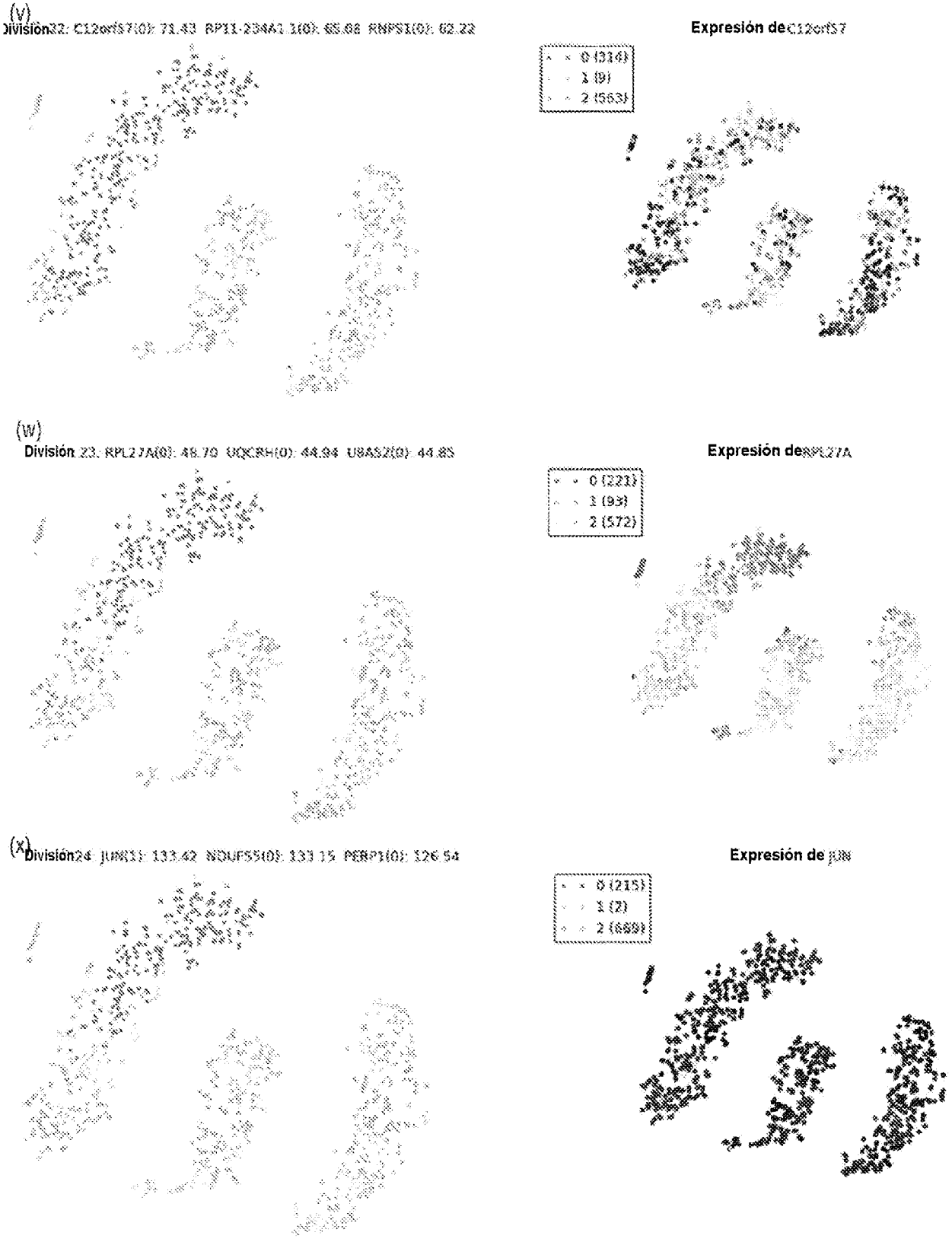


FIG. 9 (Continuación)

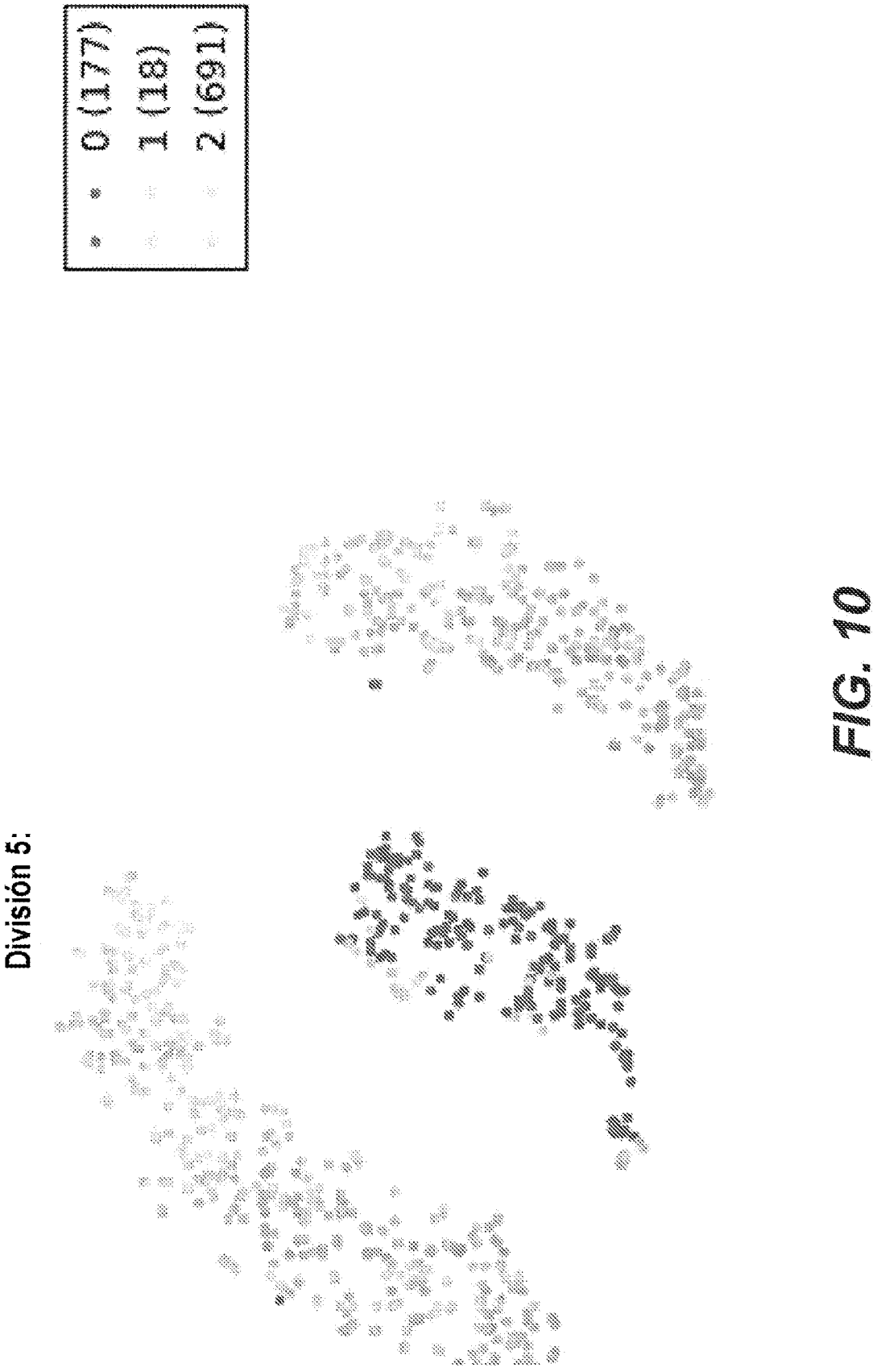


FIG. 10

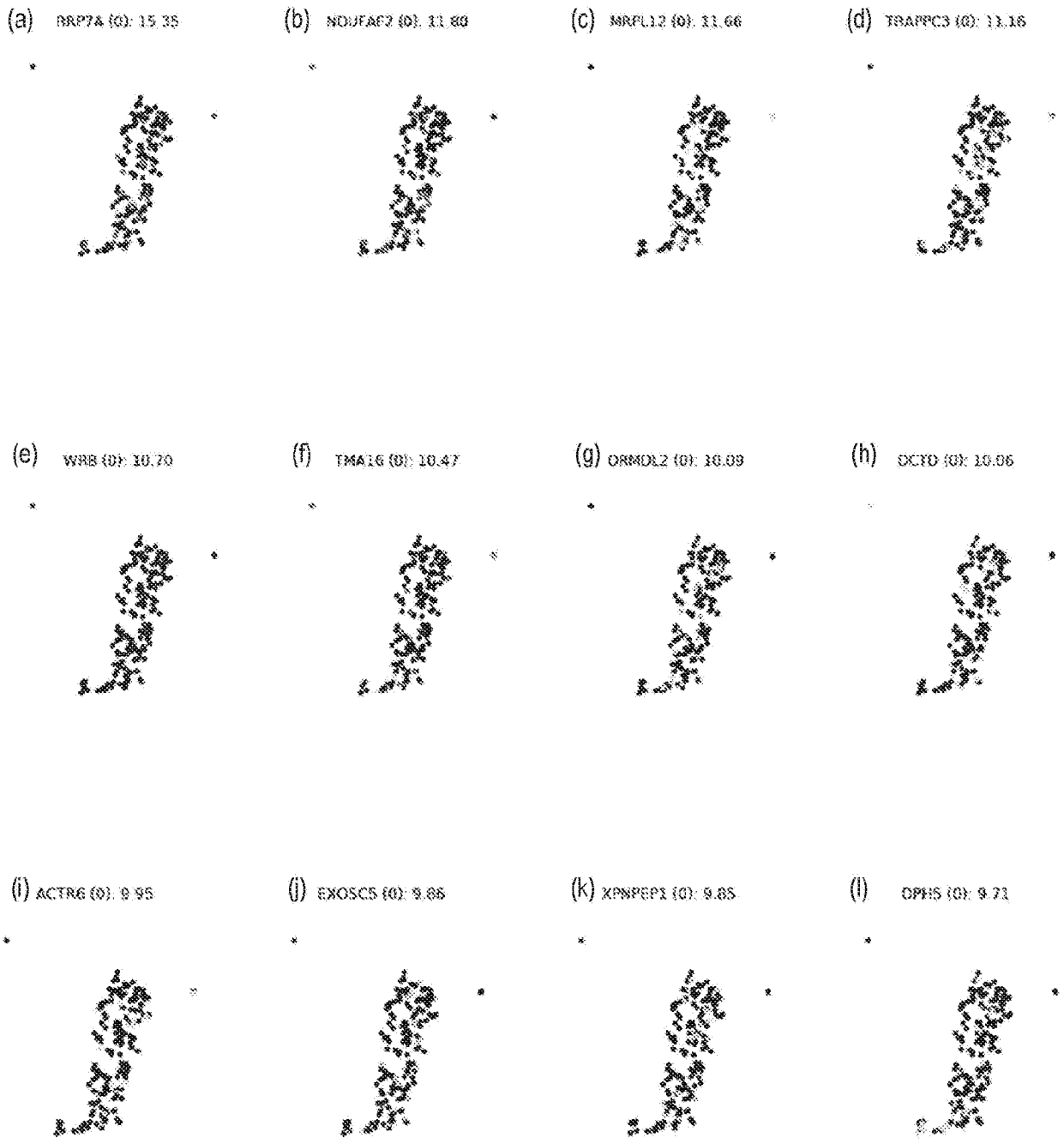

(a) BRP7A (0): 15.35
(b) NDUFAF2 (0): 11.88
(c) MRPL12 (0): 11.66
(d) TRAPPC3 (0): 11.16
(e) WRB (0): 10.70
(f) TMA16 (0): 10.47
(g) ORMDL2 (0): 10.09
(h) OCTD (0): 10.06
(i) ACTR6 (0): 9.95
(j) EXOSC5 (0): 9.86
(k) XPNPEP1 (0): 9.85
(l) OPHS (0): 9.71

FIG. 11

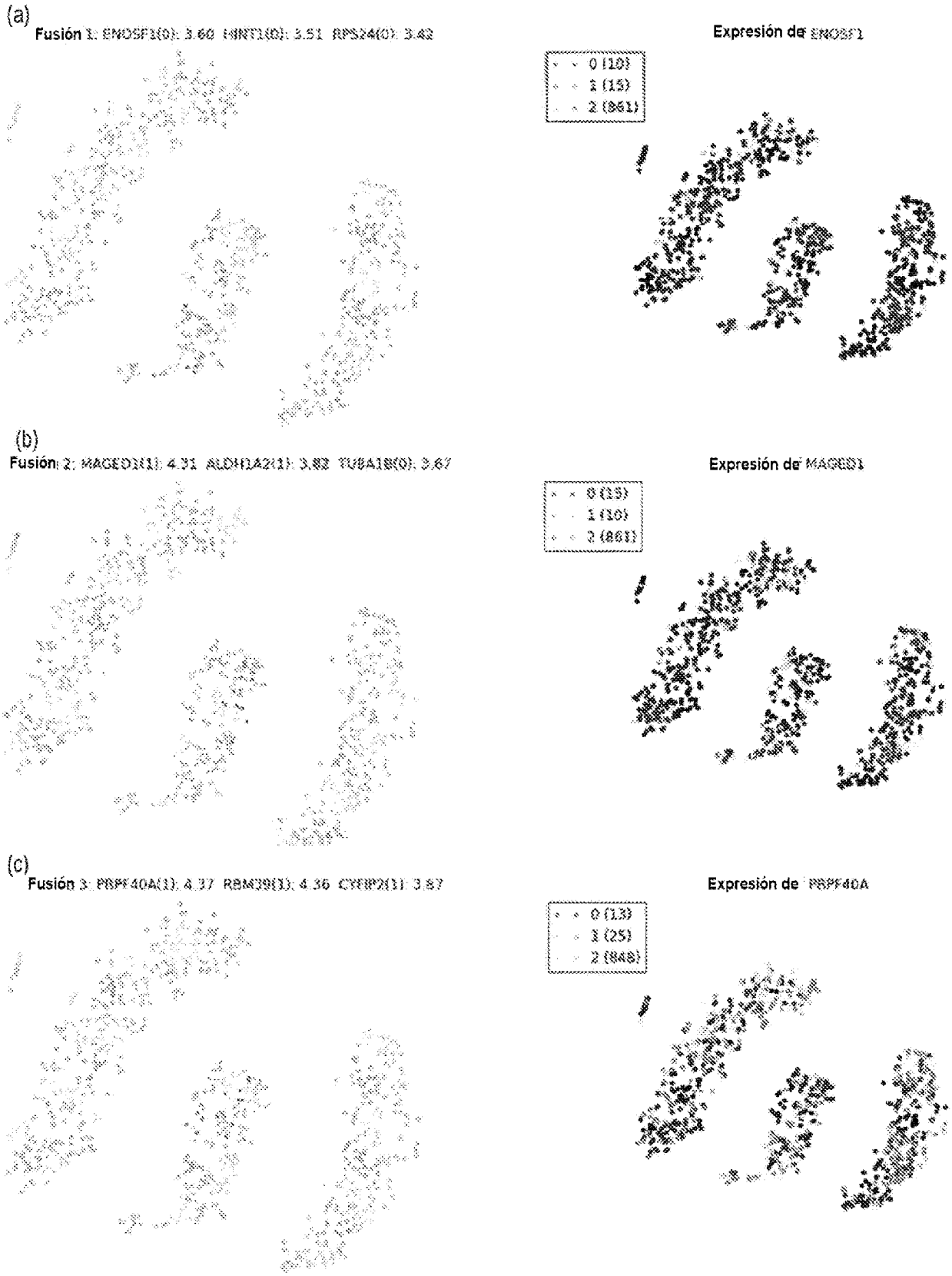

(a)
Fusión 1: ENOSF1(0): 3.60 HINT1(0): 3.51 RPS24(0): 3.42
Expresión de ENOSF1
0 (10)
1 (15)
2 (861)
(b)
Fusión 2: MAGED1(1): 4.31 ALDH1A2(1): 3.82 TUBA1B(0): 3.67
Expresión de MAGED1
0 (15)
1 (10)
2 (861)
(c)
Fusión 3: PRPF40A(1): 4.37 RBM39(1): 4.36 CYFIP2(1): 3.87
Expresión de PRPF40A
0 (13)
1 (25)
2 (848)

FIG. 12

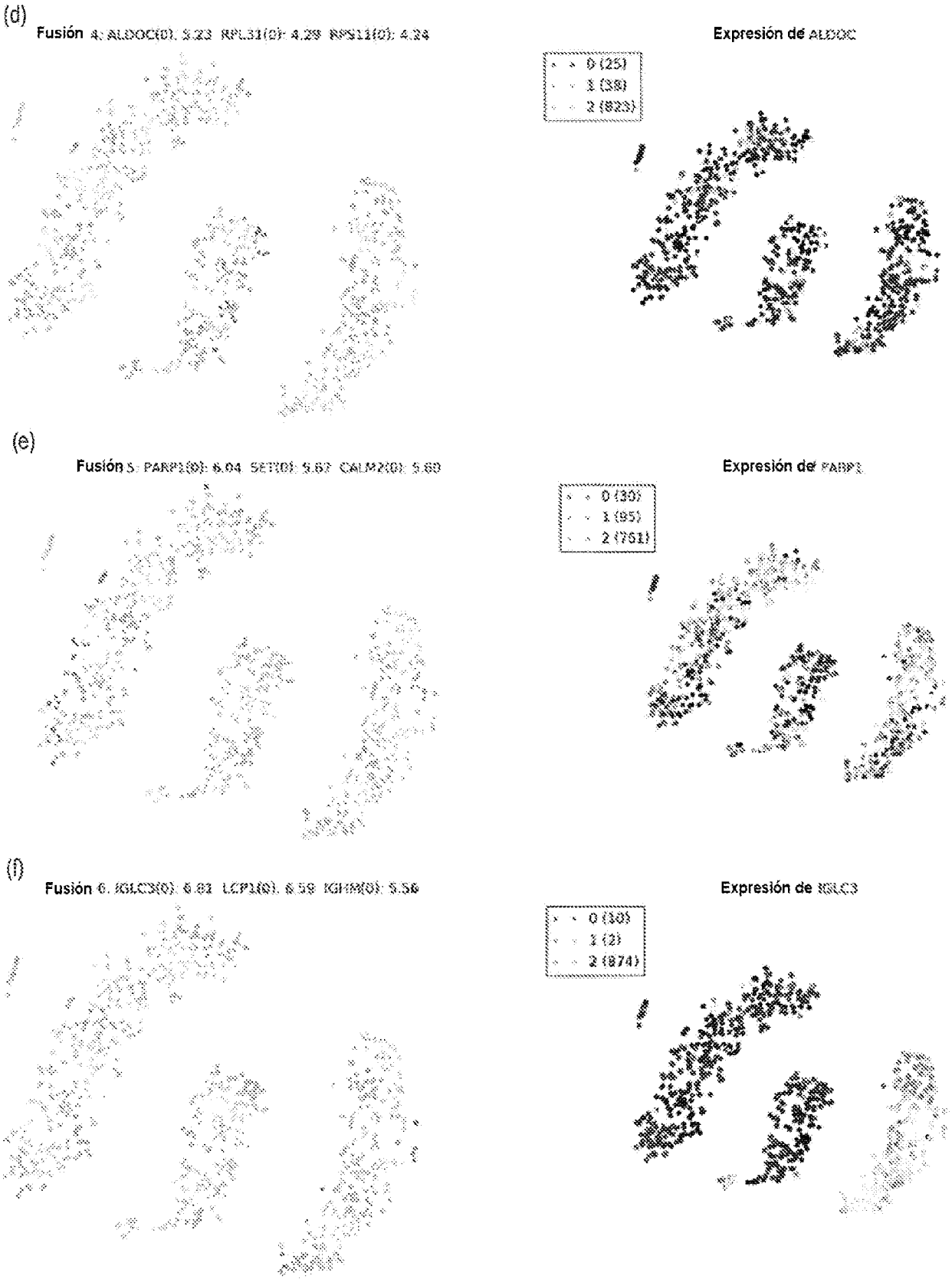


FIG. 12 (Continuación)

