

19



OFICINA ESPAÑOLA DE
PATENTES Y MARCAS

ESPAÑA



11 Número de publicación: **2 968 290**

51 Int. Cl.:

C12N 15/11 (2006.01)

C12Q 1/68 (2008.01)

C40B 50/06 (2006.01)

C12N 15/10 (2006.01)

C12Q 1/6804 (2008.01)

12

TRADUCCIÓN DE PATENTE EUROPEA

T3

86 Fecha de presentación y número de la solicitud internacional: **02.02.2018** **PCT/US2018/016587**

87 Fecha y número de publicación internacional: **09.08.2018** **WO18144813**

96 Fecha de presentación y número de la solicitud europea: **02.02.2018** **E 18747697 (3)**

97 Fecha y número de publicación de la concesión europea: **22.11.2023** **EP 3583214**

54 Título: **Métodos y composiciones para identificar o cuantificar dianas en una muestra biológica**

30 Prioridad:

05.06.2017 US 201762515180 P

02.02.2017 US 201762453726 P

23.08.2017 US 201762549189 P

21.12.2017 US 201762609332 P

15.12.2017 US 201762599450 P

15.09.2017 US 201762559228 P

45 Fecha de publicación y mención en BOPI de la traducción de la patente:
08.05.2024

73 Titular/es:

NEW YORK GENOME CENTER, INC. (100.0%)

101 Avenue Of The Americas, 7th Floor

New York, NY 10013, US

72 Inventor/es:

STOECKIUS, MARLON;

SMIBERT, PETER y

HOUCK-LOOMIS, BRIAN

74 Agente/Representante:

DEL VALLE VALIENTE, Sonia

Observaciones:

Véase nota informativa (Remarks, Remarques o Bemerkungen) en el folleto original publicado por la Oficina Europea de Patentes

ES 2 968 290 T3

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

DESCRIPCIÓN

Métodos y composiciones para identificar o cuantificar dianas en una muestra biológica

5 Solicitudes de patente relacionadas

Esta solicitud de patente reivindica el beneficio de la Solicitud de patente provisional de los EE. UU. núm. 62/609332 presentada el 21 de diciembre de 2017, la Solicitud de patente provisional de los EE. UU. núm. 62/599450 presentada el 15 de diciembre de 2017, la Solicitud de patente provisional de los EE. UU. núm. 62/559228 presentada el 15 de septiembre de 2017, la Solicitud de patente provisional de los EE. UU. núm. 62/549189 presentada el 23 de agosto de 2017, la Solicitud de patente provisional de los EE. UU. núm. 62/515180 presentada el 5 de junio de 2017 y la Solicitud de patente provisional de los EE. UU. núm. 62/453726 presentada el 2 de febrero de 2017.

15 Declaración sobre la investigación o desarrollo financiado con fondos federales

Esta invención se realizó con financiación gubernamental mediante la subvención núm. R21-HG-009748 otorgada por los National Institutes of Health. El gobierno tiene algunos derechos sobre esta invención.

20 Incorporación por referencia de material presentado en formato electrónico

Listado de secuencias, material adjunto presentado en formato electrónico. Este archivo se ha marcado como "NYG LIPP35PCT_ST25.txt", con fecha 23 de enero de 2018 y contiene 11 kB.

25 Antecedentes de la invención

La capacidad de caracterizar células individuales en una población heterogénea es cada vez más importante en la investigación biológica y el diagnóstico clínico. La naturaleza no sesgada y de alto rendimiento de los modernos enfoques de secuenciación de ARN de células individuales (scRNA-seq) ha demostrado su valor para describir poblaciones de células heterogéneas¹⁻³. Antes de la genómica de células individuales, los estados celulares se describían rutinariamente mediante el uso de paneles depurados de anticuerpos marcados de forma fluorescente dirigidos a proteínas de la superficie celular, que a menudo son indicadores fiables de la actividad y la función celular⁷. Estudios recientes^{8,9} han demostrado el potencial de acoplar las mediciones de "clasificación de índice" desde un clasificador celular a la transcriptómica de células individuales; este proceso permite cartografiar inmunofenotipos sobre agrupaciones derivadas transcriptómicamente. Sin embargo, enfoques masivamente paralelos basados en microfluidodinámica de gotículas¹⁻³, micropocillos^{47,48} o indexación combinatoria^{20,30} son incompatibles con la citometría y, por lo tanto, no pueden mejorarse con la información de la proteína. Los métodos dirigidos para medir simultáneamente transcritos y proteínas en células individuales están limitados en escala o solo pueden perfilar unos pocos genes y proteínas en paralelo¹⁰⁻¹⁴.

Tradicionalmente, la mayoría de los métodos de clasificación se han basado en la detección óptica de proteínas de la superficie celular. El análisis posterior de las células clasificadas proporciona una capa adicional de información para el fenotipado y la caracterización celular. Con los costes decrecientes de secuenciación de alto rendimiento vistos durante los últimos años, han aparecido una variedad de métodos de laboratorio para aislar y secuenciar el contenido de ARN de células individuales (secuenciación de ARN unicelular, scRNA-seq). Los enfoques iniciales de secuenciación de células individuales emplearon la clasificación de células activadas por fluorescencia (FACS) para aislar y repartir las células de una población entre pocillos individuales de una placa de microtitulación para correlacionar el contenido de sus transcriptomas con la expresión de marcadores específicos de la superficie celular. Si bien son eficaces, los enfoques FACS/scRNA-seq tienen un rendimiento relativamente bajo y un sesgo experimental ya que solamente los tipos de células seleccionados *a priori* se clasifican y secuencian. Por lo tanto, estos métodos no son muy adecuados para el descubrimiento de nuevas poblaciones de células o para caracterizar tejidos complejos que requieren el análisis de decenas de miles de células.

La transición de enfoques basados en placas a métodos de microfluidos/nanopocillos desarrollada por Fluidigm y Wafergen permitió a los investigadores escalar hasta un elevado número de células, aliviar los cuellos de botella de rendimiento, solventar el sesgo experimental que aparecía en FACS y automatizar los procesos de captura de células y preparación de genotecas requeridos para scRNA-seq. La reciente adopción de enfoques de microfluidodinámica de gotículas como Drop-seq¹, InDrop², 10x Genomics³ y un producto de Illumina/Bio-Rad, ha permitido que la sRNA-seq A escalara hasta un número masivo de células. Las actuales plataformas de microfluidodinámica de gotículas producen emulsiones acuosas en aceite de tamaño de nanolitros a velocidades superiores a 1000 gotículas por segundo. Las micropartículas con códigos de barras moleculares únicos coencapsuladas con células en gotículas permiten agrupar transcritos que se originan en la misma célula. Este enfoque tiene un rendimiento significativamente mejorado generando decenas de miles de reacciones individuales de una sola célula por experimento consiguiendo a la vez reducciones de costes significativas asociadas con el uso de reactivos de volumen de nanolitro. Aunque los avances basados en gotículas de la genómica de células individuales han cambiado drásticamente la escala de experimentos de scRNA-seq, estos métodos tienen una desventaja clave: Todos los métodos de secuenciación de

ARN de células individuales basados en gotículas pierden información fenotípica importante no relacionada con los niveles de proteína en general o la expresión de proteínas de superficie celular en particular (Tabla 1).