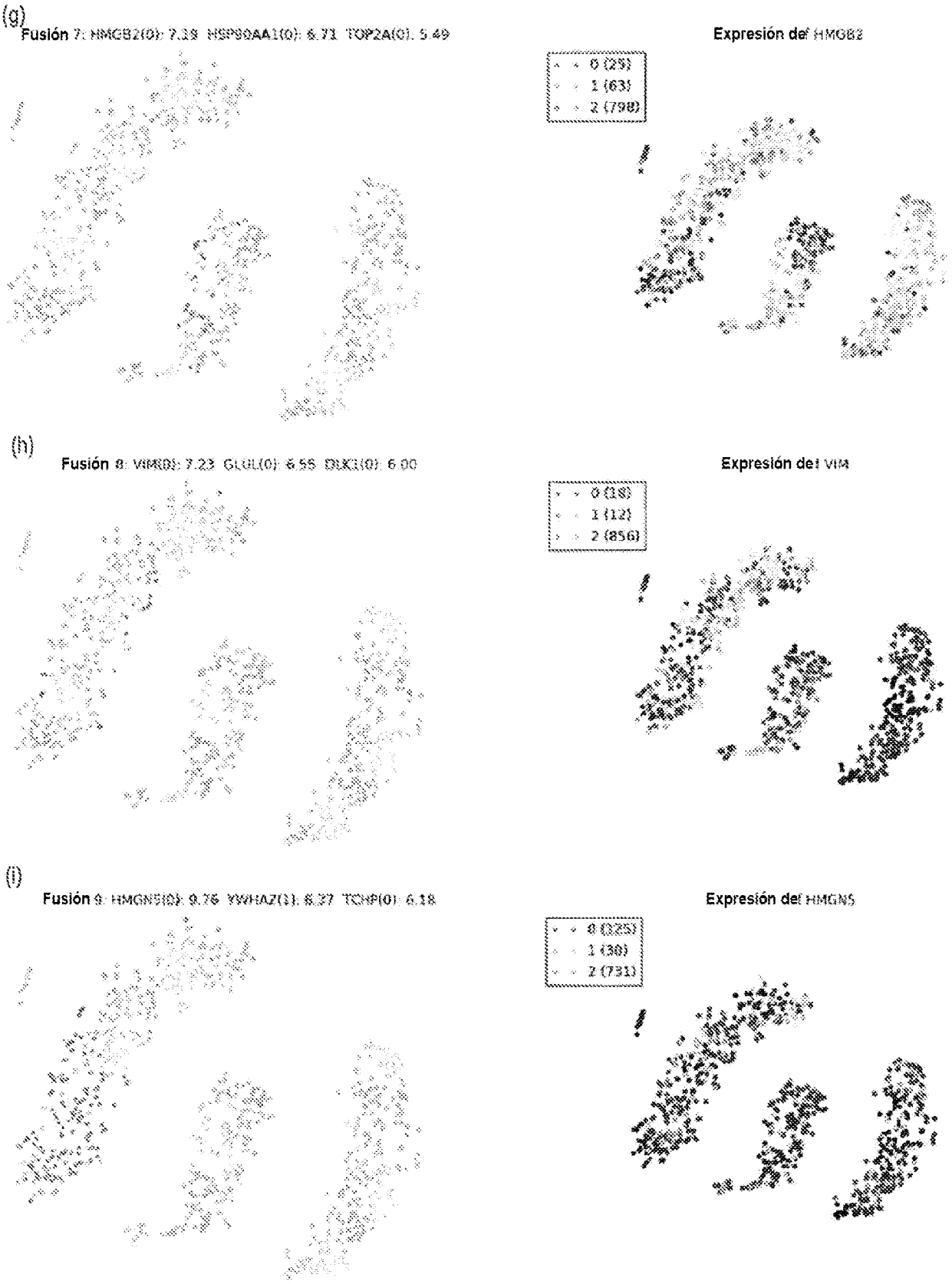


FIG. 12 (Continuación

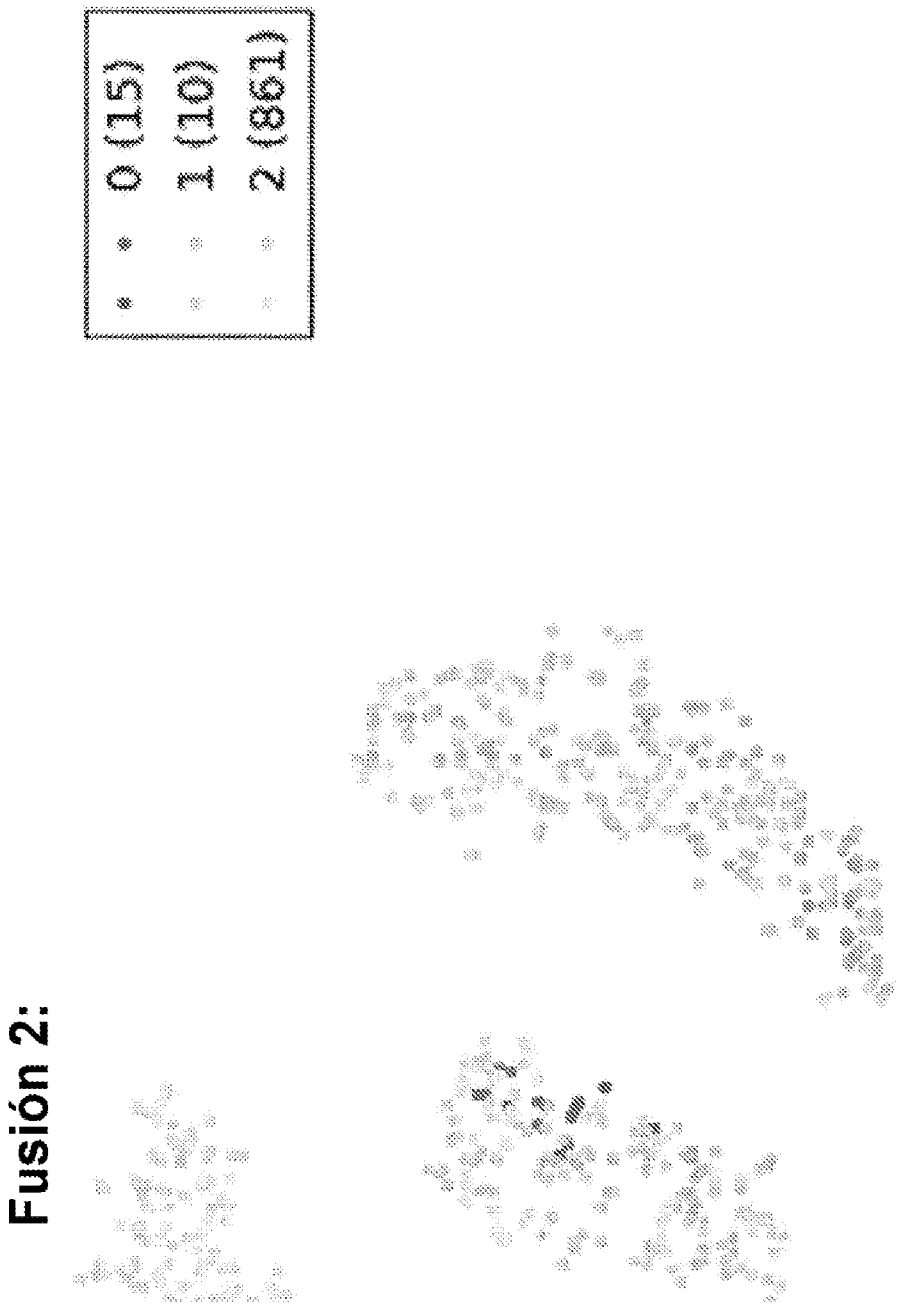


FIG. 13

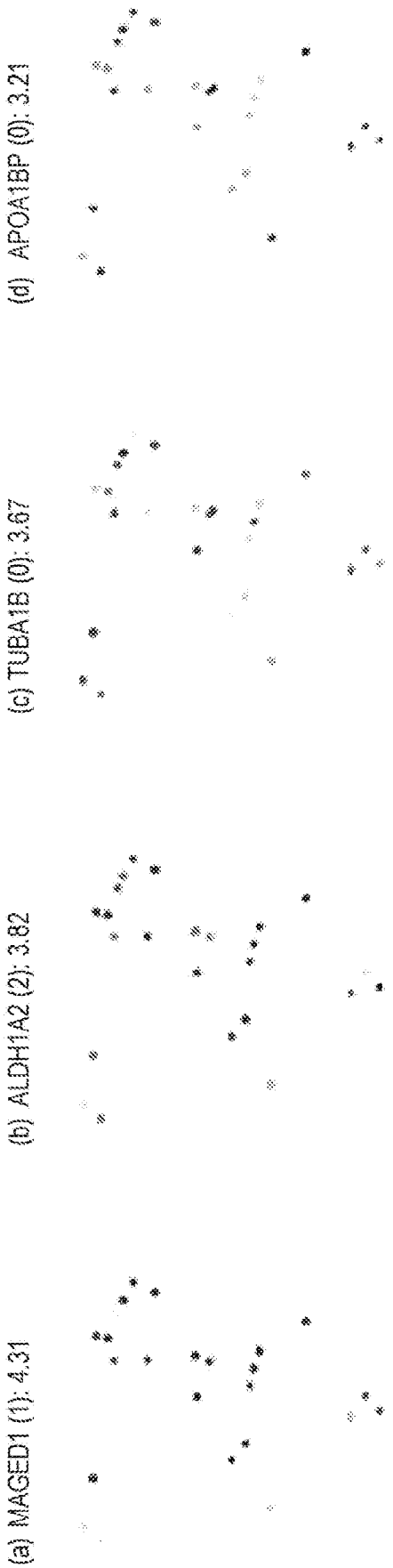
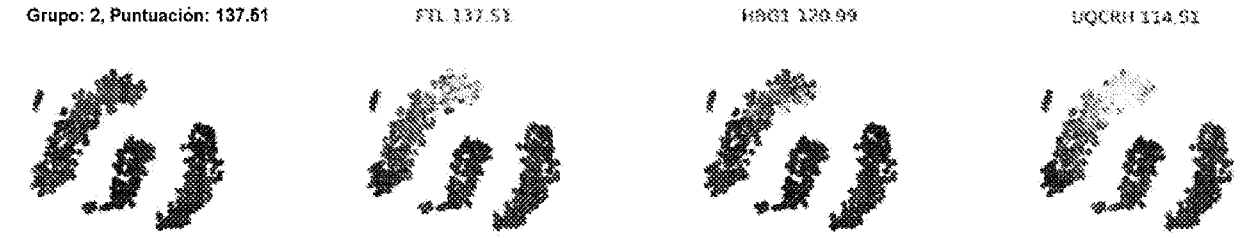


FIG. 14

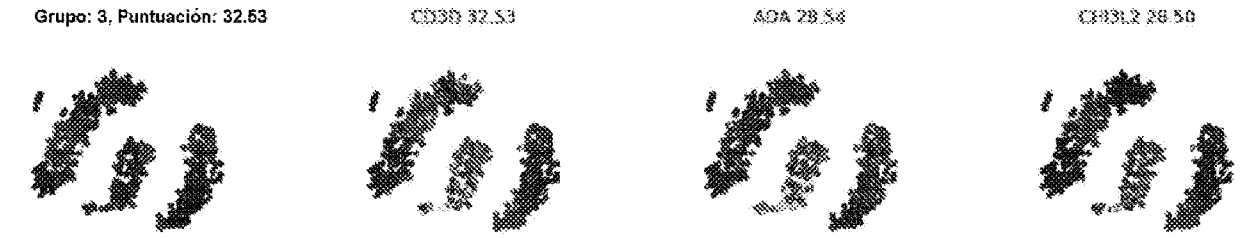

(a)
Grupo: -1, Puntuación:3.60
MALAT1 3.60
MTATP6P1 2.08
RP5-857K21.11 3.54
(b)
Grupo: 0, Puntuación: 47.69
CD3D 47.59
CHI3L2 44.91
TBBC2 39.72
(c)
Grupo: 1, Puntuación: 216.73
IGLC3 216.72
IGJ 190.36
CD74 168.36

FIG. 15

(d)



(e)



(f)

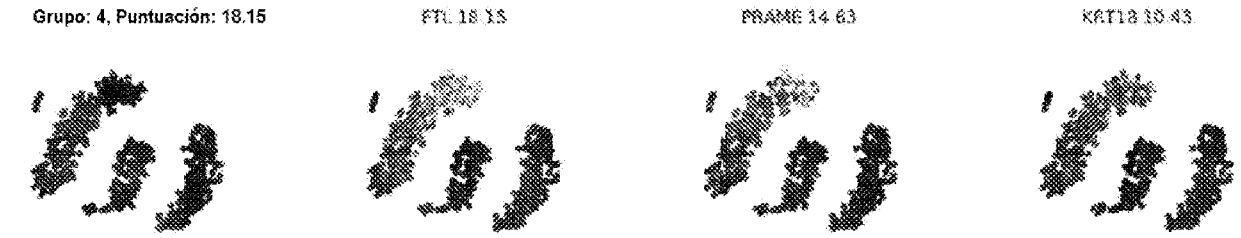


FIG. 15 (Continuación)

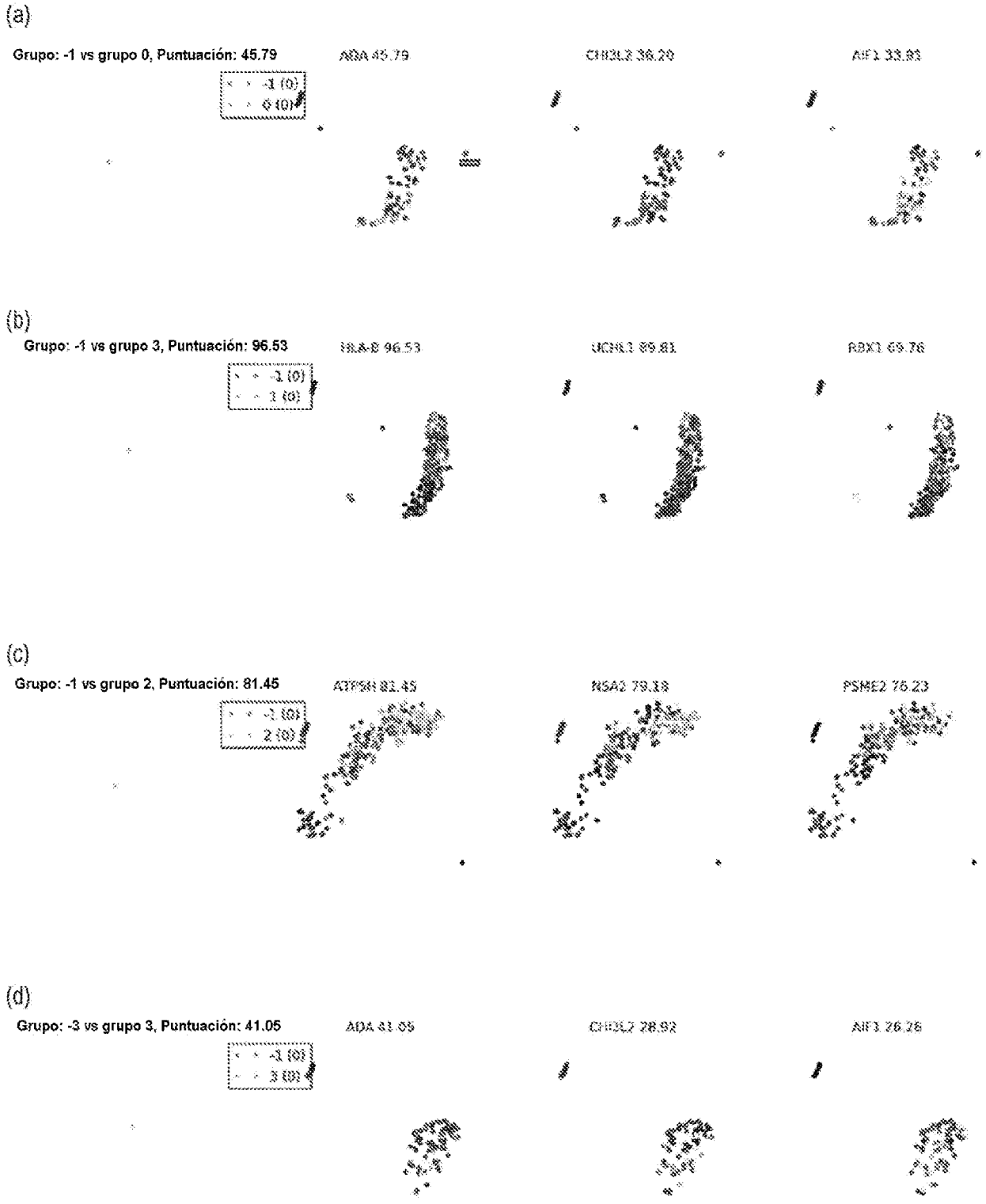

(a)
Grupo: -1 vs grupo 0, Puntuación: 45.79
ADA 45.79
CHI3L2 36.20
AIF1 33.81
-1 (0)
0 (0)
(b)
Grupo: -1 vs grupo 3, Puntuación: 96.53
HLA-B 96.53
IGHG1 89.81
RBX1 69.76
-1 (0)
3 (0)
(c)
Grupo: -1 vs grupo 2, Puntuación: 81.45
ATP5H 81.45
RSA2 79.18
PSME2 76.23
-1 (0)
2 (0)
(d)
Grupo: -3 vs grupo 3, Puntuación: 41.05
ADA 41.05
CHI3L2 28.92
AIF1 26.26
-3 (0)
3 (0)

FIG. 16

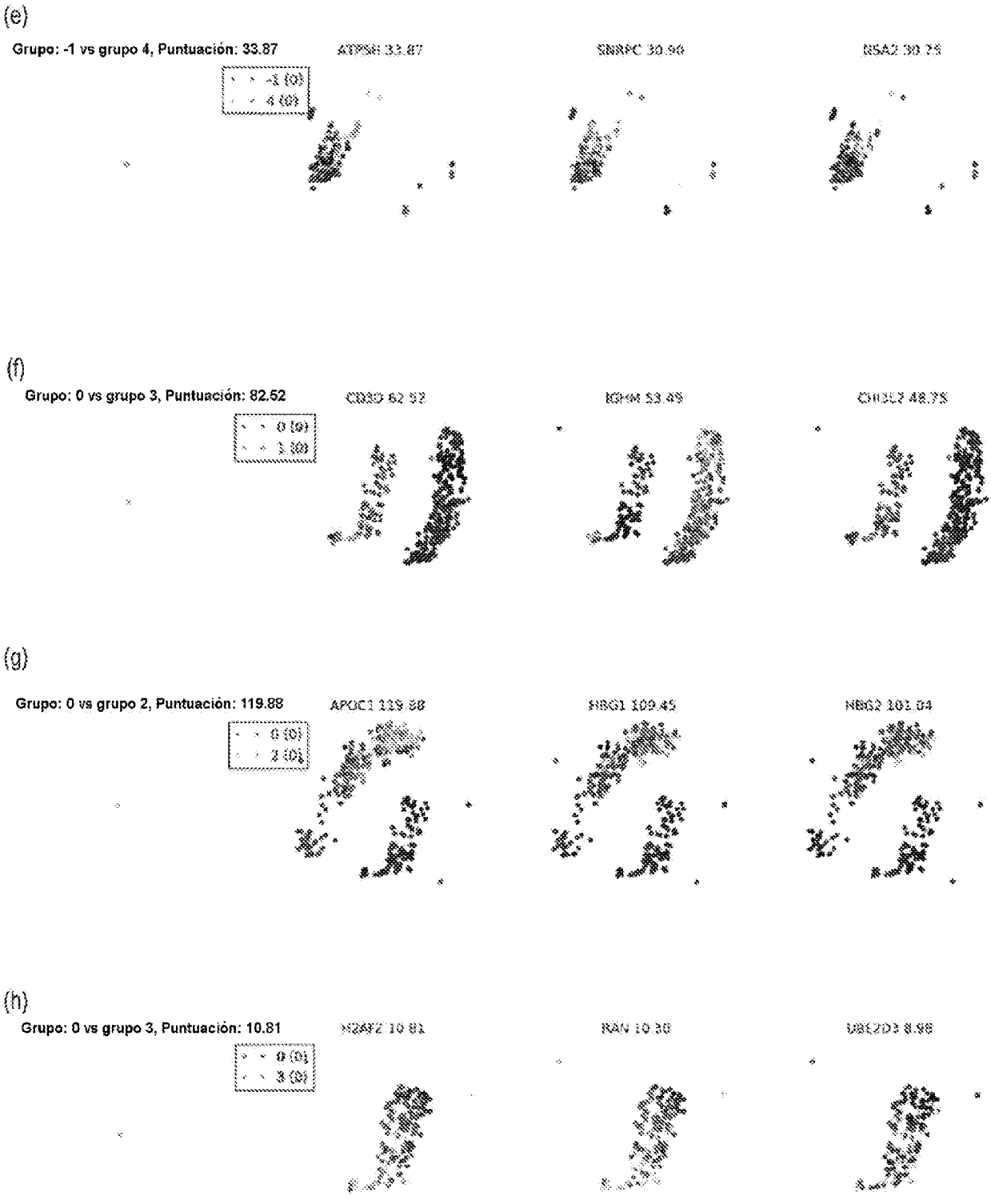


FIG. 16 (Continuación)

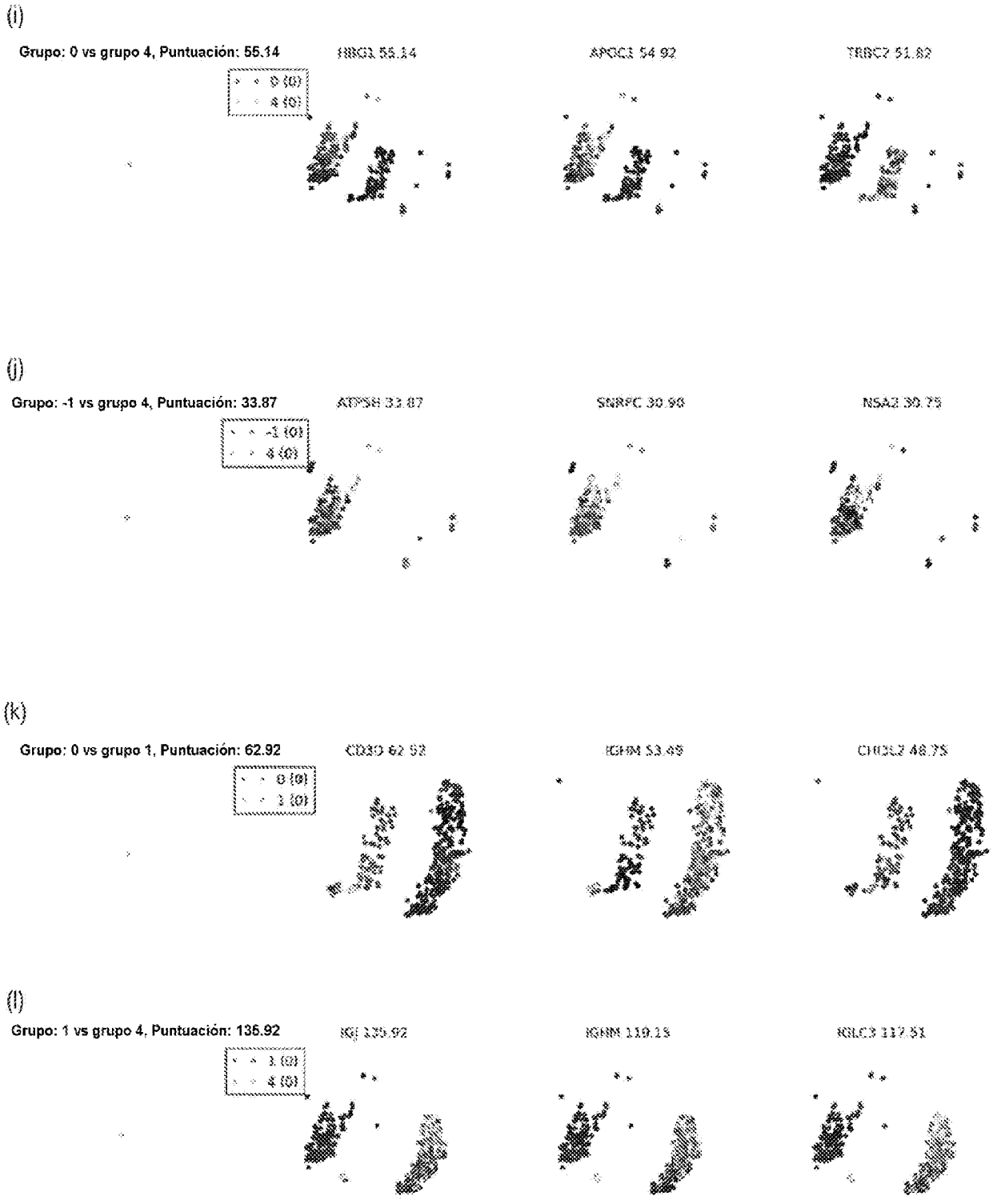
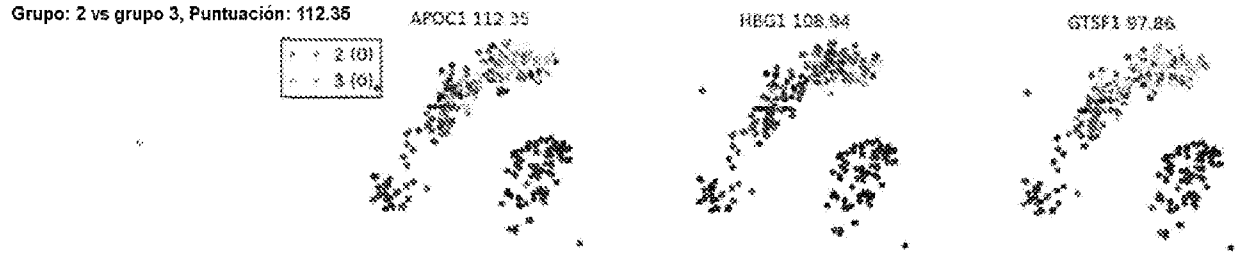
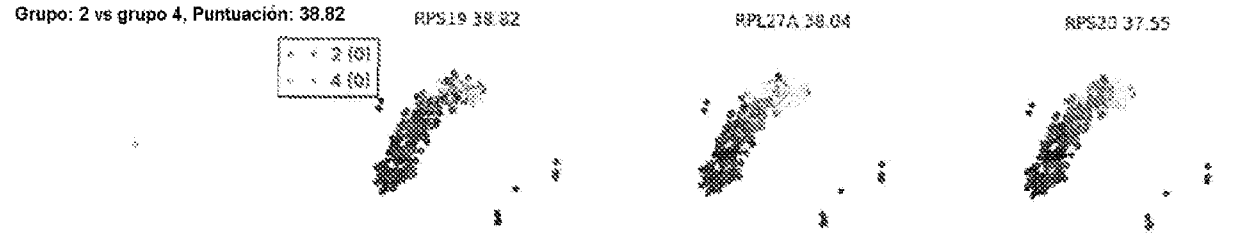


FIG. 16 (Continuación)

(m)



(n)



(o)

FIG. 16 (Continuación)

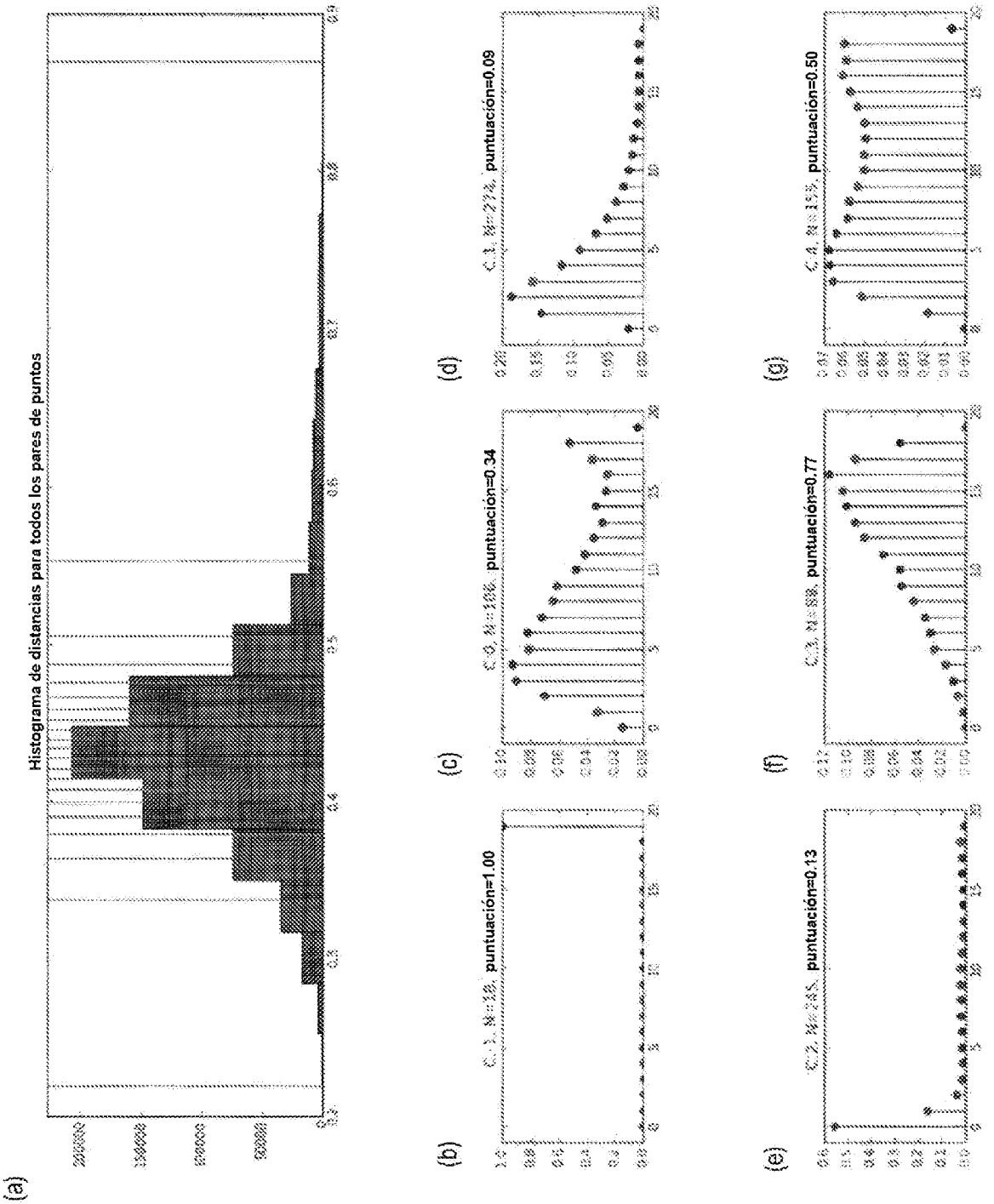

(a)
Histograma de distancias para todos los pares de puntos
(b)
(c)
puntuación=0.34
(d)
puntuación=0.09
(e)
puntuación=0.13
(f)
puntuación=0.77
(g)
puntuación=0.50

FIG. 17

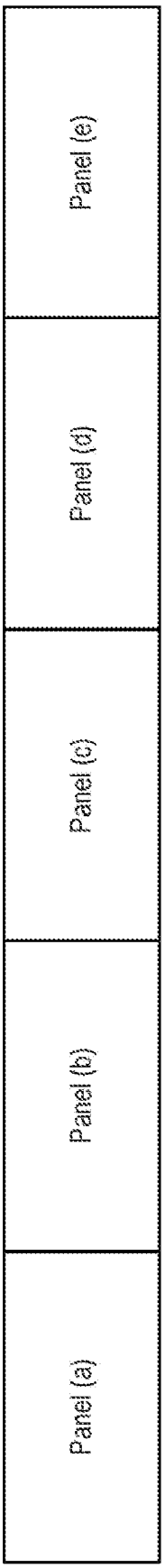


FIG. 18

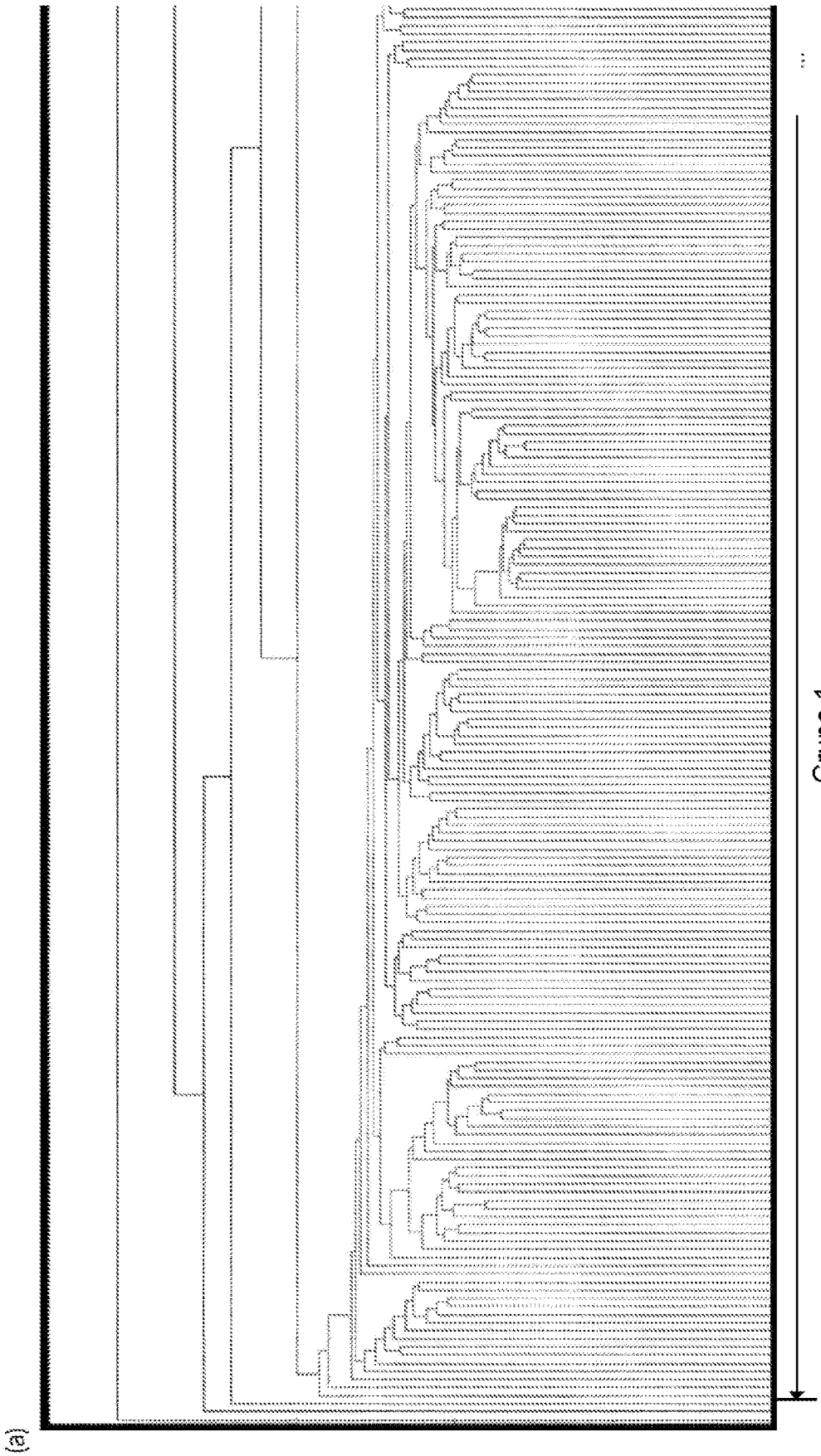


FIG.18 (continuación)

FIG.18 (continuación)

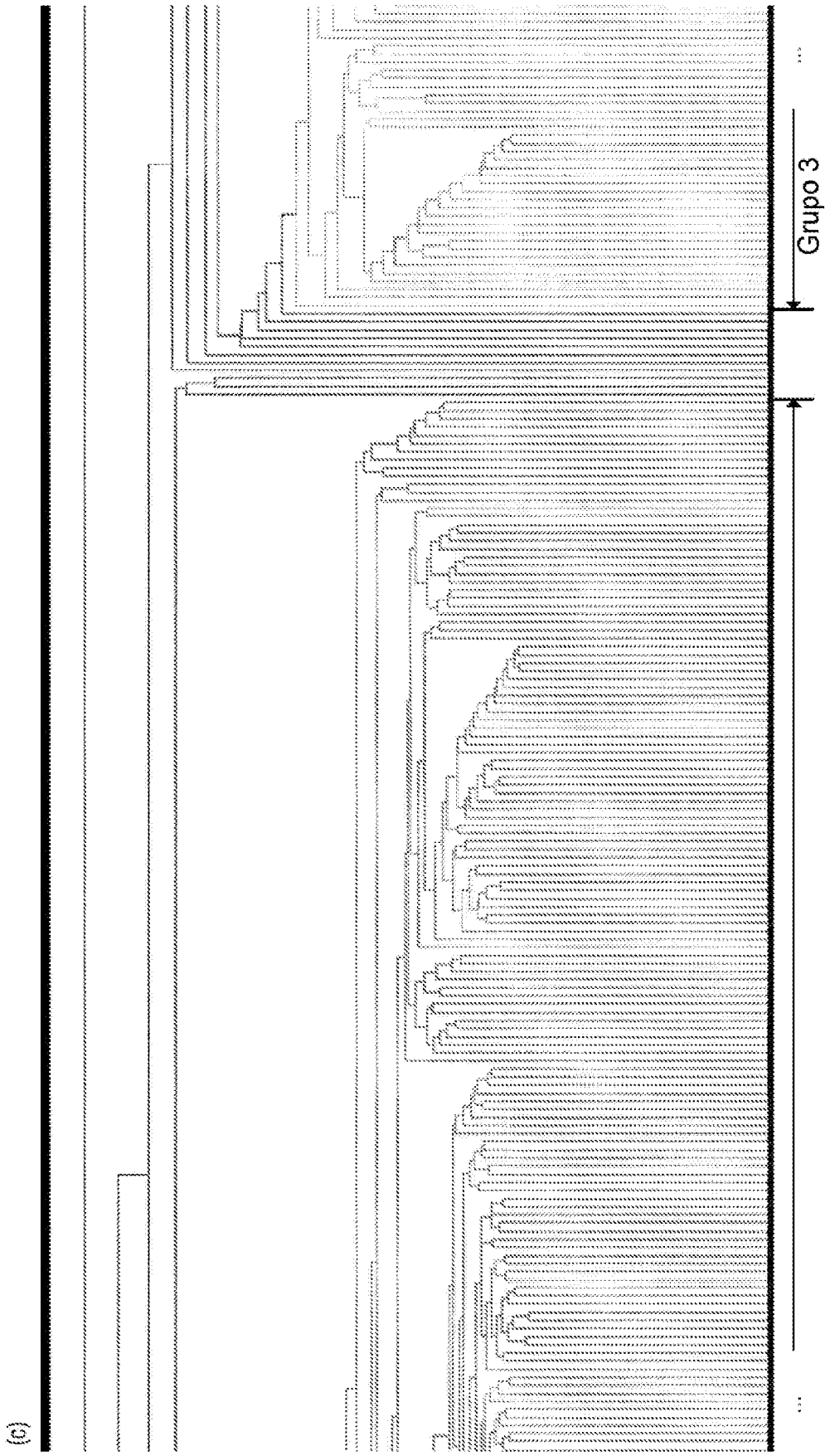


FIG.18 (continuación)