- 5 Los enfoques actuales para detectar y/o medir simultáneamente transcritos y proteínas en células individuales se basan en la clasificación celular indexada mediante el uso de un número limitado de marcadores junto con la secuenciación de ARN en placas^{8,9} o ensayo de ligadura de proximidad (PLA y sus derivados) junto con la PCR digital¹⁰⁻¹³ o la citometría de masa¹⁴. Estos ensayos tienen una escala limitada y/o solamente pueden perfilar unos pocos genes y proteínas en paralelo (véase la Tabla 1 para la comparación de diferentes tecnologías).

Tabla 1 Métodos para el perfilado de proteínas y ARN de células individuales			
TECNOLOGÍA (referencias en superíndice)	Número		
	TRANSCRITOS	PROTEÍNAS	CÉLULAS
Sec. por ARN de células individuales (basado en gotículas) ^{1,2,3}	>4.000	-	>10.000
CytoF ^{17,21}	-	<100	>10.000
FACS ^{17,21}	-	<50	>10.000
Sec. por ARN clasificado por índice ^{12,13}	>4.000	<50	-96 - cientos
Wafergen ICELL8	>4.000	-4	-1.800
Fluidigm C 1	>4.000	-4	~96- ~800
PLA/PEA y qPCR ^{13,14,15,16}	~ 96	~36	-96 - cientos
PLAYR ¹⁷	~20	-20	>10.000

- 10 Aunque el transcriptoma puede servir como una lectura detallada del estado celular, se ha demostrado que la abundancia de ARNm a menudo es una medida derivada deficiente para establecer niveles de proteínas, especialmente en procesos de desarrollo⁴⁻⁶. La expresión de marcadores de superficie celular se mide tradicionalmente con anticuerpos marcados con fluorescencia mediante citometría de células, y las poblaciones de células complejas se pueden caracterizar por la combinación de marcadores que expresan. Por ejemplo, en los últimos años se han determinado mapas elaborados de tipos de células basados en marcadores de proteínas de los sistemas inmunitario y nervioso⁷. Esto condujo al uso de la citometría de células como herramienta de diagnóstico y monitorización en una serie de campos patológicos, más especialmente en oncología e inmunología. Sin embargo, los enfoques basados en FACS están limitados en lo que respecta al número de marcadores que pueden analizarse simultáneamente y por el hecho de que las células elegidas para el análisis están sesgadas por la selección de marcadores de superficie conocidos.

- 20 Por lo tanto, se necesitan composiciones y métodos más eficientes para el análisis cualitativo y cuantitativo de diversas dianas celulares (y de otro tipo) para aplicaciones de diagnóstico e investigación.

25 Resumen de la invención

- La invención se define mediante las reivindicaciones adjuntas. En un aspecto, una composición comprende una construcción que comprende un ligando unido o conjugado a una construcción polimérica, es decir, una secuencia de oligonucleótidos, mediante un enlazador. El ligando está diseñado para unirse específicamente a una diana de una muestra biológica. La construcción polimérica, por ejemplo, una secuencia de oligonucleótidos, comprende una sonda de amplificación; un código de barras que identifica específicamente el ligando, una etiqueta molecular aleatoria (*Random Molecular Tag*, RMT) opcional, o un identificador molecular único (*Unique Molecular Identifier*, UMI), denominado en lo sucesivo como "UMI" que se coloca adyacente al código de barras en su extremo 5' o 3'; y un anclaje para la hibridación a una secuencia de captura que comprende una secuencia complementaria del anclaje y para la posterior generación de secuencias bicatenarias.

- En otro aspecto, el enlazador entre ligando y construcción polimérica puede ser un enlace covalente escindible.

- 40 En otro aspecto, la composición puede contener además una o más construcciones "adicionales" que difieren en al menos uno de diana, ligando y código de barras, así como por el UMI, de cualquier otra construcción en la composición. En otro aspecto adicional, una composición comprende una o más construcciones "sustancialmente idénticas". En ciertas realizaciones, cada construcción "sustancialmente idéntica" difiere de cualquier otra construcción de referencia (por ejemplo, la "primera" construcción o una construcción "adicional") en la composición solo por la identidad de la secuencia del UMI opcional o la ausencia de un UMI a partir de la construcción de referencia.

En otro aspecto más, un kit que comprende una o más de las composiciones y realizaciones descritas en la presente memoria y reactivos opcionales para la realización de uno o más métodos.

En otro aspecto, un método para detectar una o más dianas en una muestra biológica usa una o más de las composiciones y construcciones descritas en la presente memoria. En un aspecto, la diana es un antígeno o epítipo de la superficie celular y la composición contiene una única construcción dirigida a esa diana, es decir, una “primera” construcción. En otra realización, la composición contiene múltiples construcciones “sustancialmente idénticas”, es decir, sustancialmente idénticas a la construcción “primera”, o una o más construcciones “adicionales” dirigidas a diferentes dianas y en consecuencia con diferentes componentes, como se ha descrito anteriormente y se define a continuación. El método implica poner en contacto una muestra biológica con una o más de las composiciones descritas anteriormente. Los pasos adicionales implican lavar para eliminar las construcciones no unidas y/o hibridar cada secuencia de anclaje en construcciones individuales a una secuencia de captura. Otro paso implica extender la captura hibridada a la secuencia de anclaje para copiar el código de barras de la construcción, el UMI y la sonda de amplificación en secuencias bicatenarias. A continuación, las secuencias del código de barras de la construcción polimérica se amplifican o detectan para identificar si la muestra biológica expresa o contiene una única diana, una o más dianas adicionales o una combinación de múltiples dianas. Alternativamente, el nivel de expresión de las dianas en la muestra se determina detectando la cantidad de los correspondientes códigos de barras de la construcción polimérica normalizados por una cantidad de cualquier UMI o la cantidad media de dos o más UMI en la muestra tratada.

En otro aspecto, un método como se ha descrito anteriormente incluye aislar células individuales, fragmentos celulares o poblaciones de células, de la muestra biológica unida a una o más de las construcciones dirigidas para detectar una o más dianas después del paso de lavado. Otro paso más implica amplificar las secuencias bicatenarias con cebadores hibridados a las sondas de amplificación.

En otro aspecto adicional, un método usa las composiciones descritas en la presente memoria para caracterizar una célula mediante la detección simultánea de uno o más epítipos ubicados en o sobre la célula y/o su transcriptoma. Un método de este tipo comprende poner en contacto una muestra biológica que contiene células con una o más de las composiciones descritas en la presente memoria. En una realización de este método, los ligandos son anticuerpos o fragmentos de los mismos que se unen específicamente a epítipos diana ubicados en una célula o en la superficie de una célula. Dicho método puede usar los pasos de la técnica Drop-seq¹ (secuenciación en gotículas), por ejemplo, encapsular una única célula individual unida a una o más construcciones en una gotícula acuosa que contiene una perla microfluidica. Cada perla está conjugada con la secuencia de un oligonucleótido de captura. Después de la lisis celular, los ARNm de la célula y de la secuencia de oligonucleótidos de la construcción se hibridan con las secuencias de poli T del oligonucleótido de captura en la perla. A partir de las secuencias hibridadas con la perla, se generan ADNc bicatenarios que contienen la secuencia de código de barras de la perla y los transcritos inversos del ARNm celular y del ADN bicatenario que contiene la secuencia de código de barras de la perla y la secuencia de oligonucleótidos de la construcción. Se genera una genoteca de amplificación que contiene el ADNc de los transcritos celulares y el ADN que contiene la secuencia de oligonucleótidos de la construcción. En este método, el transcriptoma de la genoteca está asociado con la célula identificada por el anticuerpo en una construcción identificada específicamente simultáneamente. Usando las composiciones descritas en la presente memoria, las secuencias de código de barras de la construcción polimérica se usan para identificar si la célula individual expresa el epítipo diana. El transcriptoma de la genoteca está simultáneamente asociado con la célula identificada que expresa la diana.