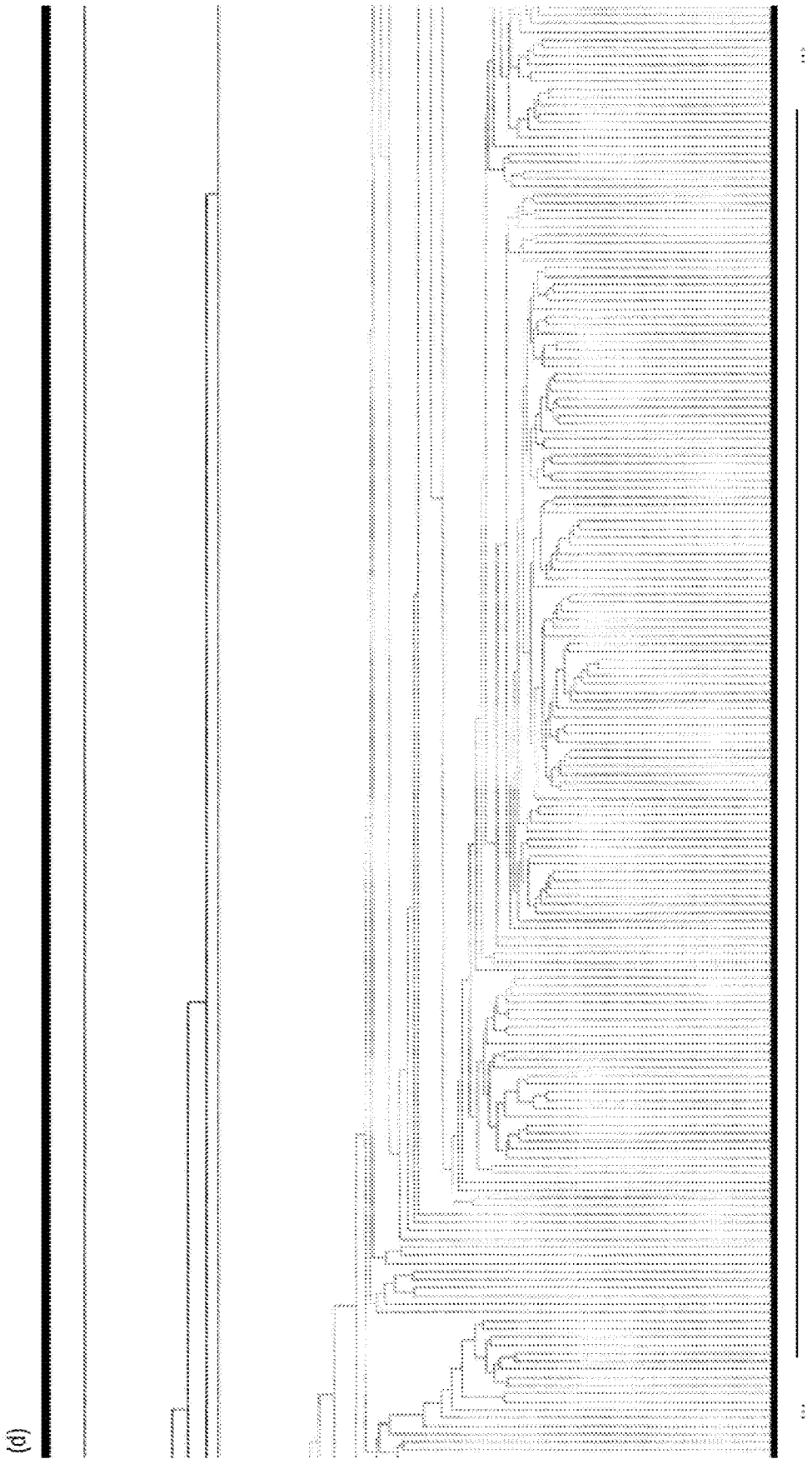


FIG.18 (continuación)

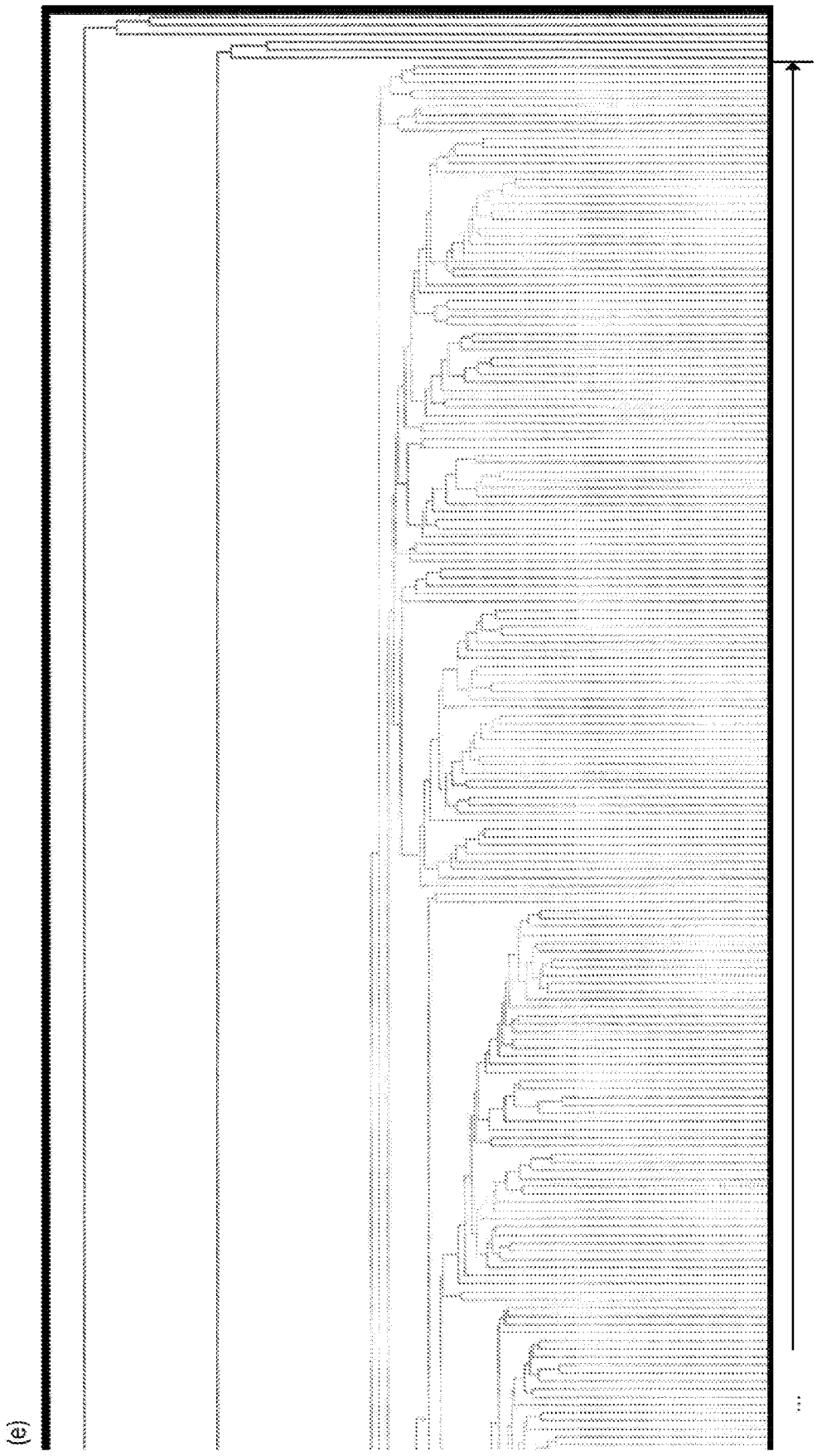
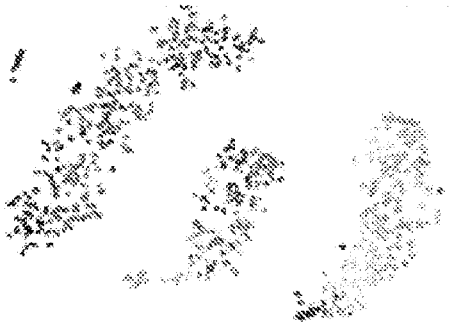


FIG.18 (continuación)

(a)

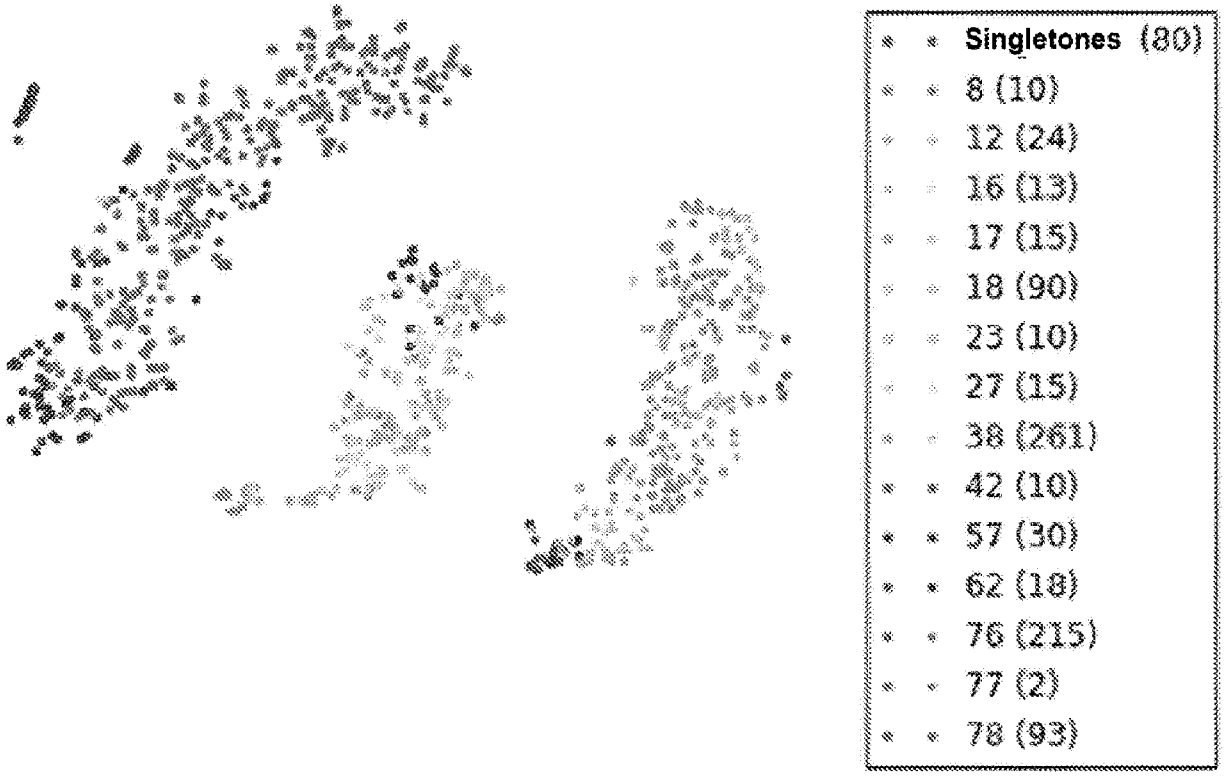
Resultado de agrupación usando un umbral de 5.00



- Singletones (205)
- 12 (6)
- 16 (24)
- 20 (33)
- 26 (36)
- 27 (2)
- 31 (2)
- 32 (36)
- 37 (2)
- 41 (11)
- 42 (2)
- 46 (5)
- 47 (2)
- 48 (2)
- 52 (7)
- 53 (2)
- 54 (8)
- 55 (5)
- 59 (2)
- 65 (2)
- 73 (8)
- 77 (28)
- 88 (282)
- 90 (13)
- 93 (12)
- 94 (2)
- 95 (22)
- 97 (2)
- 98 (12)
- 112 (2)
- 131 (2)
- 143 (2)
- 166 (2)
- 167 (9)
- 181 (2)
- 184 (2)
- 185 (23)
- 187 (2)
- 192 (8)
- 193 (2)
- 196 (2)
- 198 (19)
- 203 (2)
- 211 (2)
- 213 (8)
- 214 (28)
- 215 (9)
- 216 (5)
- 220 (5)
- 221 (24)
- 224 (2)
- 225 (8)
- 228 (15)
- 229 (2)
- 230 (28)
- 231 (2)
- 238 (2)
- 240 (26)
- 244 (2)
- 248 (50)

FIG. 19

(b) **Resultado de agrupación usando un umbral de 10.000**



(c) **Resultado de agrupación usando un umbral de 15.000**

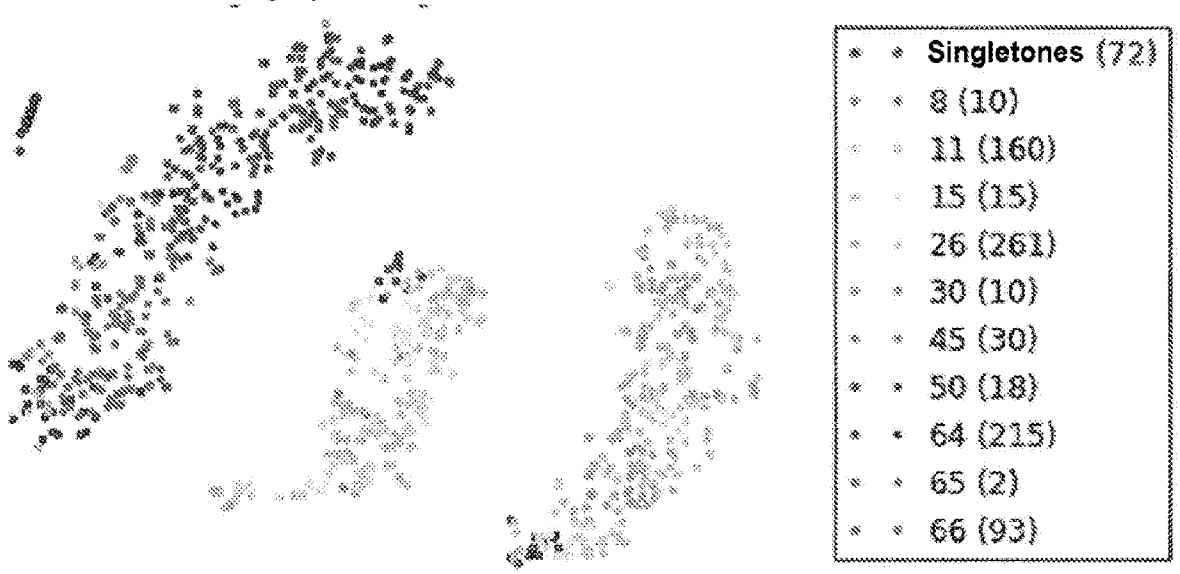
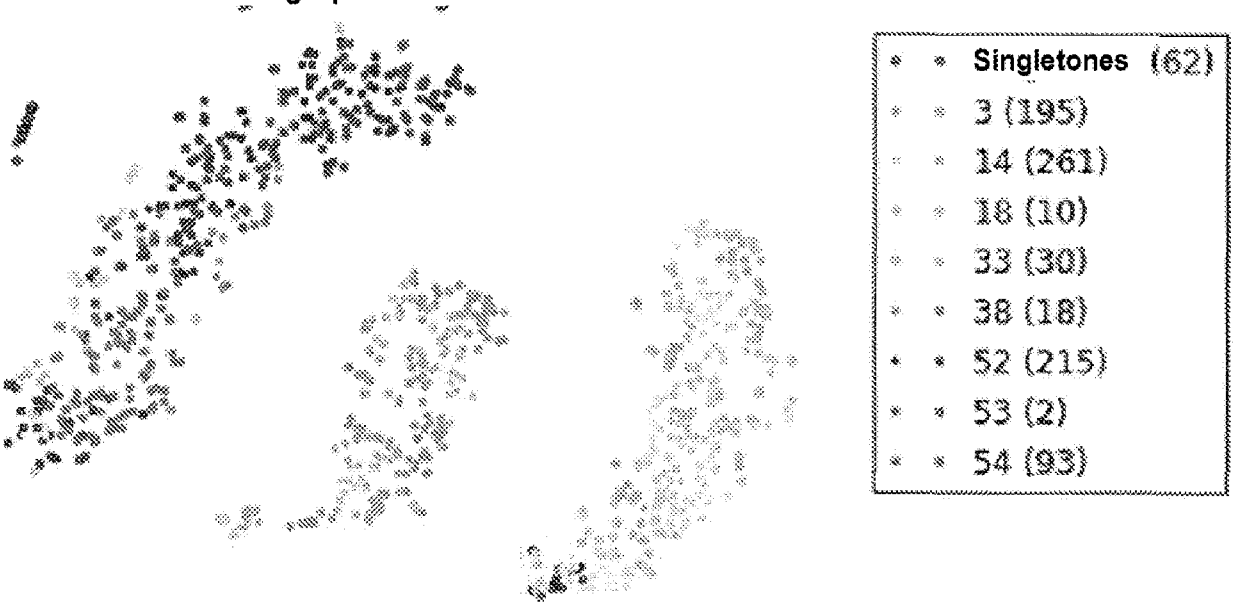


FIG. 19 (Continuación)

(d) **Resultado de agrupación usando un umbral de 20.000**



(e) **Resultado de agrupación usando un umbral de 25.000**

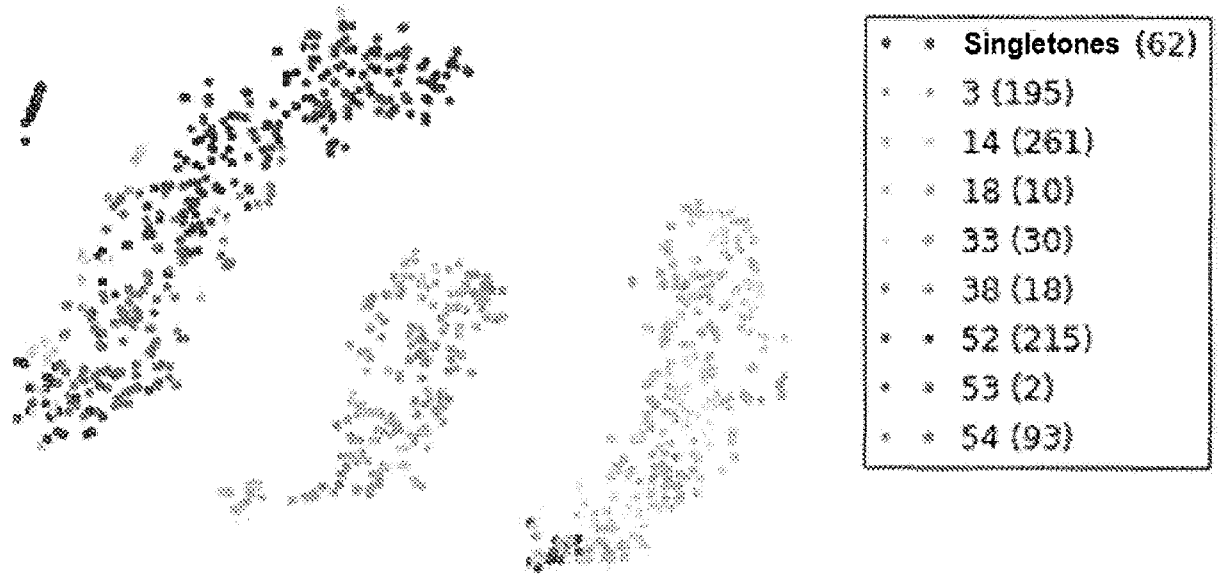
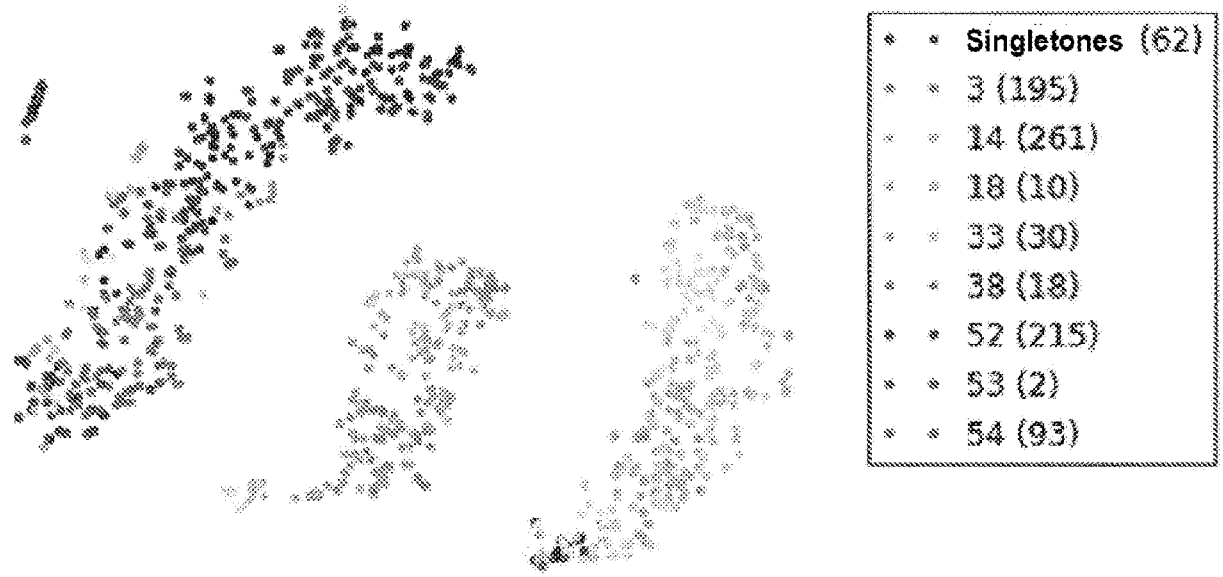


FIG. 19 (Continuación)