En otro aspecto más, las construcciones descritas anteriormente se usan en un método de código de barras por lotes o “*hashtaging*” celular. Una construcción descrita anteriormente, por ejemplo, un anticuerpo o cualquier ligando que se une a una célula, conjugado o asociado con una secuencia de oligonucleótidos que comprende una sonda de amplificación; un código de barras que identifica específicamente el ligando, una etiqueta molecular aleatoria (*Random Molecular Tag*, RMT) opcional, o un identificador molecular único (*Unique Molecular Identifier*, UMI), denominado en lo sucesivo como “UMI” que se coloca adyacente al código de barras en su extremo 5' o 3'; y un anclaje, por ejemplo, la secuencia de poli A) como se describen en la presente memoria se usan para etiquetar cada célula dentro de una muestra antes de la agrupación. A continuación varias muestras, marcadas con tales construcciones, se agrupan y luego se analizan por métodos tales como scRNA-seq o CITE-seq como se describe en la presente memoria. Las construcciones utilizadas para marcar cada célula dentro de una muestra tienen una secuencia de sonda de amplificación diferente de la que se usa en los métodos scRNA-seq o CITE-seq. Dicho multiplexado del marcado permite una determinación inequívoca de la mayoría de los dobletes y la capacidad de controlar los efectos de lote.

En algunos aspectos presentados en la presente memoria, un método es para detectar una muestra o diana en un ensayo multiplexado, comprendiendo el método: a) poner en contacto una primera muestra con una primera construcción que comprende un primer ligando unido a un primer oligonucleótido, en donde el primer ligando se une específicamente a una primera diana, y el primer oligonucleótido comprende: i) una primera sonda de amplificación, ii) un primer código de barras que identifica específicamente la primera muestra y iii) un primer anclaje. En algunos aspectos, el método comprende además: b) poner en contacto una segunda muestra con una segunda construcción que comprende un segundo ligando unido a un segundo oligonucleótido, en donde el segundo ligando se une específicamente a una segunda diana, y el segundo oligonucleótido comprende: i) una segunda sonda de amplificación, ii) un segundo código de barras que identifica específicamente la segunda muestra y iii) un segundo

anclaje. En algunas realizaciones, la primera diana y la segunda diana son la misma diana y, opcionalmente, la primera sonda de amplificación y la segunda sonda de amplificación son sustancialmente idénticas y, opcionalmente, el primer anclaje y el segundo anclaje son sustancialmente idénticos. En algunos aspectos, el método comprende, además: c) poner en contacto la primera y la segunda muestras con una tercera construcción que comprende un tercer ligando unido a un tercer oligonucleótido, en donde el tercer ligando se une específicamente a una tercera diana, y el tercer oligonucleótido comprende: (i) una tercera sonda de amplificación, (ii) un tercer código de barras que identifica específicamente el tercer ligando y (iii) un tercer anclaje. En algunos aspectos, el método comprende además d) poner en contacto la primera y la segunda muestras con una cuarta construcción que comprende un cuarto ligando unido a un cuarto oligonucleótido, en donde el cuarto ligando se une específicamente a una cuarta diana, y el cuarto oligonucleótido comprende: i) una cuarta sonda de amplificación, ii) un cuarto código de barras que identifica específicamente el cuarto ligando, y iii) un cuarto anclaje.

En algunas realizaciones, la tercera sonda de amplificación y la cuarta sonda de amplificación son sustancialmente idénticas y son diferentes de la primera sonda de amplificación y la segunda sonda de amplificación. En algunas realizaciones, el primer anclaje, el segundo anclaje, el tercer anclaje y el cuarto anclaje son sustancialmente idénticos y opcionalmente comprenden una secuencia de poli A de al menos 10 nucleótidos de longitud. En algunas realizaciones, la tercera diana y la cuarta diana son dianas diferentes y, opcionalmente, la tercera diana es diferente de las dianas primera o segunda y, opcionalmente, la cuarta diana es diferente de la primera o segunda dianas.

En algunos aspectos, el método comprende además e) poner en contacto una tercera muestra con una quinta construcción que comprende un quinto ligando que se une específicamente a una quinta diana, en donde la quinta diana es opcionalmente la misma que la primera diana, y el quinto ligando está unido a un quinto oligonucleótido que comprende: i) una quinta sonda de amplificación, opcionalmente sustancialmente igual a la primera sonda de amplificación, ii) un quinto código de barras que identifica específicamente la tercera muestra, y iii) un quinto anclaje, opcionalmente sustancialmente igual al primer anclaje, y que comprende opcionalmente una secuencia de poli A.

En aspectos, el método comprende además f) poner en contacto la primera y la segunda muestras y, opcionalmente muestras adicionales, con una sexta construcción que comprende un sexto ligando, en donde el sexto ligando se une específicamente a una sexta diana, y está unido a un sexto oligonucleótido que comprende: i) una sexta sonda de amplificación, opcionalmente sustancialmente igual a la tercera sonda de amplificación, ii) un sexto código de barras que identifica específicamente la sexta diana, y iii) un sexto anclaje, opcionalmente el mismo que el tercer anclaje, y que comprende opcionalmente una secuencia de poli A.