(f) **Resultado de agrupación usando un umbral de 30.000**



(g) **Resultado de agrupación usando un umbral de 35.000**

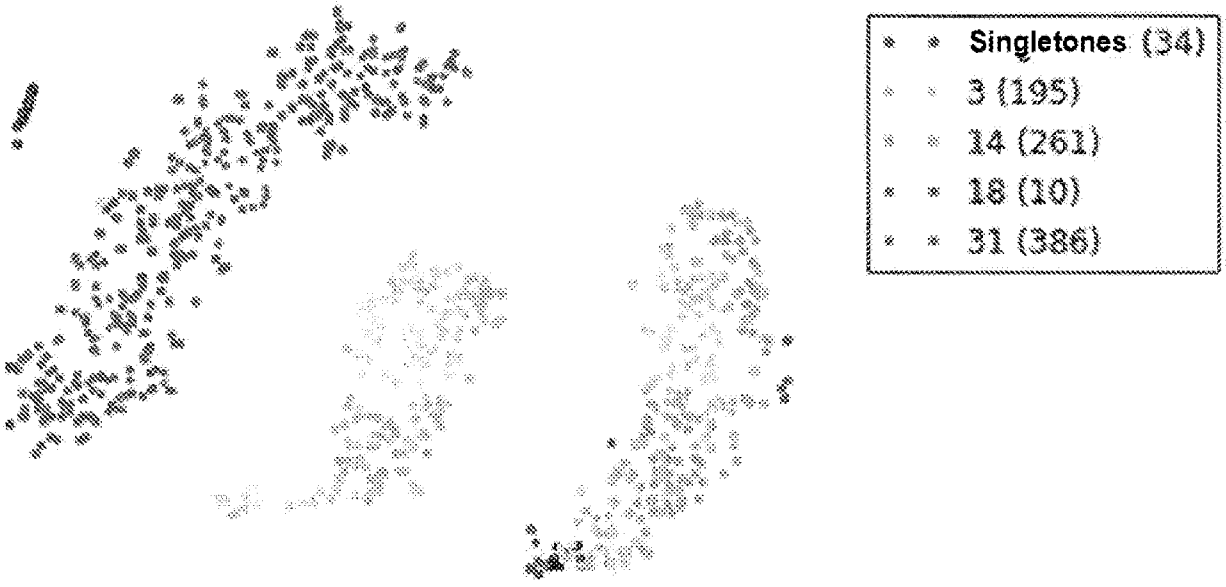
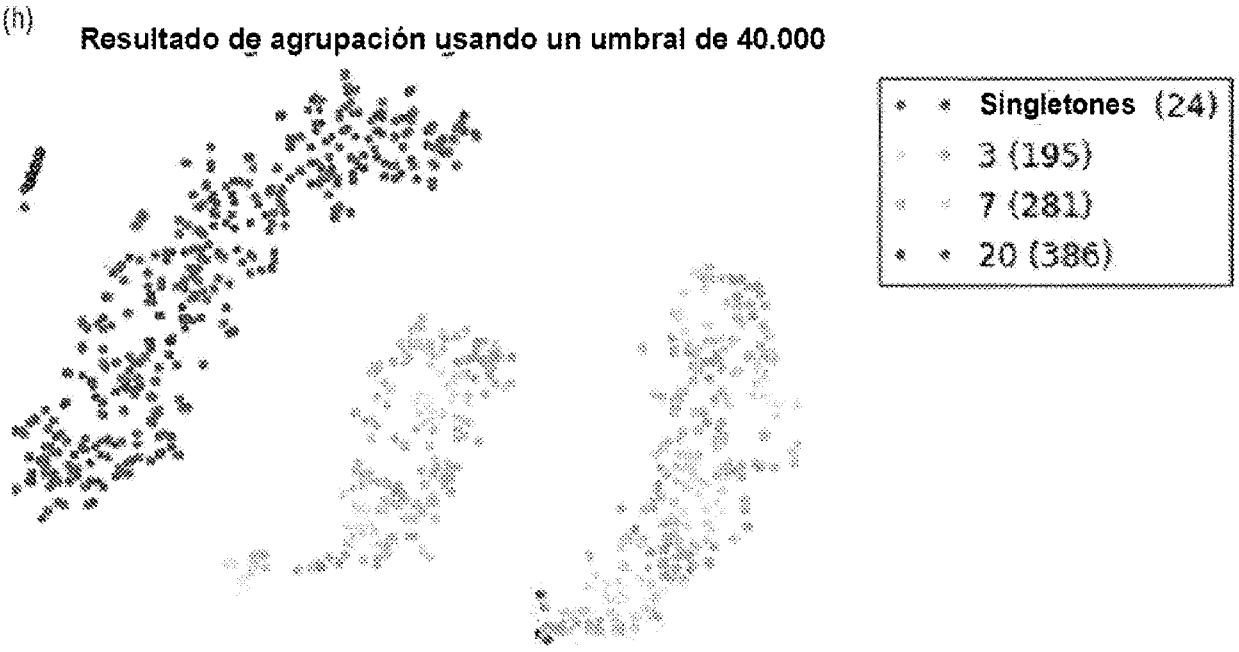


FIG. 19 (Continuación)

(h)



(i)

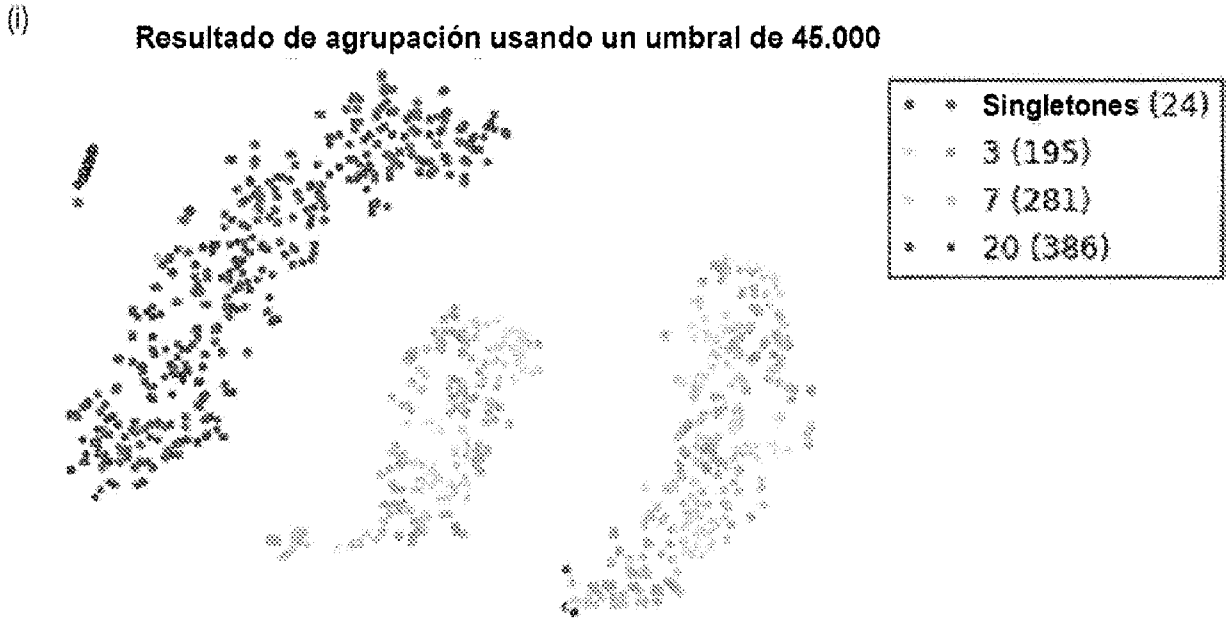
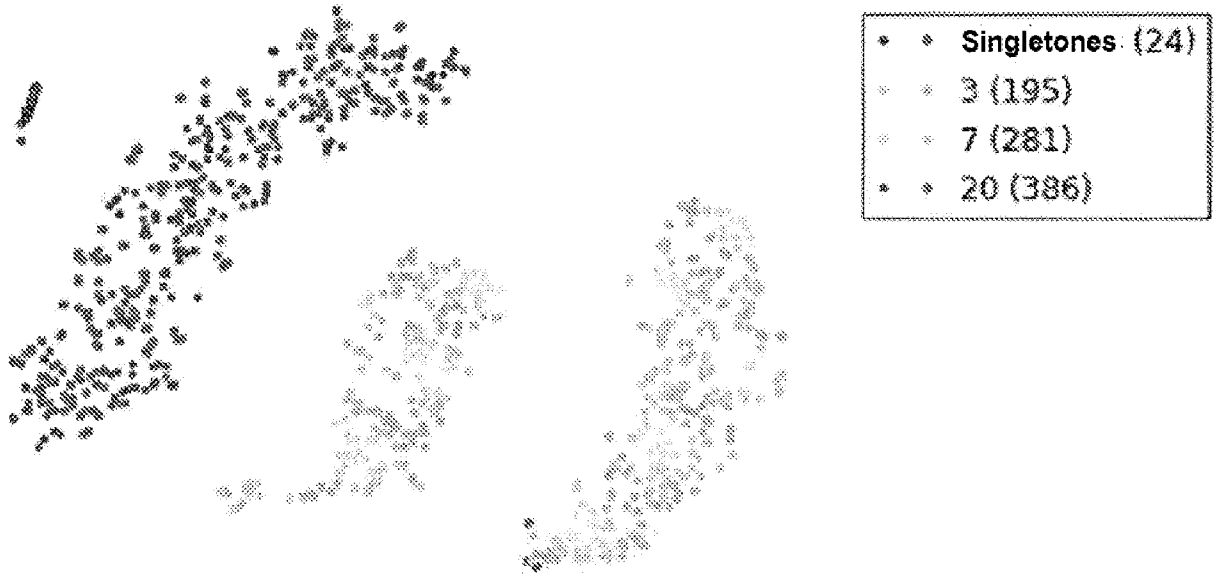


FIG. 19 (Continuación)

(j) Resultado de agrupación usando un umbral de 50.000



(k) Resultado de agrupación usando un umbral de 55.000

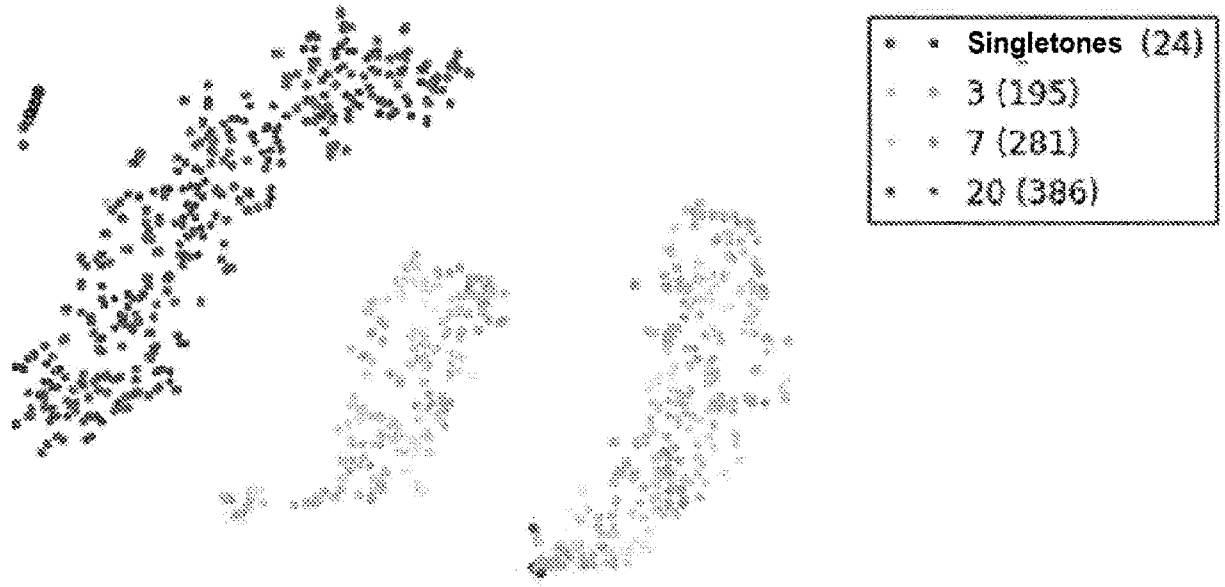
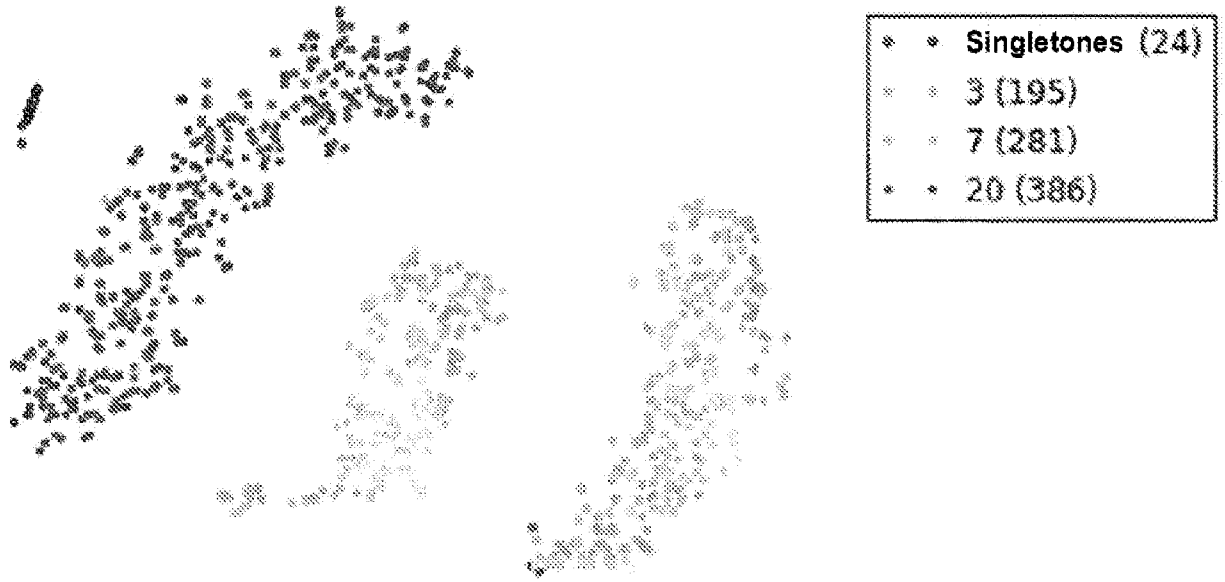


FIG. 19 (Continuación)

(l) **Resultado de agrupación usando un umbral de 60.000**



(m) **Resultado de agrupación usando un umbral de 65.000**

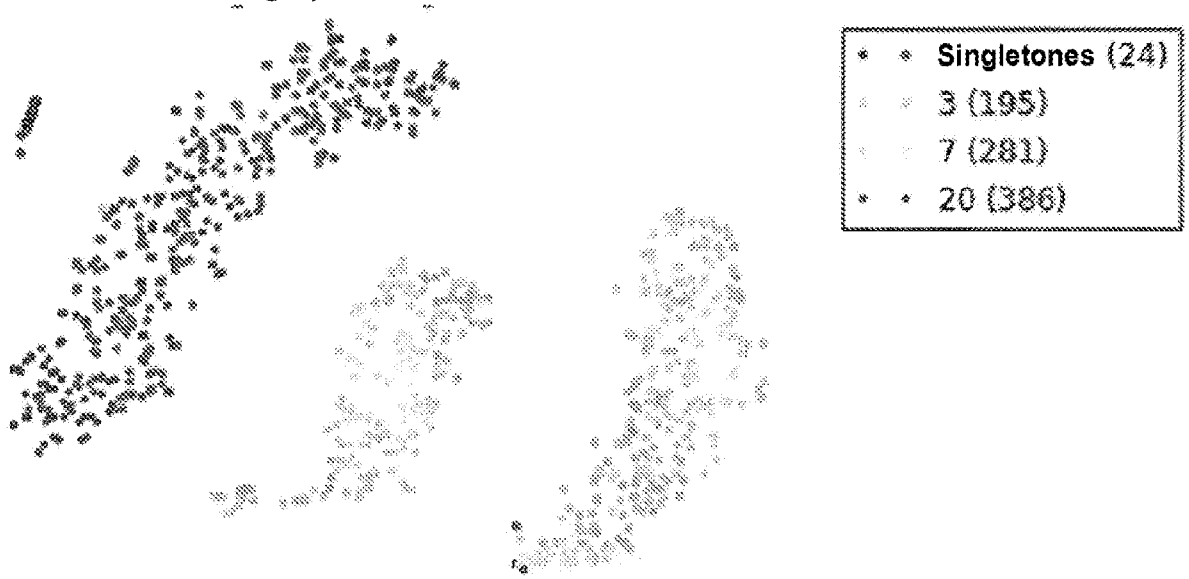
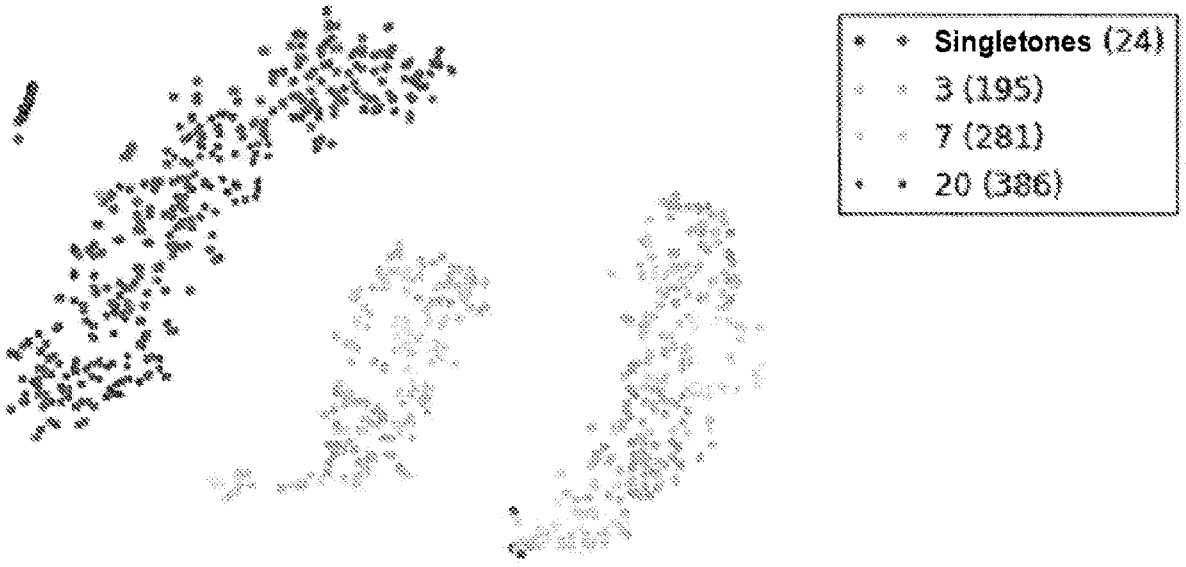


FIG. 19 (Continuación)