En algunas realizaciones, la primera y la segunda muestra y, opcionalmente, una o más muestras adicionales, comprenden una o más células, y las dianas primera, segunda, tercera, cuarta, quinta y sexta están presentes en, o sobre la superficie de, al menos una de las una o más células. En algunas realizaciones, la puesta en contacto de (a), (b), (c), (d), (e) o (f) comprende poner en contacto la una o más células de la primera muestra, la segunda muestra y muestras adicionales opcionales con la primera, segunda, tercera, cuarta, quinta o sexta construcciones. En algunas realizaciones, la primera y la segunda muestra, y opcionalmente una o más muestras adicionales, comprenden uno o más orgánulos celulares, mitocondrias, exosomas, liposomas, vesículas sintéticas o naturales, microvesículas, ectosomas, núcleos, bacterias, virus, perlas, partículas, micropartículas, nanopartículas, macromoléculas, y esferas de lípidos, fosfolípidos o membranas sintéticas o naturales, y las dianas primera, segunda, tercera, cuarta, quinta y sexta están presentes en, o sobre la superficie de, al menos uno de los uno o más orgánulos celulares, mitocondrias, exosomas, liposomas, vesículas sintéticas o naturales, microvesículas, ectosomas, núcleos, bacterias, virus, perlas, partículas, micropartículas, nanopartículas, macromoléculas y las esferas de lípidos, fosfolípidos o membranas naturales o sintéticas. En algunas realizaciones, la puesta en contacto de (a), (b), (c), (d), (e) o (f) comprende poner en contacto el uno o más orgánulos celulares, mitocondrias, exosomas, liposomas, vesículas sintéticas o naturales, microvesículas, ectosomas, núcleos, bacterias, virus, perlas, partículas, micropartículas, nanopartículas, macromoléculas y esferas de lípidos, fosfolípidos o membranas sintéticas o naturales de la primera muestra, la segunda muestra y muestras adicionales opcionales con la primera, segunda, tercera, cuarta, quinta o sexta construcciones. En algunas realizaciones, la puesta en contacto de (a) y (b), y opcionalmente (e) tiene lugar antes de la puesta en contacto de una cualquiera de (c), (d) o (f). En algunas realizaciones, la puesta en contacto de (c), (d) o (f) comprende poner en contacto una mezcla de la primera muestra, la segunda muestra y opcionalmente muestras adicionales con la tercera, cuarta o sexta construcciones. En algunas realizaciones, el primer, segundo, tercer, cuarto, quinto o sexto ligandos comprenden un anticuerpo, o fragmento de unión a antígeno del mismo. En algunas realizaciones (i) el primer, segundo, tercer, cuarto, quinto o sexto anclaje se ubica en 3' de la primera, segunda, tercera, cuarta, quinta o sexta sonda de amplificación, respectivamente, y en 3' del primer, segundo, tercer, cuarto, quinto o sexto código de barras, respectivamente; y opcionalmente, (ii) la primera, segunda, tercera, cuarta, quinta o sexta sonda de amplificación está ubicada en 5' del primer, segundo, tercer, cuarto, quinto o sexto código de barras, respectivamente, y en 5' del primer, segundo, tercer, cuarto, quinto o sexto anclaje, respectivamente. En algunas realizaciones, el método comprende además lavar la primera muestra, la segunda muestra, o una mezcla de la primera muestra y la segunda muestra, y opcionalmente muestras adicionales después de una cualquiera o más de los pasos (a), (b), (c), (d), (e) o (f) para eliminar las construcciones no unidas. En algunas realizaciones, después de (a), (b), (c), (d), (e) o (f), encapsular una primera célula individual de una de la primera, segunda o tercera muestras en una primera gotícula que comprende una primera perla conjugada con una pluralidad de un primer oligonucleótido de captura que comprende, de 5' a 3', una séptima sonda de amplificación, un séptimo código de barras que identifica la primera perla,

y una secuencia complementaria a la primera, segunda, tercera, cuarta, quinta o sexta secuencia de anclaje, y opcionalmente encapsular una segunda célula individual de una de la primera, segunda o tercera muestras en una segunda gotícula que comprende una segunda perla conjugada con una pluralidad de un segundo oligonucleótido de captura que comprende, de 5' a 3', la séptima sonda de amplificación, un octavo código de barras que identifica la segunda perla y una secuencia complementaria a la primera, segunda, tercera, cuarta, quinta o sexta secuencia de anclaje. En algunas realizaciones, el método comprende además lisar la primera y segunda células individuales proporcionando de este modo un primer lisado encapsulado en la primera gotícula y un segundo lisado encapsulado en la segunda gotícula, en donde el primer y el segundo lisados comprenden opcionalmente ARNm. En algunas realizaciones, el método comprende además poner en contacto el lisado de la primera y segunda células con una polimerasa. En algunas realizaciones, el método comprende además generar ADNc y secuencias oligonucleotídicas bicatenarias del primer, segundo, tercer, cuarto, quinto o sexto oligonucleótidos.

En algunos aspectos, en la presente memoria se presenta un método para detectar una o más dianas en una muestra biológica, comprendiendo el método poner en contacto la muestra biológica con uno o más de: a) una composición que comprende una primera construcción que comprende un primer ligando unido o conjugado a una construcción polimérica mediante un enlazador, uniéndose específicamente dicho primer ligando a una primera diana y comprendiendo dicha construcción polimérica: una sonda de amplificación; un código de barras que identifica específicamente dicho primer ligando; un identificador molecular único opcional que se coloca adyacente al código de barras en su extremo 5' o 3'; y un anclaje para hibridarse a una secuencia de captura que comprende una secuencia complementaria de dicho anclaje; b) una composición que comprende al menos una construcción adicional, construcción que comprende un ligando adicional unido o conjugado a una construcción polimérica adicional mediante un enlazador, uniéndose específicamente dicho ligando adicional a una diana adicional y comprendiendo dicha construcción polimérica adicional una sonda de amplificación; un código de barras adicional que identifica específicamente dicho ligando adicional; un identificador molecular único adicional opcional que se coloca adyacente al código de barras adicional en su extremo 5' o 3' y un anclaje para hibridarse a una secuencia de captura que comprende una secuencia complementaria de dicho anclaje; y c) una composición que comprende una o más construcciones sustancialmente idénticas, difiriendo cada una de las construcciones sustancialmente idénticas de cualquier otra construcción de referencia primera o adicional por la secuencia de su identificador molecular único (UMI) opcional o ausencia del UMI.

En algunos aspectos, en la presente memoria se presenta un método de alto rendimiento para detectar uno o más epítomos en una muestra biológica, comprendiendo el método poner en contacto una muestra biológica con una o más de (i) una composición que comprende una primera construcción que comprende un primer anticuerpo o fragmento del mismo que se une específicamente a un primer epítipo, dicho primer anticuerpo o fragmento unido o conjugado a una primera construcción polimérica mediante un enlazador, en donde la primera construcción polimérica comprende: una sonda de amplificación; una secuencia de código de barras que identifica específicamente dicho primer anticuerpo o fragmento de cualquier otro anticuerpo o fragmento que reconoce un epítipo diferente, una secuencia de identificador molecular único opcional que se coloca adyacente al extremo 5' o 3' del código de barras y una secuencia de anclaje para hibridarse a una secuencia de captura que comprende una secuencia complementaria de dicho anclaje; (ii) una composición de (i) que comprende al menos una construcción adicional, que comprende un anticuerpo adicional o fragmento del mismo unido o conjugado a una construcción polimérica adicional mediante un enlazador, uniéndose específicamente dicho anticuerpo adicional o fragmento del mismo a un epítipo adicional, y comprendiendo dicha construcción polimérica adicional: una sonda de amplificación; un código de barras adicional que identifica específicamente dicho anticuerpo adicional o fragmento del mismo; un identificador molecular único adicional opcional que se coloca adyacente al extremo 5' o 3' del código de barras adicional y una secuencia de anclaje de (i), en donde dicha construcción adicional difiere de cualquier otra construcción por la composición de su anticuerpo, epítipo, código de barras y UMI; y (iii) una composición de (i) o (ii) que comprende una o más construcciones sustancialmente idénticas, difiriendo cada una de las construcciones sustancialmente idénticas de cualquier otra construcción de referencia primera o adicional por la secuencia de su identificador molecular único (UMI) opcional o ausencia del UMI.

En algunos aspectos presentados en la presente memoria, un método es para detectar al menos dos dianas en al menos una primera y una segunda muestra, comprendiendo el método: a) poner en contacto la primera muestra con una primera construcción que comprende un primer ligando unido a un primer oligonucleótido, en donde el primer ligando se une específicamente a una primera diana, y el primer oligonucleótido comprende: i) una primera sonda de amplificación, ii) un primer código de barras que identifica específicamente la primera muestra y iii) un anclaje que comprende una secuencia de poli A; b) poner en contacto la segunda muestra con una segunda construcción que comprende el primer ligando unido a un segundo oligonucleótido, en donde el segundo oligonucleótido comprende: i) la primera sonda de amplificación, ii) un segundo código de barras que identifica específicamente la segunda muestra y iii) el anclaje; c) poner en contacto la primera y la segunda muestras con una tercera construcción que comprende un segundo ligando unido a un tercer oligonucleótido, en donde el segundo ligando se une específicamente a una segunda diana, y el tercer oligonucleótido comprende: (i) una segunda sonda de amplificación, (ii) un tercer código de barras que identifica específicamente el segundo ligando y (iii) el anclaje; y d) poner en contacto la primera y la segunda muestras con una cuarta construcción que comprende un tercer ligando unido a un cuarto oligonucleótido, en donde el tercer ligando se une específicamente a una tercera diana, y el cuarto oligonucleótido comprende: i) la segunda sonda de amplificación, ii) un cuarto código de barras que identifica específicamente el tercer ligando y iii) el anclaje.

En algunos aspectos, se presenta en la presente memoria un kit que comprende: a) una primera construcción que comprende un primer ligando unido a un primer oligonucleótido, en donde el primer ligando se une específicamente a una primera diana, y el primer oligonucleótido comprende: i) una primera sonda de amplificación, ii) un primer código de barras único configurado para identificar específicamente una primera muestra y iii) un anclaje que comprende una secuencia de poli A; b) una segunda construcción que comprende el primer ligando unido a un segundo oligonucleótido, en donde el segundo oligonucleótido comprende: i) la primera sonda de amplificación, ii) un segundo código de barras único configurado para identificar específicamente una segunda muestra y iii) el anclaje; c) una tercera construcción que comprende un segundo ligando unido a un tercer oligonucleótido, en donde el segundo ligando se une específicamente a una segunda diana, y el tercer oligonucleótido comprende: (i) una segunda sonda de amplificación, (ii) un tercer código de barras único configurado para identificar específicamente el segundo ligando y (iii) el anclaje; y d) una cuarta construcción que comprende un tercer ligando unido a un cuarto oligonucleótido, en donde el tercer ligando se une específicamente a una tercera diana, y el cuarto oligonucleótido comprende: i) la segunda sonda de amplificación, ii) un cuarto código de barras único configurado para identificar específicamente el tercer ligando y iii) el anclaje.

En algunos aspectos, se presenta en la presente memoria una composición que comprende una construcción que comprende un ligando unido a un oligonucleótido, en donde el ligando se une específicamente a una diana, y el oligonucleótido comprende: i) una sonda de amplificación, ii) un código de barras único configurado para identificar específicamente una primera muestra y iii) un anclaje que comprende opcionalmente una secuencia de poli A.

En otro aspecto más, los métodos descritos en la presente memoria son métodos de alto rendimiento y emplean otras técnicas conocidas de detección y secuenciación.

Otros aspectos y ventajas de estas composiciones y métodos se describen adicionalmente en la siguiente descripción detallada de las realizaciones preferidas de la misma.

Breve descripción de los dibujos

Las Figuras 1A a 1C son representaciones esquemáticas que muestran el proceso CITE-seq y los componentes que permiten la detección simultánea de transcriptomas de células individuales y marcadores de proteínas.

La Figura 1A es una ilustración que muestra una realización de una construcción descrita en la presente memoria en la que los anticuerpos (ligandos) están unidos a una construcción polimérica, que en esta realización es una secuencia de oligonucleótidos con un puente disulfuro (enlazador) y que contiene componentes de secuencia funcional (sonda de amplificación y sonda de PCR) y un código de barras identificador de anticuerpo único seguido de una cola de poli A (Anclaje).

La Figura 1B es una ilustración que muestra que las perlas de Drop-seq son micropartículas que contienen una o más secuencias de oligonucleótidos de la construcción polimérica con las características funcionales de una sonda de amplificación (sonda de PCR), un código de barras de célula individual seguido de un identificador molecular único (UMI) y una cola poli T (Anclaje).

La Figura 1C es un diagrama esquemático de una realización del protocolo CITE-seq. Las células se incuban con anticuerpos (1), se lavan (2) y se pasan a través de un chip de microfluidos donde una célula individual y una perla se encapsulan en una gotícula (3) y (4). Después de la lisis celular (5), los ARNm y las construcciones de anticuerpo-oligonucleótido se unen a las perlas de Drop-seq (6). La retrotranscripción y el interruptor de molde se realizan a granel después de la rotura de la gotícula de emulsión (7). Después de la PCR SMART, el ADNc de longitud completa (8a) y los productos de construcción de oligo-anticuerpo (8b) pueden separarse por tamaños y amplificarse independientemente.

La Figura 1D es una ilustración que muestra una realización de una construcción descrita en la presente memoria en la que los anticuerpos (ligandos) están unidos a una construcción polimérica que contiene componentes de secuencia funcional (sonda de amplificación y sonda de PCR) y un código de barras identificador de anticuerpo único seguido de una cola de poli A (Anclaje).

Las Figuras 2A a 2E son representaciones que muestran que CITE-seq identifica con precisión diferentes especies en el experimento de mezcla.

La Figura 2A es un resultado de electroforesis en gel así como ilustraciones de las moléculas detectadas. Los complejos de anticuerpos-oligo (1) aparecen como un frotis de alto peso molecular en gel de agarosa y pueden escindirse reduciendo el enlace disulfuro (2).

La Figura 2B son representaciones que ilustran dos anticuerpo-oligos. Los anticuerpos dirigidos contra integrina Beta-1 (CD29) de ratón se unen al código de barras del oligo 1 que contiene un enlazador de tipo puente disulfuro, una sonda de amplificación (también denominada secuencia común o sonda de PCR), un código de barras identificador de anticuerpo único (5'-ATGTCCT-3') y un UMI que contiene 4 nt seguido de una cola de poli A (panel superior). Los

anticuerpos dirigidos contra CD29 humano están unidos al código de barras del oligo 2 que contiene un puente disulfuro, una secuencia común (sonda de amplificación, sonda de PCR), un código de barras identificador de anticuerpo único (5'-GCCATTA-3') y un UMI que contiene 4 nt seguido de una cola de poli A (panel inferior).

La Figura 2C son los resultados de la electroforesis en gel y la traza de electroforesis capilar de los ADNc y oligos de longitud completa derivados de los anticuerpo-oligos. Después de la retrotranscripción y la PCR SMART, se pueden observar dos poblaciones distintas de productos (panel derecho). Estos pueden separarse por tamaños en los ADNc de longitud completa (panel superior, traza de electroforesis capilar) y producto anticuerpo-oligo (panel inferior) y amplificarse independientemente.