(n) **Resultado de agrupación usando un umbral de 70.000**



(o) **Resultado de agrupación usando un umbral de 75.000**

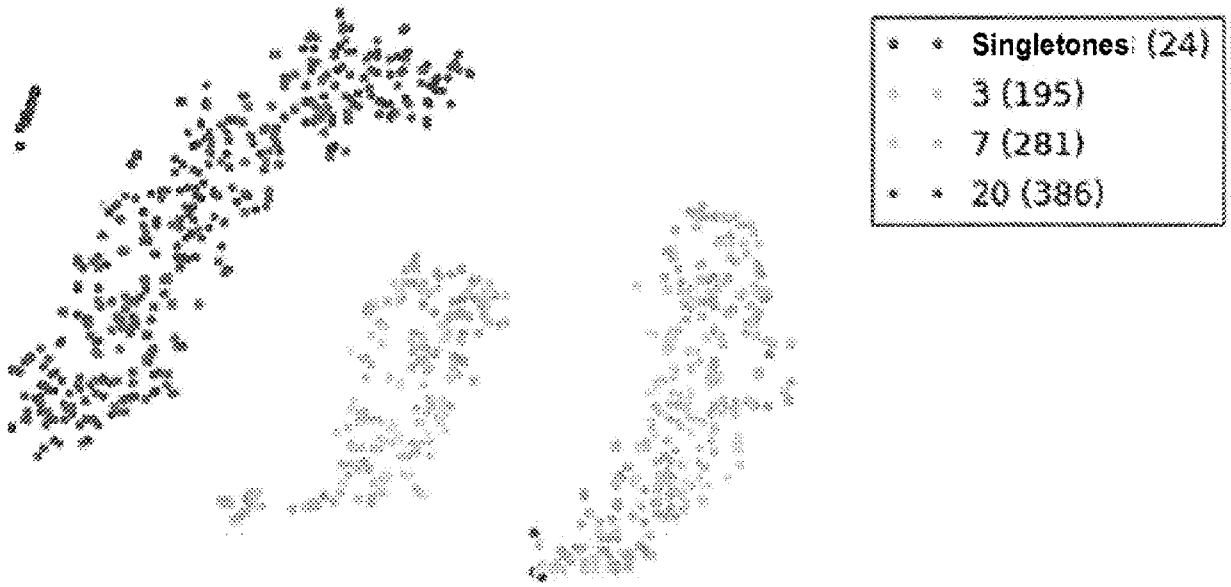
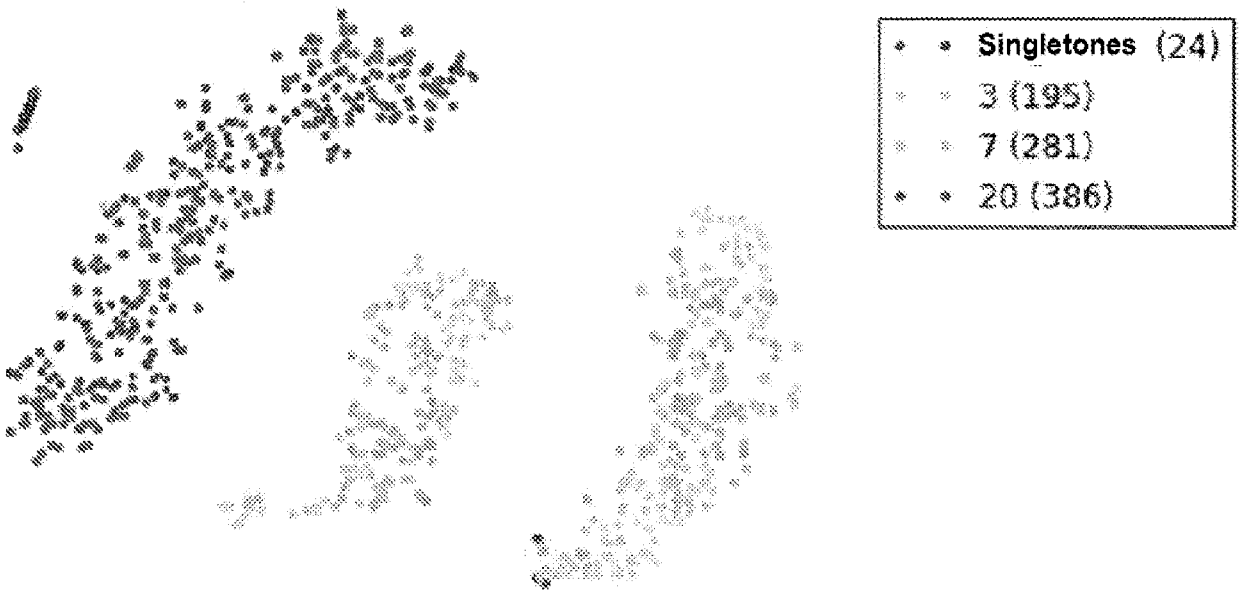


FIG. 19 (Continuación)

(p) **Resultado de agrupación usando un umbral de 80.000**



(q) **Resultado de agrupación usando un umbral de 85.000**

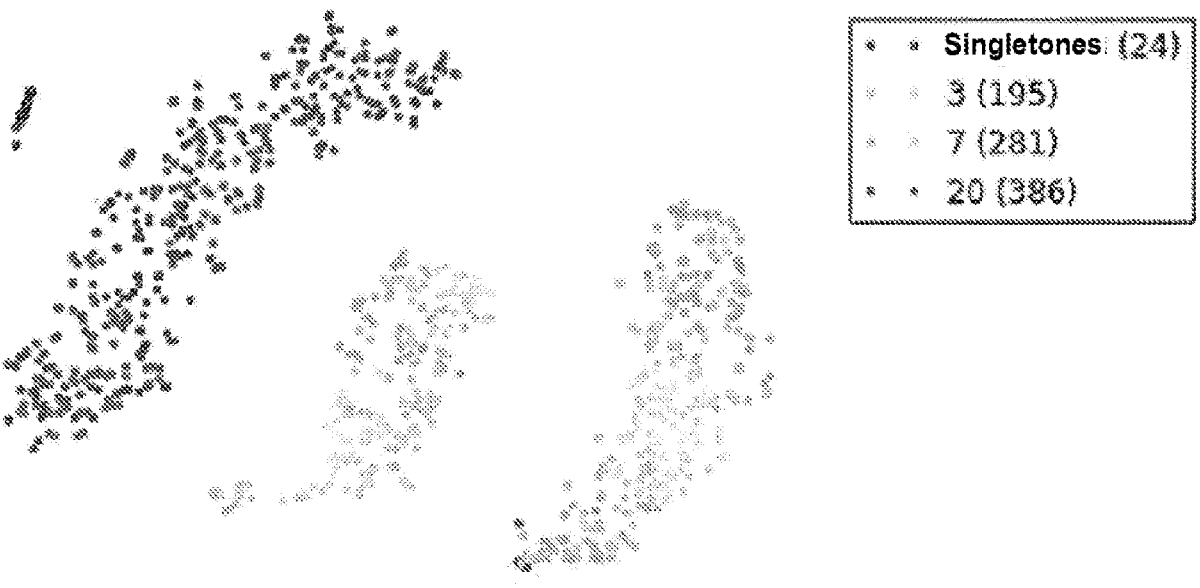
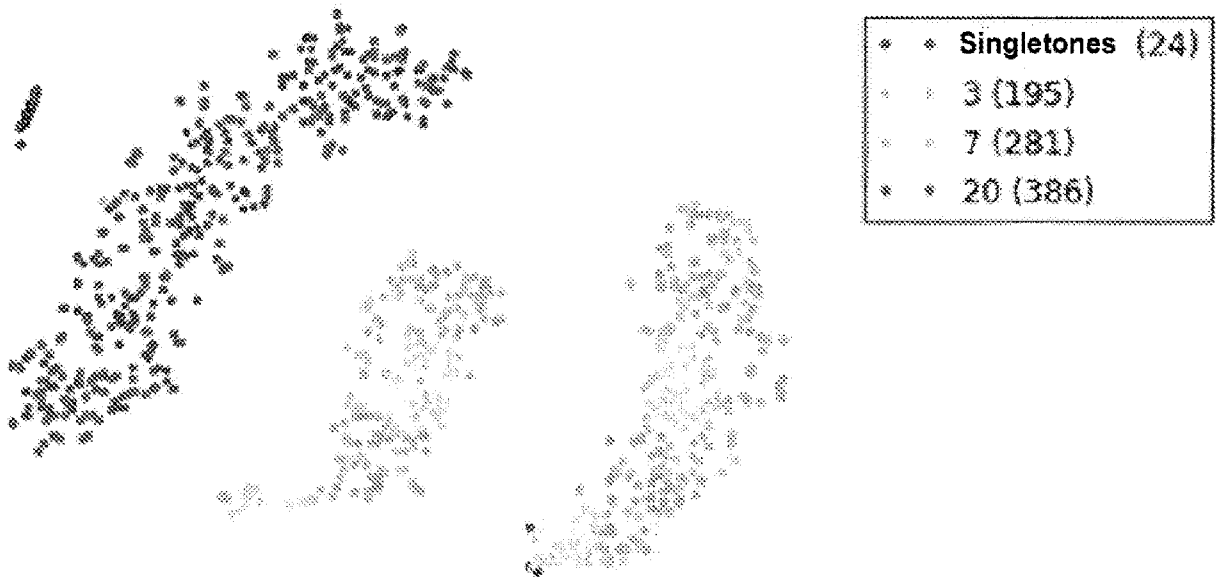


FIG. 19 (Continuación)

(r) **Resultado de agrupación usando un umbral de 90.000**



(s) **Resultado de agrupación usando un umbral de 95.000**

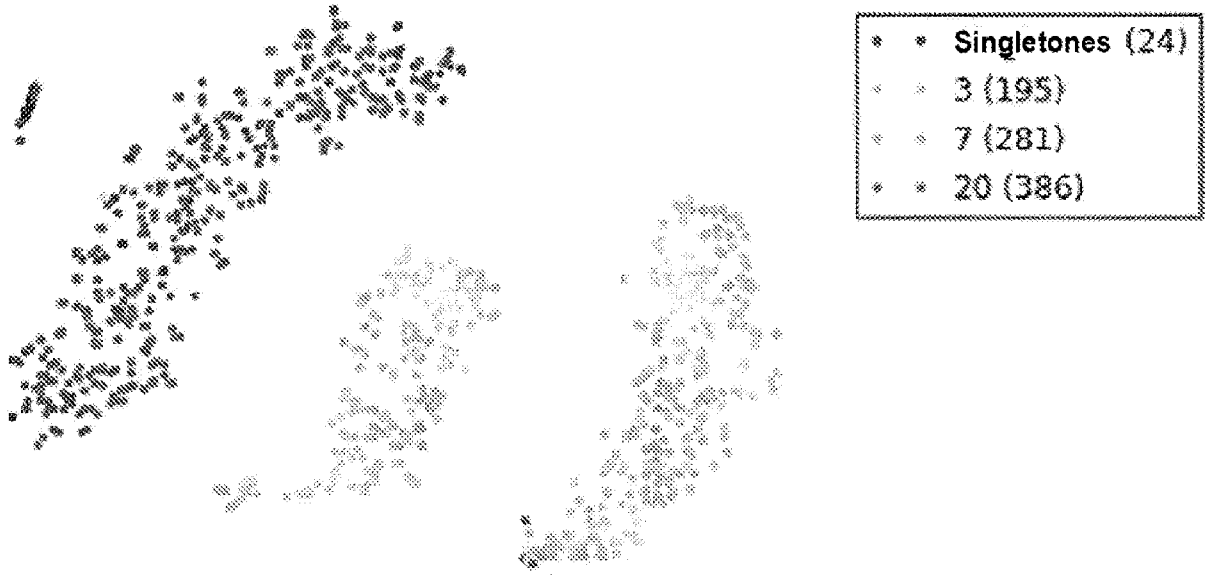


FIG. 19 (Continuación)

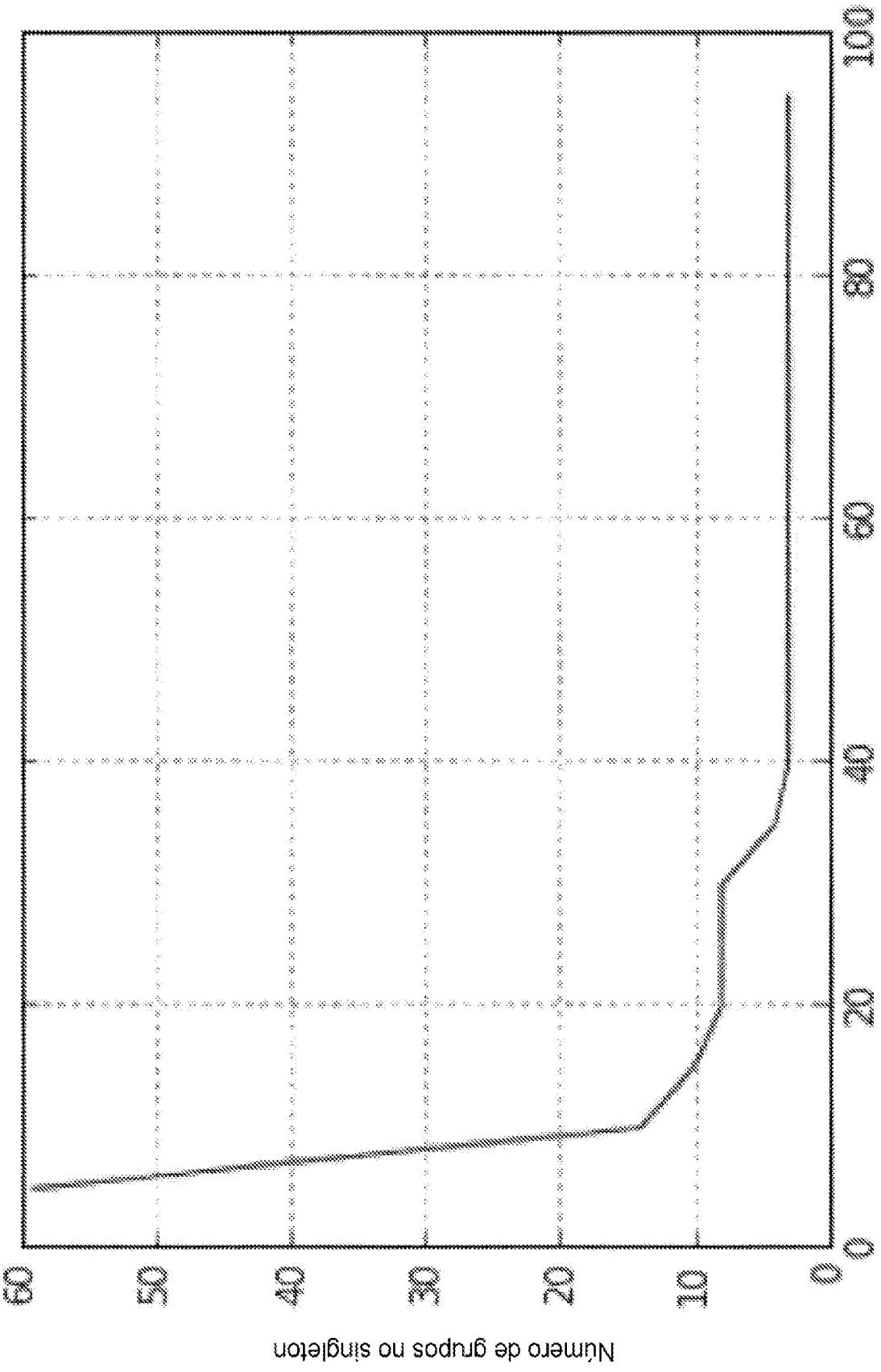

Número de grupos no singleton
0
10
20
30
40
50
60
0
20
40
60
80
100
Umbrales (-log10(valor p)

FIG. 20

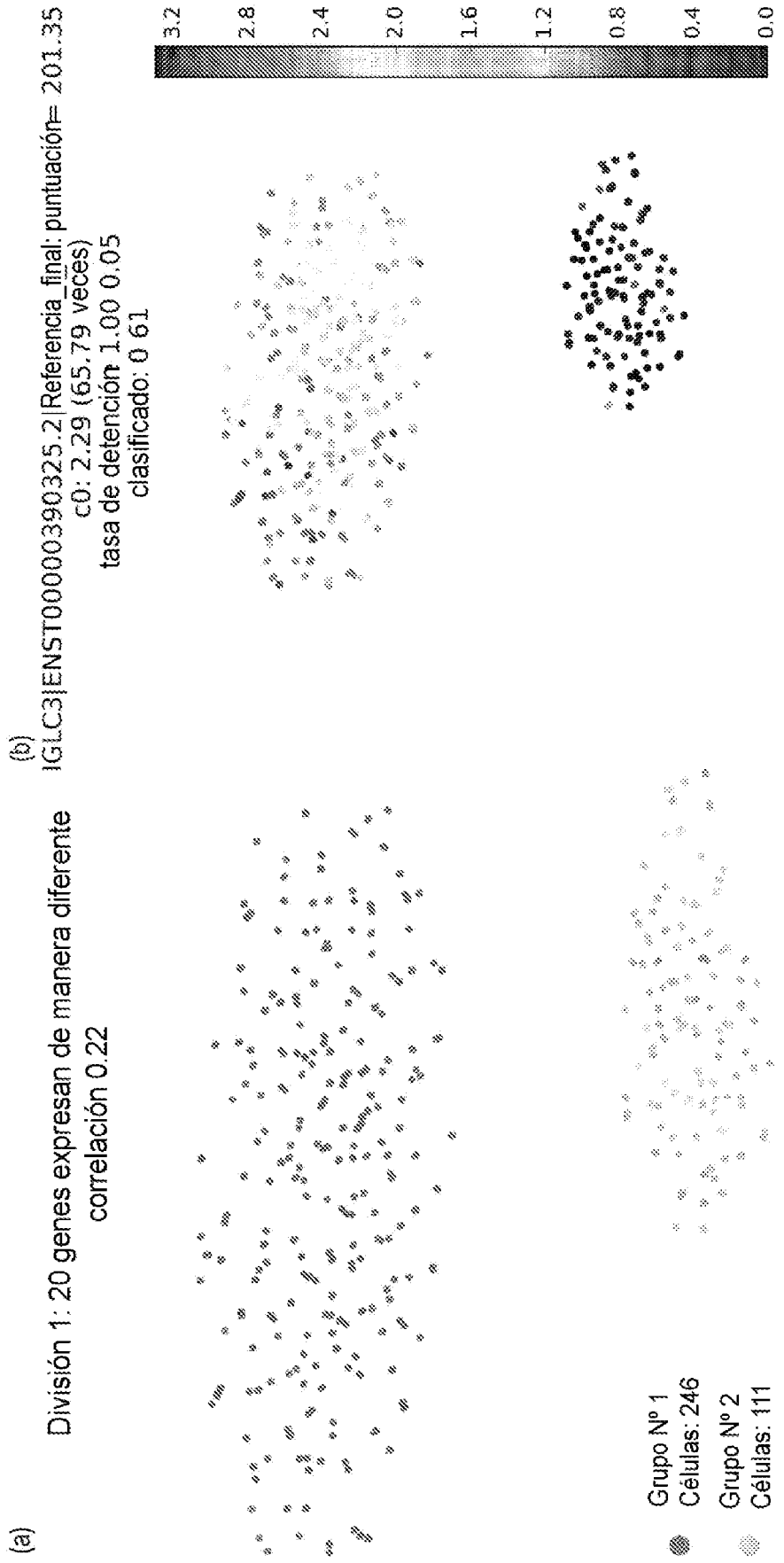

(a)
División 1: 20 genes expresan de manera diferente
correlación 0.22
Grupo Nº 1
Células: 246
Grupo Nº 2
Células: 111
(b)
IGLC3|ENST00000390325.2|Referencia final: puntuación= 201.35
c0: 2.29 (65.79 veces)
tasa de detención 1.00 0.05
clasificado: 0 61
3.2
2.8
2.4
2.0
1.6
1.2
0.8
0.4
0.0

FIG. 21

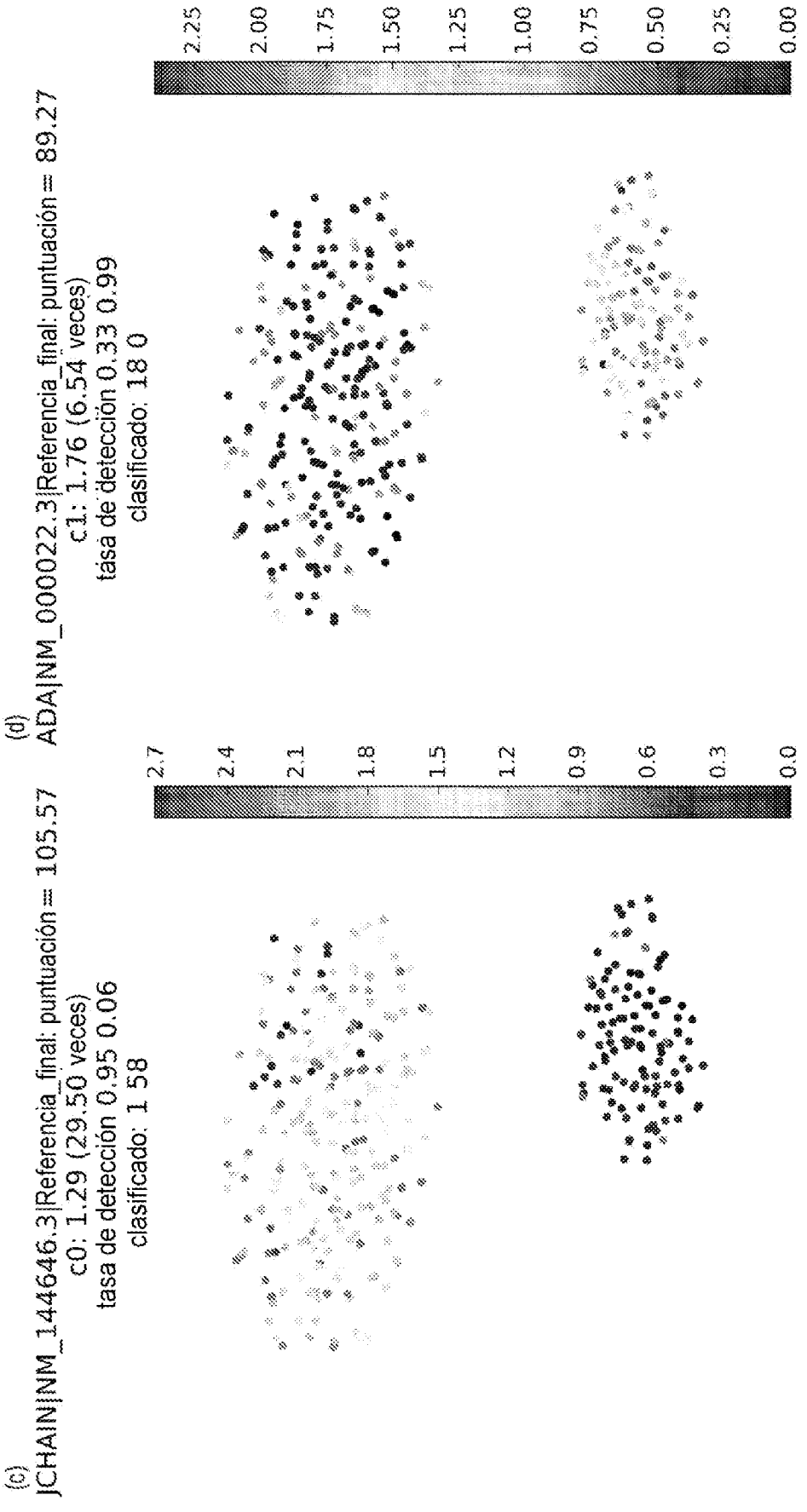


FIG. 21 (Continuación)