La Figura 2D es una gráfica de puntos que muestra la lectura de ARN-seq así como secuencias de oligo específicas de anticuerpo de ratón y humano obtenidas en el mismo ciclo de secuenciación. Las células humanas y de ratón se incubaron con anticuerpos oligoetiquetados específicos de marcadores de superficie de células humanas o de ratón (integrina beta, CD29). A continuación, las células se pasaron a través del flujo de trabajo Drop-seq a una concentración más alta para permitir la encapsulación de múltiples células. A continuación, se determinaron especies en cada gotícula (puntos en diagrama de dispersión) mediante secuenciación de ARNm (ARN humano: rodeado por una línea continua, excepto un pequeño número de salidas; ARN de ratón: rodeado por una línea discontinua, excepto por un pequeño número de valores atípicos; ARN de especies mezcladas: resto de los puntos excepto algunos valores atípicos mencionados anteriormente).

La Figura 2E es un diagrama de puntos que muestra la clasificación primaria de células contadas mediante secuenciación de ARNm y ADNc generadas a partir de las mismas. Los puntos que representan células humanas y células de ratón están marcados con un círculo sólido y un círculo discontinuo respectivamente.

La Figura 3A es una representación que muestra los resultados de un análisis CITE-seq de 8700 células mononucleares de la sangre marcadas con 10 construcciones de anticuerpo CITE-seq como se describe en la presente memoria que tiene los componentes incluidos en la Tabla 2 (véase el Ejemplo 7 más adelante), tSNE (incrustación estocástica de vecinos distribuida en t)³⁴ y la agrupación se realiza usando análisis de correlación canónico, que integra mediciones de proteínas y ARN. Estos datos muestran que CITE-seq permite una mayor agrupación de células y clasificación de células mononucleares de sangre de cordón umbilical.

La Figura 3B es un análisis CITE-seq del mismo conjunto de datos de la Figura 3A mediante el uso de datos de ARN en solitario. Los símbolos de la figura son Mono (para monocitos), B para células B, T para células T, NK para células citolíticas naturales, DC para células dendríticas convencionales, pDC para plasmacitoides DC, Pre para precursores, y Ery para eritroblastos. Comparando la Figura 3A y la Figura 3B se muestra una resolución mejorada cuando se usan datos multimodales.

La Figura 3C muestra diagramas biaxiales de datos de anticuerpos CITE-seq para anticuerpos seleccionados, es decir, las construcciones núms. 1, 3, 4, 6, 7 y 9 de la Tabla 2 (véase el Ejemplo 7). Estos datos muestran que, a diferencia de la información obtenida por citometría de flujo, mediante el uso de la metodología CITE-seq y las composiciones pone a disposición el transcriptoma para cada célula individual (cada punto) dentro del gráfico. Por lo tanto, las células se pueden analizar y clasificar adicionalmente en función de sus datos de ARN, datos de proteínas o ambos.

La Figura 4 es una serie de representaciones biaxiales generadas mediante la multiplexación de 8700 células mononucleares de sangre marcadas con las 10 construcciones de anticuerpos de la Tabla 2 (Ejemplo 7) en un análisis CITE-seq como se describe en la presente memoria. Se muestran representaciones biaxiales de datos de anticuerpos CITE-seq para todos los 10 anticuerpos. Estos datos son comparables a la información obtenida por citometría de flujo, mediante el uso de la metodología CITE-seq y las composiciones también pone a disposición el transcriptoma para cada célula individual (cada punto) dentro del gráfico. Por lo tanto, las células se pueden analizar y clasificar adicionalmente en función de sus datos de ARN, datos de proteínas o ambos.

La Figura 5A es una gráfica que muestra el agrupamiento de ARN de aproximadamente 4000 células mononucleares de sangre periférica (PBMC) que contienen células B, células citolíticas NK, células de ratón, células T citolíticas naturales, monocitos, monocitos CD16, células CD4 y células CD8. La tSNE (Incrustación estocástica de vecinos distribuida en t)³⁴ y la agrupación se realizan usando análisis de correlación canónico usando datos de expresión de ARN.

La Figura 5B muestra 6 perfiles de histograma de niveles de ADT (etiqueta derivada de anticuerpos) transformados mediante CLR (cociente logarítmico centrado) en agrupaciones de células B, células citolíticas NK, células de ratón, células T citolíticas naturales, monocitos, monocitos CD16, células CD4 y células CD8 expuestas a las composiciones como se describe en la Tabla 3. Una de estas composiciones comprende un ligando que es un anticuerpo anti-CD3, anticuerpo anti-CD4 o anticuerpo anti-CD8, covalentemente (directamente) unido a una construcción polimérica, en este ejemplo un oligonucleótido de ADN, que contiene una sonda de amplificación compatible con el ARN pequeño de Illumina Truseq, un código de barras de 10 nucleótidos que es único para cada anticuerpo destinado a identificar el ligando y un anclaje de cola de poli A de 30 nucleótidos para hibridarse a una secuencia de captura que comprende una secuencia complementaria del anclaje. Otras composiciones de este tipo comprenden un ligando que es un

anticuerpo anti-CD3, un anticuerpo anti-CD4 o un anticuerpo anti-CD8, unido a través de un enlace estreptavidina-biotina (SAV) como se utiliza en los experimentos de prueba de concepto (Figuras 1-4) a un polímero, en este ejemplo un oligonucleótido de ADN, que contiene una sonda de amplificación compatible con el ARN pequeño Illumina Truseq, un código de barras de 10 nucleótidos que es único para cada anticuerpo destinado a identificar el ligando y un anclaje de cola poli A de 30 nucleótidos para hibridarse a una secuencia de captura que comprende una secuencia complementaria del anclaje. Los perfiles de histograma en las diferentes poblaciones (por ejemplo, células NK, CD4, CD8) son comparables entre la SAV y la conjugación directa.

Las Figuras 6A-6F muestran la multiplexación de muestras usando anticuerpos que tienen un código de barras de ADN.

La Figura 6A es una visión general esquemática de la multiplexación de muestras mediante *hashing* celular. Las células de diferentes muestras se incuban con anticuerpos que tienen un código de barras de ADN que reconocen proteínas de superficie celular ubicuas. Diferentes códigos de barras (denominados oligos-*“hashtag”*, HTO) en los anticuerpos permiten la combinación de múltiples muestras en un experimento de secuenciación de ARN de célula individual. Después de la secuenciación, las células se pueden clasificar según su muestra de origen basándose en los niveles de HTO.

La Figura 6B es un diagrama de dispersión representativo que muestra recuentos sin procesar para HTO A y HTO B, para todos los códigos de barras de las células. Ambos ejes se fijan al cuantil 99,9 % de cuantiles para excluir los valores atípicos visuales.

La Figura 6C es un mapa de calor de todos los niveles de HTO normalizados y escalados, basándose en las clasificaciones de los inventores. Los dobletes y los multipletes expresan más de un HTO. Las poblaciones negativas contienen células HEK-293T y células NIH-3T3 de ratón que se agregaron a los experimentos como controles negativos. Las células con múltiples señales de *“hashtag”* son probablemente dobletes, y la frecuencia de estas células coincide con las tasas de multiplete esperadas para el ensayo descrito en el Ejemplo 10.