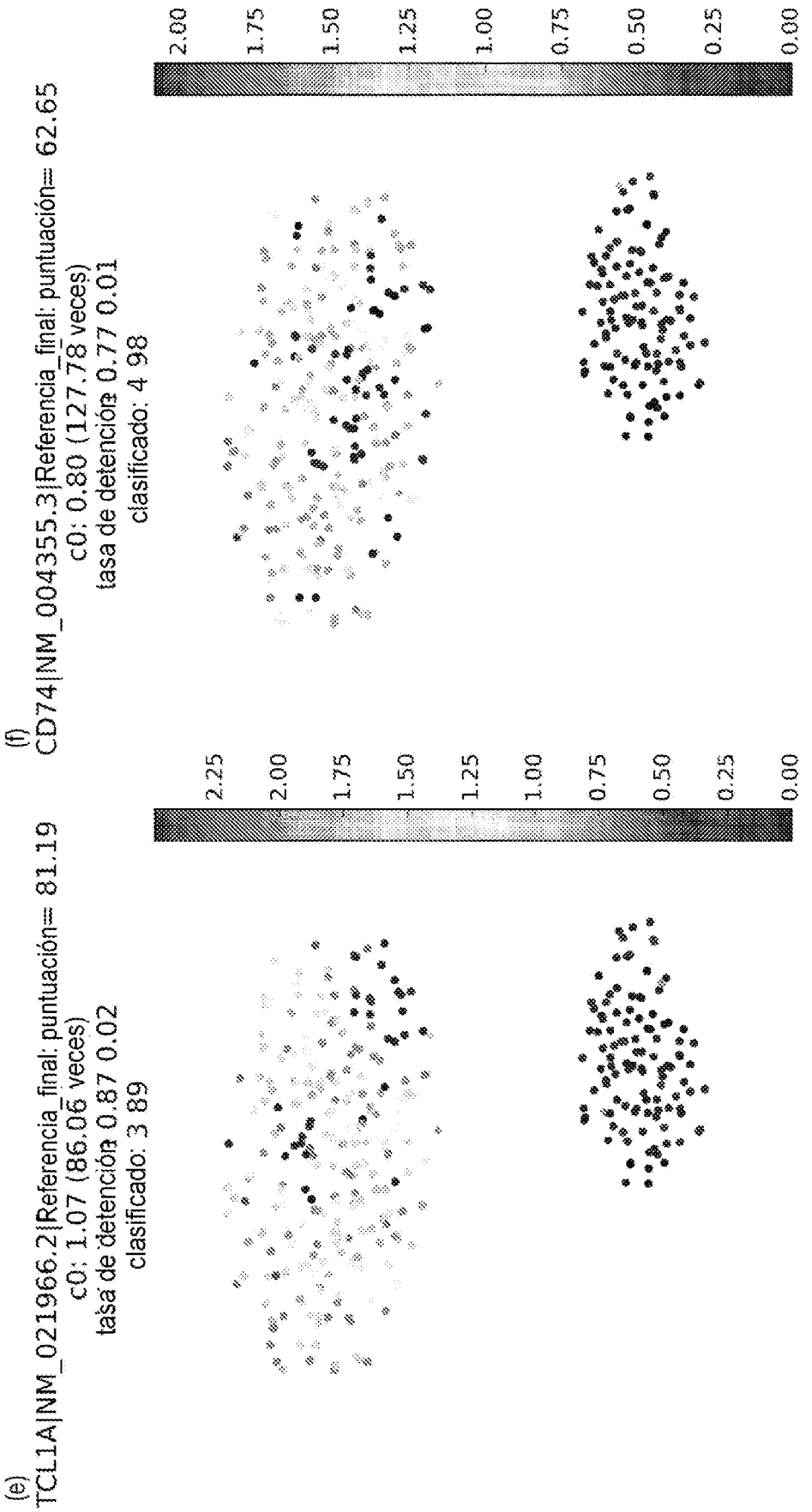


FIG. 21 (Continuación)

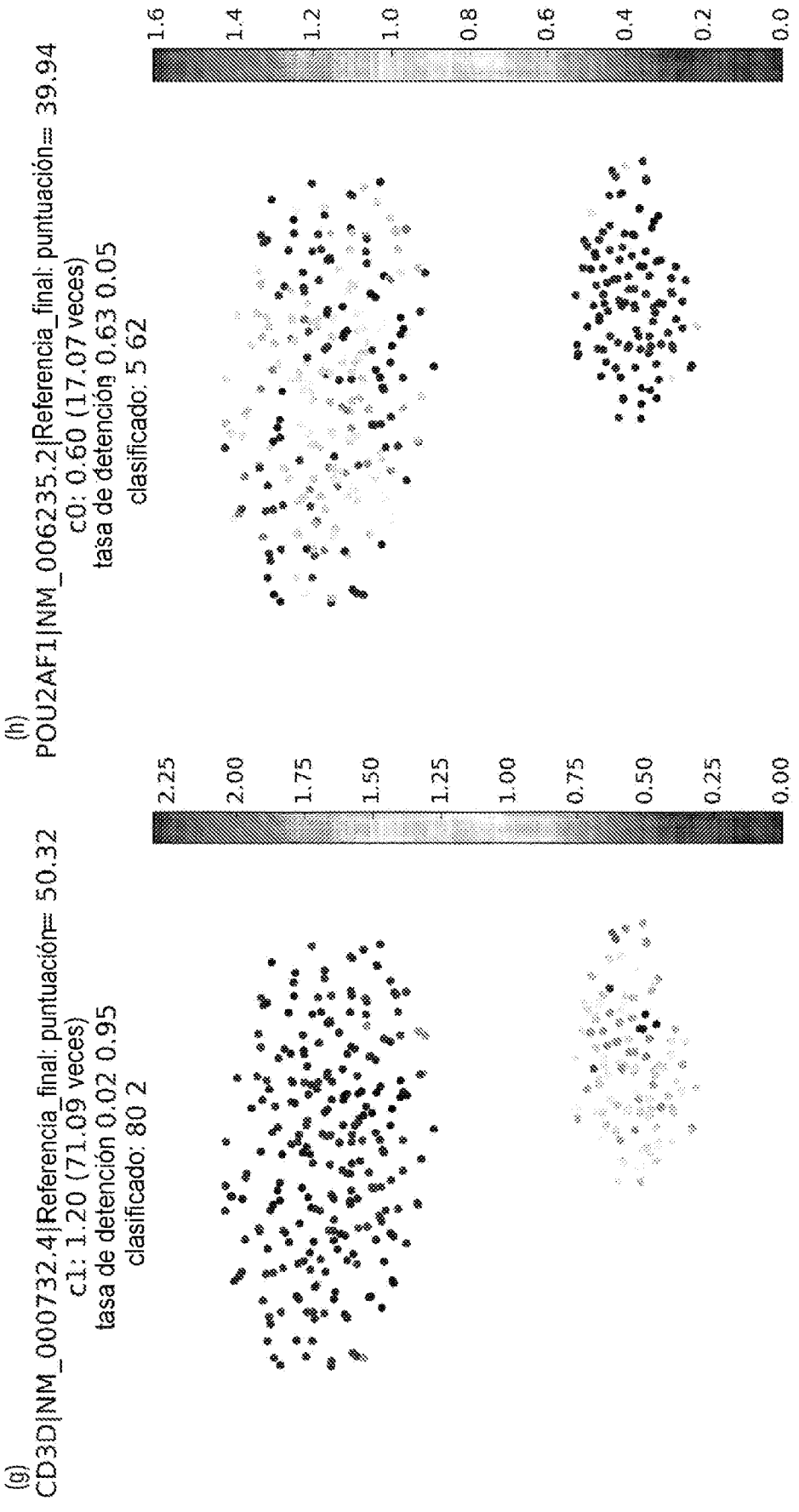


FIG. 21 (Continuación)

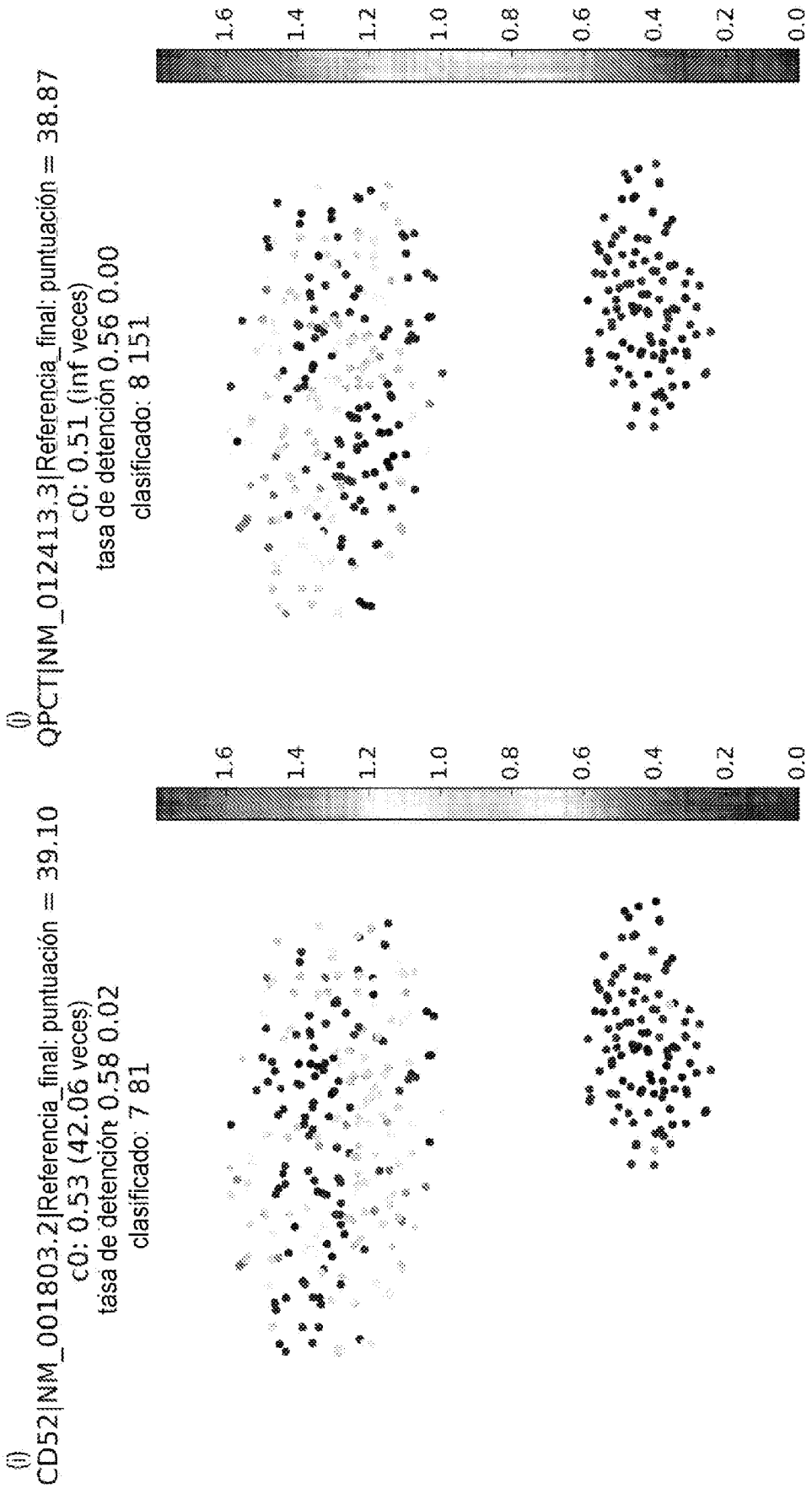


FIG. 21 (Continuación)

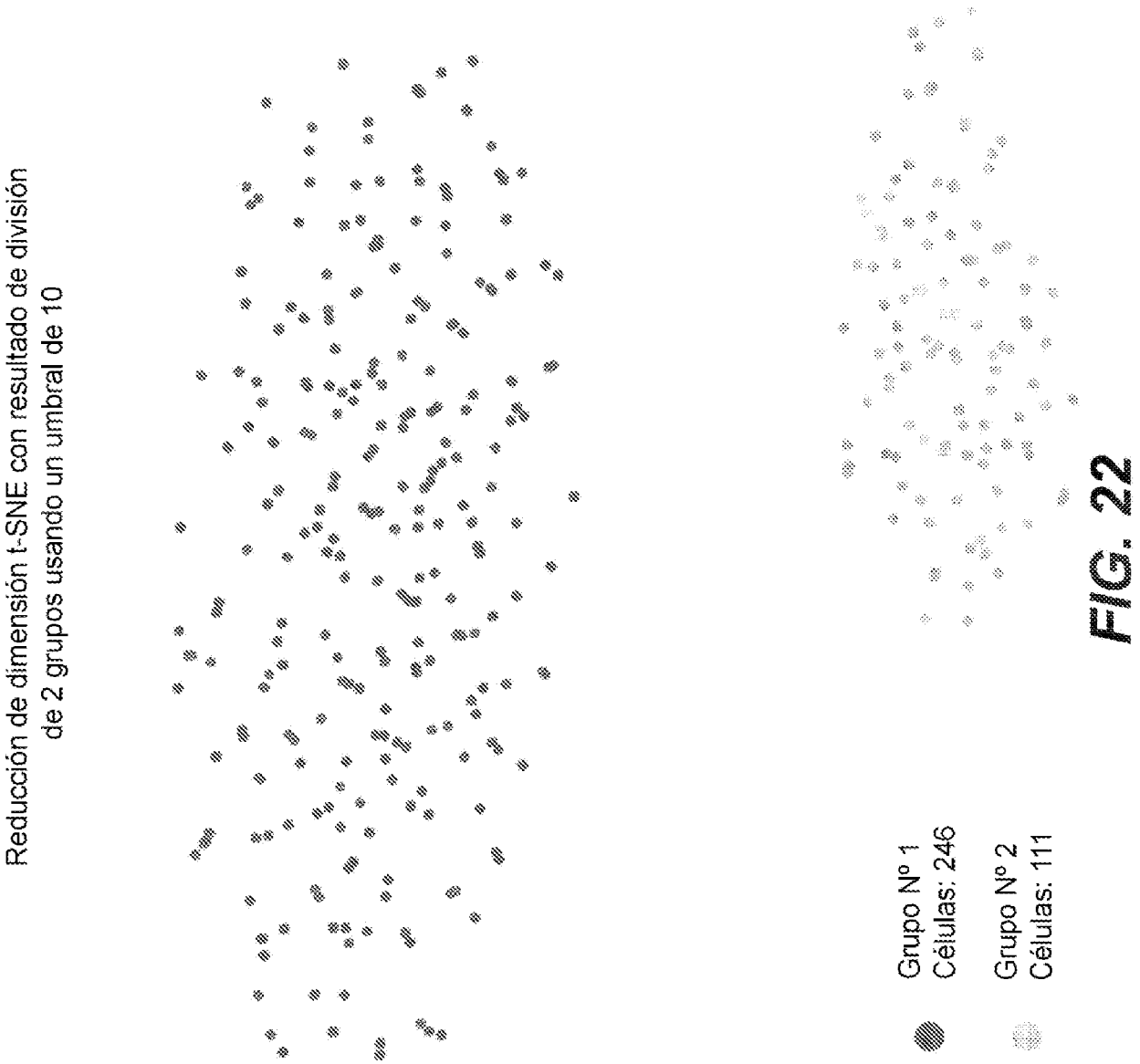

Reducción de dimensión t-SNE con resultado de división
de 2 grupos usando un umbral de 10
Grupo Nº 1
Células: 246
Grupo Nº 2
Células: 111

FIG. 22

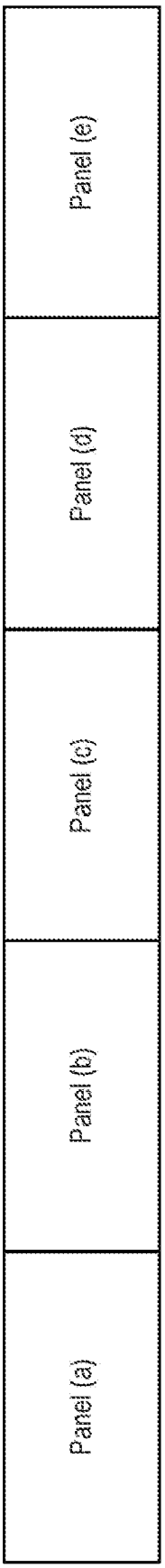


FIG. 23

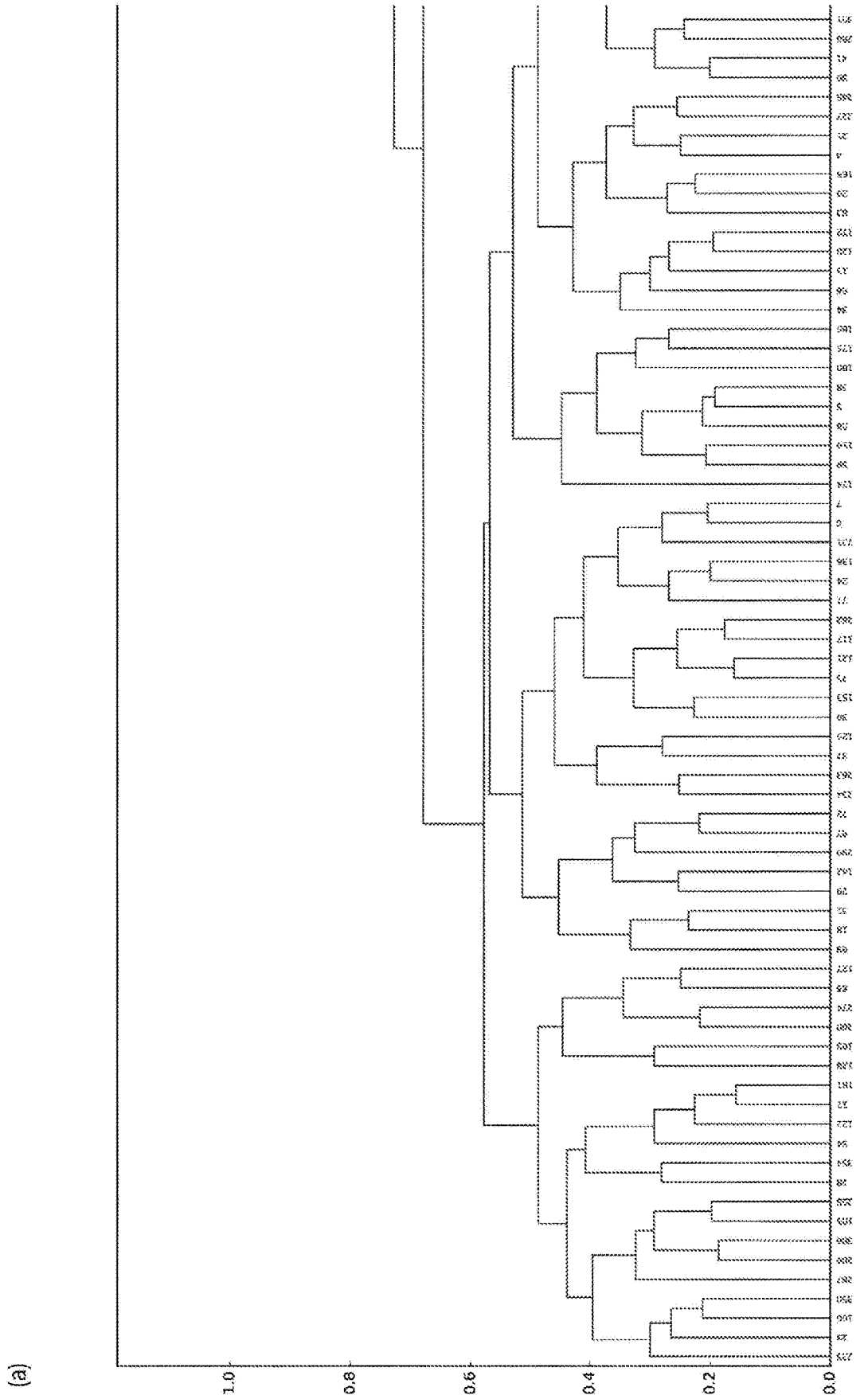


FIG. 23 (Continuación)