La Figura 6D muestra la integración de tSNE del conjunto de datos de HTO. Las células están coloreadas y marcadas basándose en las clasificaciones de los inventores. Ocho grupos singlete y los 28 grupos de dobletes de muestra cruzada están claramente presentes.

La Figura 6E muestra una distribución de los UMI de ARN según el código de barras de la célula para células caracterizadas como singletes (rojo), dobletes (violeta) o negativas (gris).

La Figura 6F muestra la agrupación basada en transcriptoma de perfiles de expresión de células individuales que revela distintas poblaciones de células inmunitarias intercaladas entre donantes. B, células B; T, células T; NK, células citolíticas naturales; mono, monocitos; DC, células dendríticas; pDC, células dendríticas plasmacitoides; y células plasmáticas. Las células se colorean basándose en su clasificación de HTO (ID de donante), como en la Figura 6D.

Las Figuras 7A-7E muestran la validación de *“hashing”* celular usando demuxlet.

La Figura 7A muestra una “matriz de confusión” normalizada por filas que compara las clasificaciones según demuxlet y HTO. Cada valor de la diagonal representa la fracción de códigos de barras para una clasificación de HTO dada que recibió una clasificación idéntica según demuxlet.

La Figura 7B es una distribución de recuento de los HTO más expresados para grupos de singletes concordantes y discordantes. Ambos grupos tienen una intensidad de clasificación idéntica basada en *“hashing”* celular.

La Figura 7C muestra que los singletes discordantes tienen recuentos de UMI inferiores, lo que sugiere que la falta de profundidad de secuenciación contribuyó a llamadas “ambiguas” desde demuxlet.

La Figura 7D son distribuciones de UMI de ARN para multipletes discordantes y concordantes. Solamente los multipletes concordantes presentan mayor complejidad molecular, lo que sugiere que ambos tienen de forma conservadora un exceso de llamadas multipletes en los casos discordantes.

La Figura 7E muestra que demuxlet asigna probabilidades posteriores de multipletes inferiores a las llamadas discordantes.

Las Figuras 8A-8F muestran que el *“hashing”* celular permite una optimización e identificación experimental eficaces de células de baja calidad. Las Figuras 8A a 8C son representaciones que muestran los resultados del rendimiento de una serie de titulación para evaluar concentraciones óptimas de tinción en un panel de anticuerpos de inmunofenotipado CITE-seq. Los recuentos de ADT normalizados para CD8 (Figura 8A) CD45RA (Figura 8B) y CD4 (Figura 8C) se representan para las diferentes concentraciones usadas por prueba.

La Figura 8D muestra una curva de titulación que representa el índice de tinción (SI) para estos tres anticuerpos a través de la serie de titulación. La relación señal/ruido para estos anticuerpos comienza a saturarse a niveles similares a las concentraciones de tinción recomendadas por el fabricante típicas en los anticuerpos de citometría de flujo.

- 5 La Figura 8E muestra que las células con recuentos bajos de UMI pueden distinguirse del ARN ambiental mediante el uso de clasificaciones de HTO. Grupo de singletes clasificados en poblaciones hematopoyéticas canónicas.

La Figura 8F muestra que los códigos de barras clasificados como “negativos” no se agrupan y probablemente representan gotículas “vacías” que solo contienen ARN ambiental.

10 Descripción detallada

Las composiciones descritas en la presente memoria aumentan la sensibilidad de una variedad de metodologías de ensayo. El uso de las composiciones y métodos para detectar múltiples dianas en un entorno complejo es fácilmente escalable y solamente está limitada por el número de ligandos específicos, por ejemplo, anticuerpos, que están disponibles, a diferencia de los métodos de ensayo fluorescente que están limitados por la superposición espectral de fluoróforos disponibles. Por ejemplo, la citometría de flujo permite la medición rutinaria de hasta 15 parámetros por célula^{17,18}. Las composiciones descritas en la presente memoria que emplean código de barras molecular de ligandos (por ejemplo, anticuerpos) permiten multiplexar a prácticamente cualquier cantidad e incluso podrían competir con la paralelización basada en citometría de masa (CyTOF hasta 100 etiquetas)¹⁸.

Por ejemplo, un aspecto de las composiciones y métodos descritos en detalle a continuación permite la medición simultánea de grandes cantidades de marcadores basados en anticuerpos establecidos junto con datos de transcriptoma de células individuales no sesgados, a una escala de decenas de miles de células por experimento. Los inventores denominan esta técnica como Indexación celular de transcriptoma y epítomos mediante secuenciación (CITE-seq), usando las composiciones descritas en la presente memoria. Sin embargo, otras técnicas pueden usar las composiciones descritas para mejorar el estudio y la comprensión de los tipos de células y las poblaciones celulares, tales como la catalogación de los tipos de células en personas sanas o estudiando la regulación génica postranscripcional en el desarrollo y la enfermedad. La eficiencia de cualquier número de técnicas de diagnóstico y aplicaciones para analizar diferentes patologías se puede mejorar con el uso de las composiciones descritas en la presente memoria. Los métodos y composiciones descritos en la presente memoria expanden en gran medida la potencia del fenotipado de células individuales combinando información tanto de proteínas como de transcritos de las mismas células individuales a una escala sin precedentes.

35 I. Componentes de las composiciones y métodos

En las descripciones de las composiciones y métodos descritos en la presente memoria, los diversos componentes pueden definirse mediante el uso de términos técnicos y científicos que tienen el mismo significado que entiende comúnmente el experto en la técnica a la que pertenece la presente invención y por referencia a textos publicados. Dichos textos proporcionan al experto en la técnica una guía general de muchos de los términos utilizados en la presente solicitud. Las definiciones incluidas en esta memoria descriptiva se proporcionan para mayor claridad en la descripción de los componentes y las composiciones de la presente memoria.

Como se utiliza en la presente memoria, el término “construcción” se refiere a un ensamblaje sintetizado químicamente o modificado genéticamente que comprende un ligando unido (covalentemente, no covalentemente, o de otro modo como se indica en la presente memoria) a al menos una construcción polimérica (por ejemplo, en una realización, una secuencia de oligonucleótidos) mediante un enlazador. Cada construcción polimérica comprende varios elementos funcionales: una sonda de amplificación; un código de barras que identifica específicamente el ligando adjunto, un identificador molecular único opcional que se coloca adyacente al código de barras en su extremo 5' o 3' y un anclaje para hibridarse a una secuencia de captura que comprende una secuencia complementaria del anclaje. Estos componentes de la construcción pueden producirse en cualquier orden. En una realización, los componentes se enumeran de 5' a 3': Ligando, Enlazador, Sonda de amplificación, Código de barras y Anclaje con el UMI a cada extremo del código de barras. En otra realización, los componentes se enumeran de 3' a 5': Ligando, Enlazador, Sonda de amplificación, Código de barras y Anclaje con el UMI a cada extremo del código de barras. En otras realizaciones más, estos elementos de la construcción pueden estar en cualquier otro orden. En otra realización más, una construcción comprende un único ligando unido a múltiples construcciones poliméricas idénticas. En una realización, cada construcción polimérica está directamente unida al ligando (un enlace por construcción polimérica). En otra realización, las construcciones poliméricas están unidas al ligando como concatámeros (múltiples construcciones poliméricas por enlace de ligando individual). Por ejemplo, un solo ligando (es decir, un anticuerpo monoclonal) se puede unir a de 1 a 50 construcciones poliméricas.