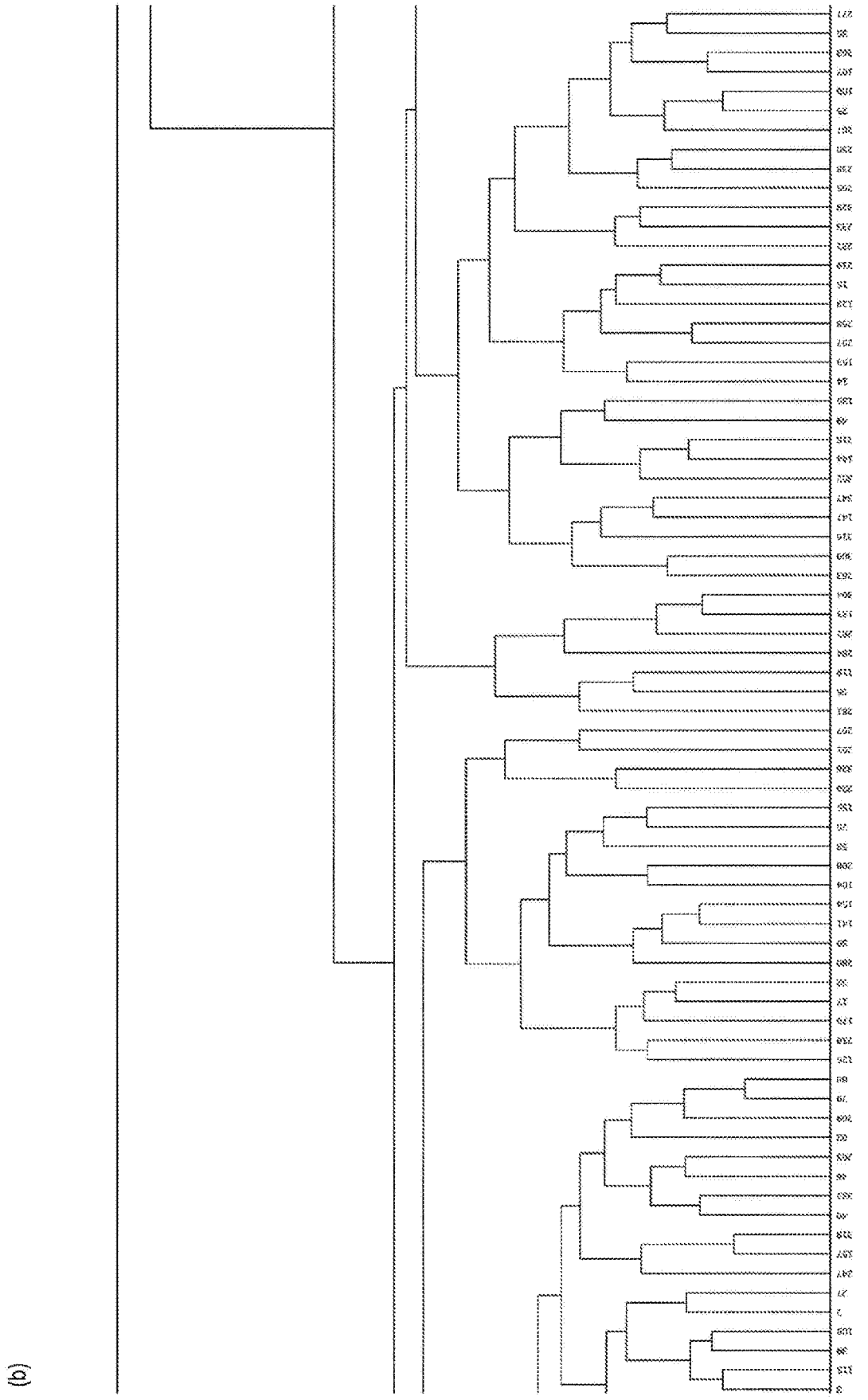


FIG. 23 (Continuación)

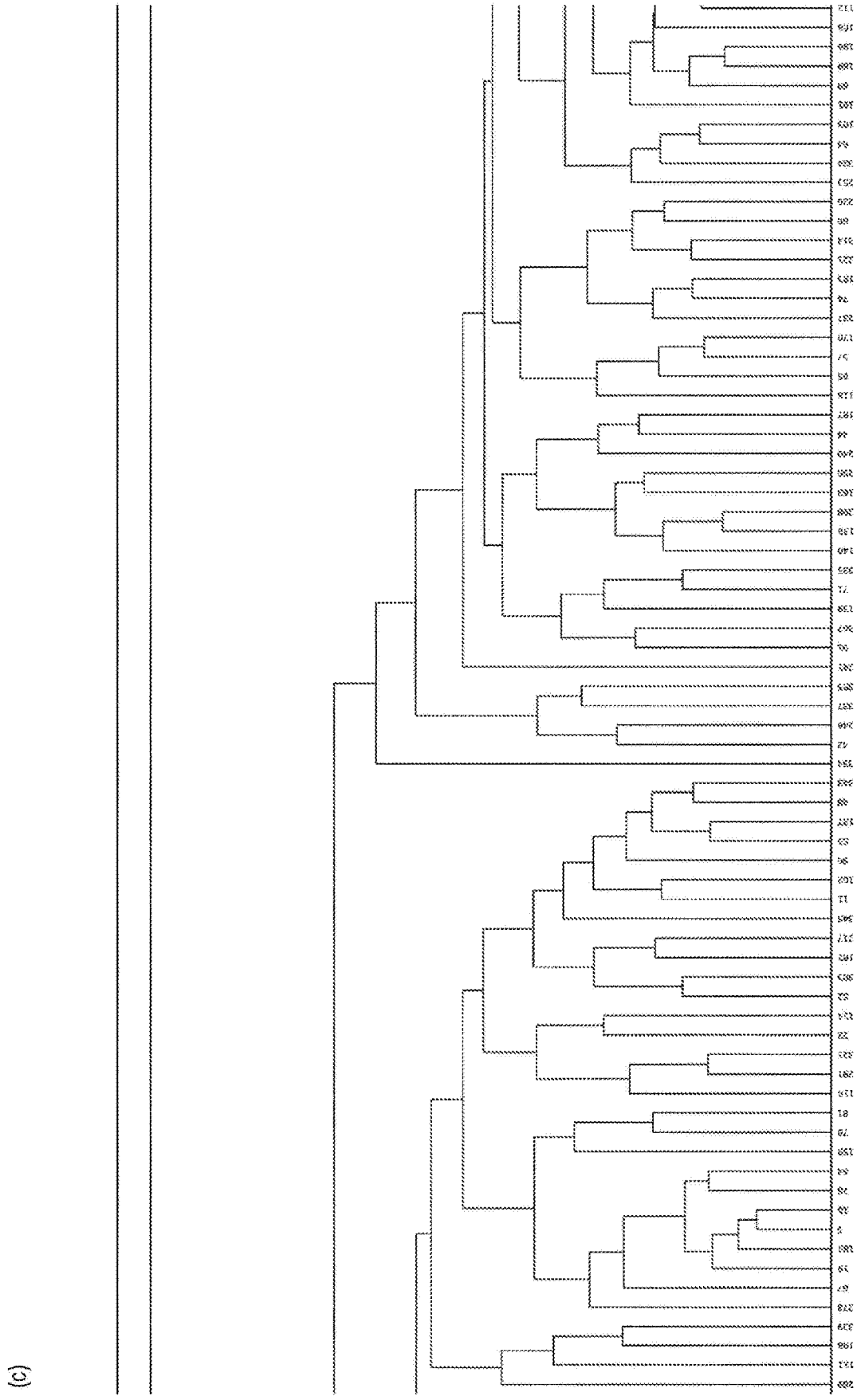


FIG. 23 (Continuación)

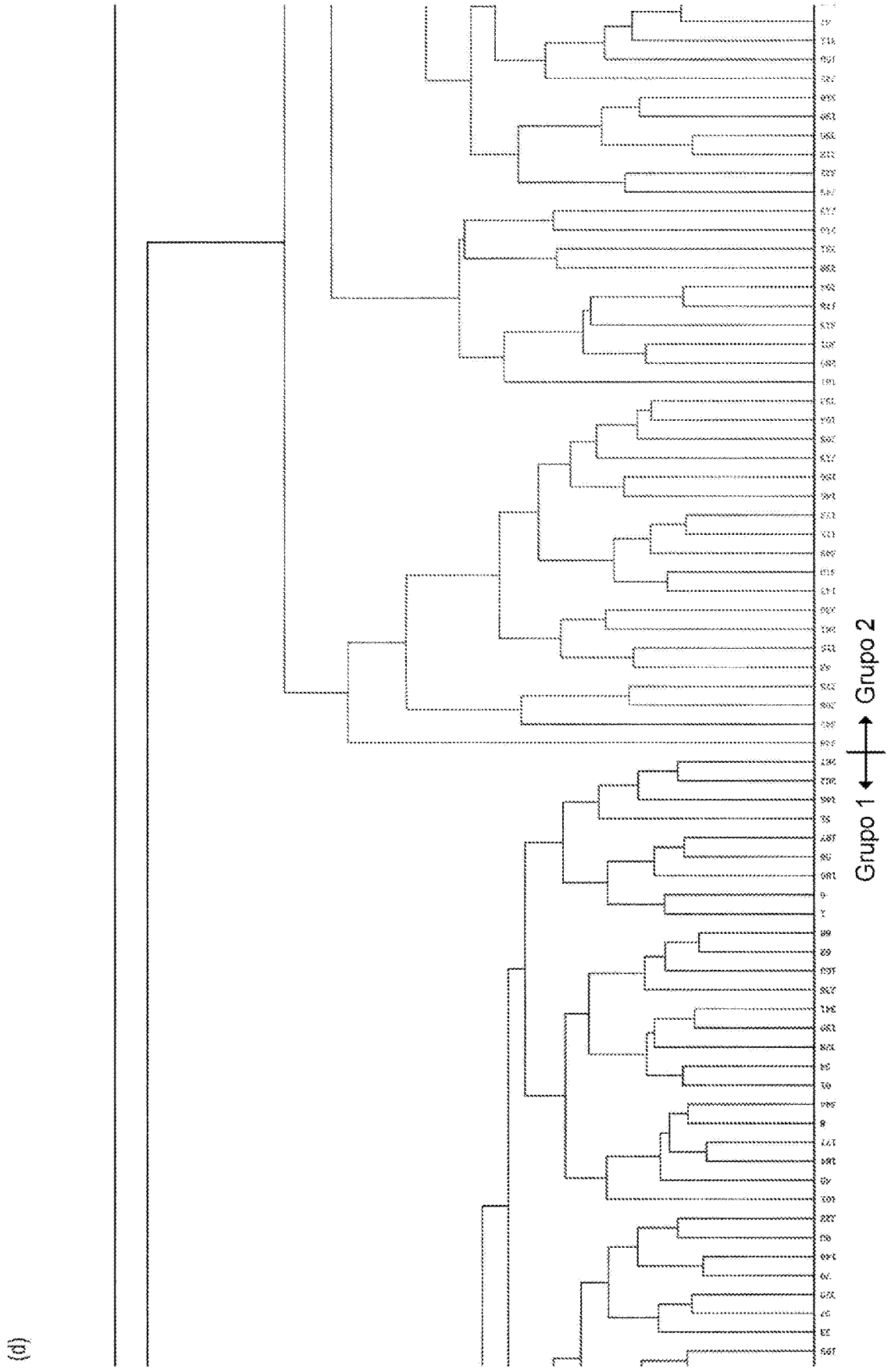


FIG. 23 (Continuación)

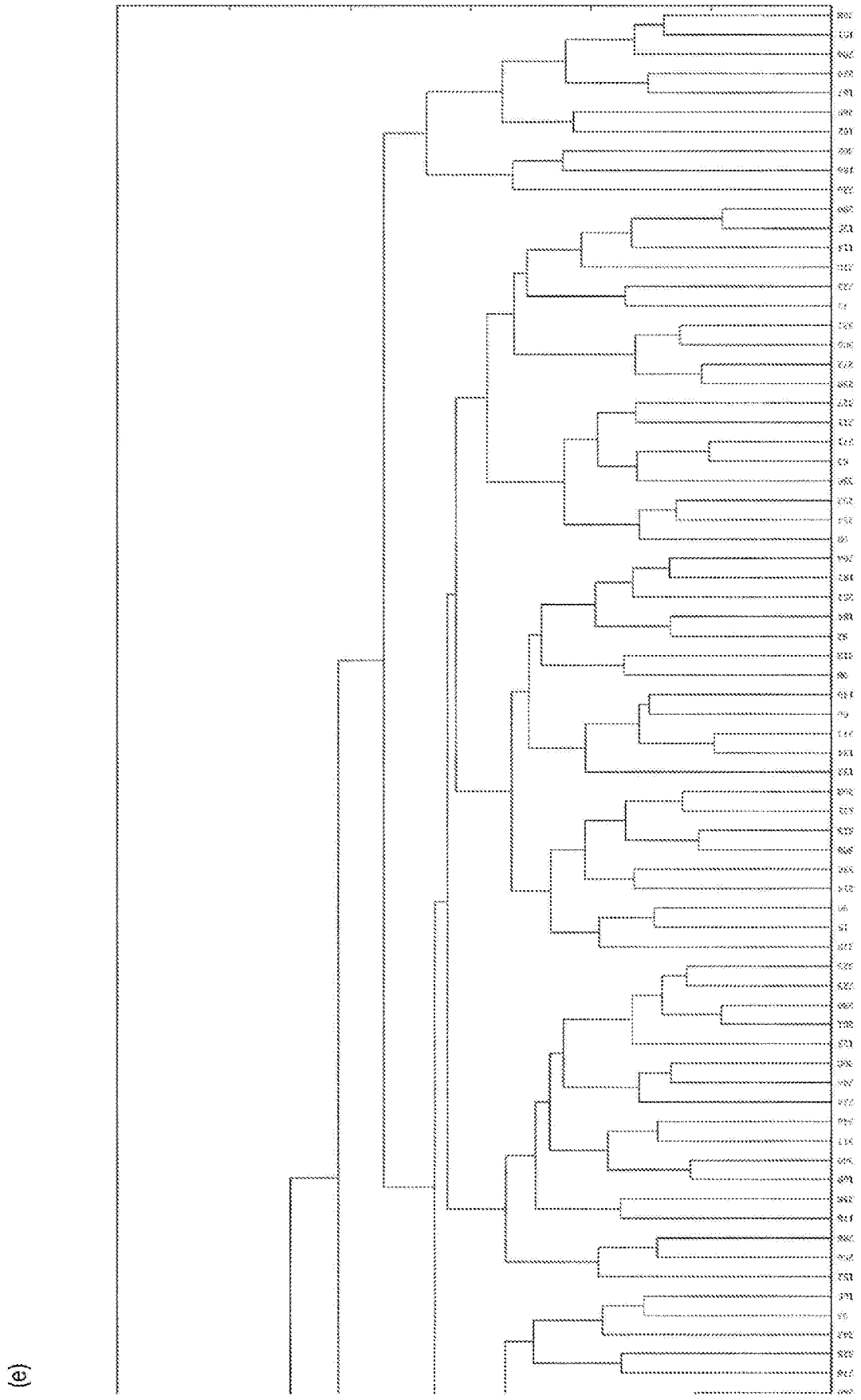


FIG. 23 (Continuación)

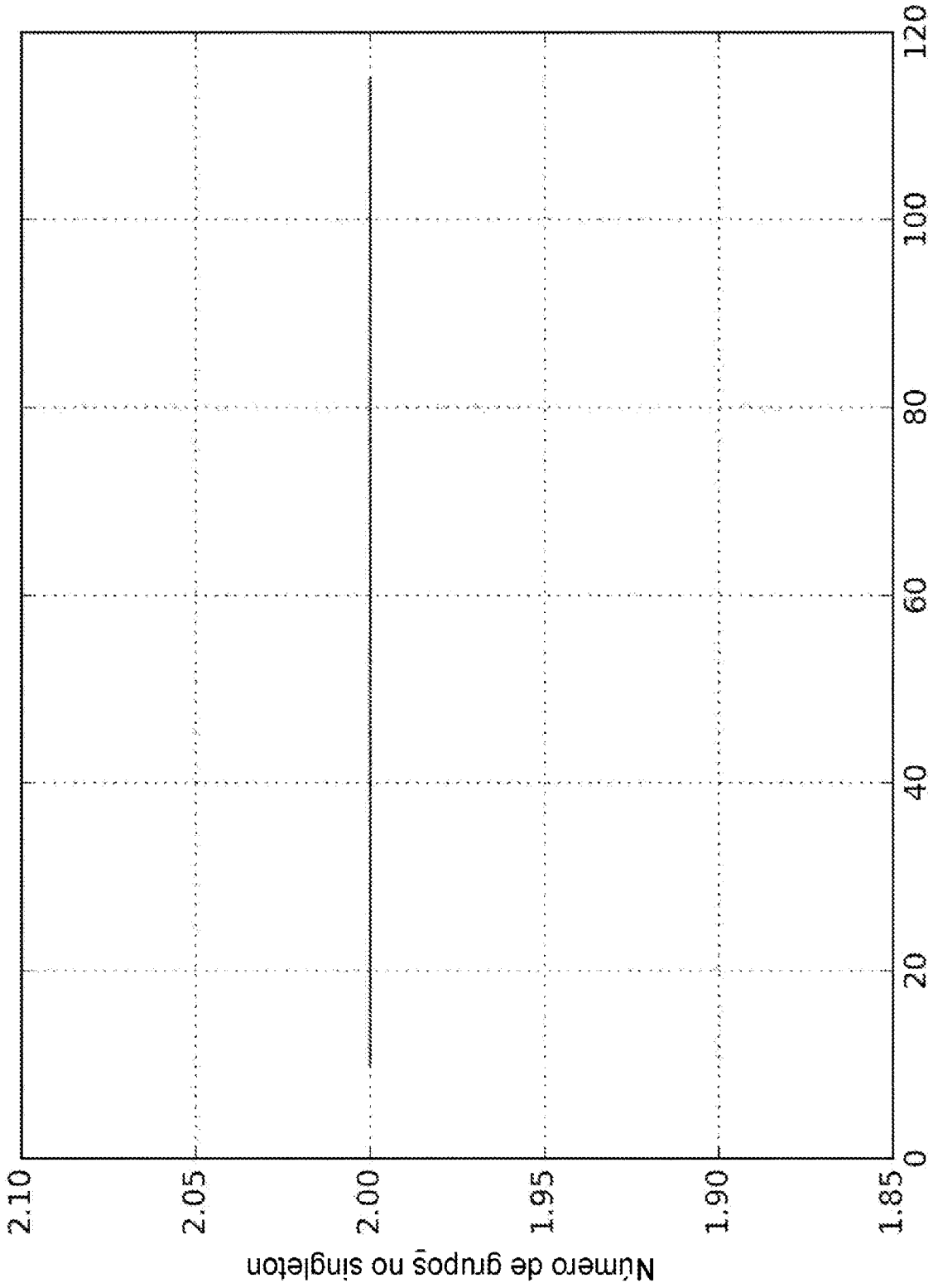

Número de grupos no singleton
2.10
2.05
2.00
1.95
1.90
1.85
0
20
40
60
80
100
120
Umbrales (-log10(valor p))

FIG. 24