Una sola cadena de un nucleico a menudo comprende un extremo 5' (5 prima) y un extremo 3' (3-prima). Por lo tanto, los términos 5' y 3' se refieren a una posición relativa en una sola cadena de un ácido nucleico. En consecuencia, la posición relativa de ciertos elementos o secuencias de un ácido nucleico (por ejemplo, una sonda, un código de barras y un anclaje) puede especificarse en un orden secuencial de 5' a 3', o alternativamente de 3' a 5'. Por ejemplo, un ácido nucleico puede incluir, de 5' a 3', una sonda, un código de barras y un anclaje y puede representarse como: 5'-

sonda-código de barras-anclaje-3'. En el ejemplo anterior, puede decirse que el código de barras y el anclaje están en 3' de la sonda. En el ejemplo anterior, puede decirse que la sonda y el código de barras están en 5' del anclaje. Además, la posición de la sonda en el ejemplo anterior también puede considerarse adyacente al código de barras. De manera similar, el código de barras puede considerarse flanqueado por la sonda y el anclaje. En consecuencia, un experto en la técnica sabría lo que significan los términos posicionales 3' y 5'. Dicho término posicional, como se utiliza en la presente memoria, salvo que se indique otra cosa explícitamente, no implica que haya secuencias de ácido nucleico adicionales intercaladas entre los elementos de referencia. Por ejemplo, en el ejemplo anterior, pueden estar presentes secuencias adicionales (por ejemplo, un UMI) entre la sonda y el código de barras.

El término “polímero” como se utiliza en la presente memoria se refiere a cualquier cadena principal de múltiples componentes monoméricos que pueden funcionar para unirse al ligando y/o componente de anclaje seleccionado y utilizarse en un ensayo posterior. Este ensayo puede utilizar la actividad de una o más enzimas, por ejemplo, transcriptasas inversas, ADN o ARN polimerasas, ADN o ARN ligasas, etc. Tales polímeros o componentes monoméricos incluyen oligonucleótidos (por ejemplo, ADN, ARN, bases de ADN o ARN sintéticas o recombinantes o análogos de bases de ADN o ARN), ácidos nucleicos peptídicos (es decir, un análogo de ácido nucleico sintético en el que las bases nucleotídicas naturales están unidas a una cadena principal similar a un péptido en lugar de a la cadena principal de azúcar-fosfato que se encuentra en ADN y ARN), ácidos nucleicos bloqueados (LNA; véase, por ejemplo, Grunweller A y Hartmann RK, “Locked nucleic acid oligonucleotides: the next generation of antisense agents?” *BioDrugs* 2007. 21(4):235-43) o polímeros de poliamida (véase por ejemplo, Dervan, PB y Burli, RW, “Sequence-specific DNA recognition by polyamides”, *Curr. Opin Chem. Biol.* 1999, 3:688-693). Por simplicidad y facilidad de comprensión, a lo largo de esta memoria descriptiva también se puede ilustrar una construcción polimérica o un componente funcional de la misma (por ejemplo, Anclaje, Código de barras, UMI o Sonda de amplificación) como un polímero específico o componente monomérico, tal como una secuencia de oligonucleótidos, un ácido nucleico, una secuencia de ácido nucleico, etc. Sin embargo, cuando se utilice el término “oligonucleótido”, “ácido nucleico” o “nucleótido” o un ejemplo específico similar de un monómero o polímero en esta memoria descriptiva, también debe entenderse que significa que la construcción o componente del polímero puede estar formado de cualquier polímero adecuado como se describe en este párrafo.

Los términos “primero/a”, “adicional” y “sustancialmente idéntico/a/” se usan a lo largo de esta memoria descriptiva como términos de referencia para distinguir entre diversas formas y componentes de las construcciones. Por ejemplo, una “primera construcción” puede definir una construcción con ciertos componentes especificados en la que un único ligando especificado “primero” se une a una “primera” diana específica. El “primer” código de barras es específico del primer ligando; el UMI identifica solamente que la “primera” construcción polimérica y el anclaje se unen a una secuencia complementaria especificada. La expresión “construcción adicional” se refiere a una construcción (por ejemplo, una segunda, tercera o cuarta construcción) que difiere de cualquier otra construcción usada en las composiciones y métodos definidos en la presente memoria por la identidad de la diana, ligando y código de barras. En una realización, una construcción adicional difiere de otras construcciones en las composiciones o métodos por la identidad de la diana, ligando, código de barras, UMI y anclaje. Cada construcción adicional comprende un ligando adicional unido o conjugado a una construcción polimérica adicional mediante un enlazador. El ligando adicional se une específicamente a una diana adicional diferente del de la primera diana. El enlazador entre el ligando y la construcción polimérica adicional puede ser el mismo o diferente del enlazador de la primera construcción. La construcción polimérica adicional también difiere en la identidad de sus elementos funcionales. La sonda de amplificación puede ser la misma o diferente de la utilizada en la primera construcción. Sin embargo, el código de barras adicional que identifica específicamente el ligando adicional no identifica ningún otro ligando. El UMI adicional opcional que se coloca adyacente al código de barras adicional en su extremo 5' o 3' es específico de la construcción polimérica adicional. En otra realización más, el anclaje adicional tiene la misma secuencia o una diferente para hibridarse a la misma secuencia o una secuencia complementaria de captura diferente a la que se une el primer anclaje. En una realización, cada ligando “adicional”, diana “adicional”, código de barras “adicional” y componentes de UMI “adicionales” de cada construcción adicional difiere del componente correspondiente en cualquier otra construcción de la composición o método descrito.

En consecuencia, salvo que se indique lo contrario, los términos “primero/a”, “segundo/a”, “tercero/a”, “cuarto/a”, “quinto/a”, “sexto/a”, “séptimo/a” y “octavo/a” se refieren a un elemento (por ejemplo, construcción, ligando, código de barras, oligonucleótido, oligonucleótido de captura, perla, diana, anclaje, sonda de amplificación y similares), donde los elementos “primero/a”, “segundo/a”, “tercero/a”, “cuarto/a”, “quinto/a”, “sexto/a”, “séptimo/a” y “octavo/a” enumerados pueden ser iguales o diferentes.

La expresión “se une específicamente” o “específicamente se une” se refiere a un ligando que se une a una diana indicada en preferencia a la unión a otras dianas (por ejemplo, otras moléculas, otros péptidos u otros antígenos) según lo determinado en, por ejemplo, un ensayo *in vitro* adecuado (por ejemplo, un Elisa, inmunotransferencia, citometría de flujo y similares). Un ligando que se une específicamente a una diana muestra una interacción de unión específica con la diana que discrimina sobre las interacciones de unión no específicas con otras dianas (por ejemplo, cualquier otra proteína, antígeno, molécula, etc.) en aproximadamente 2 veces o más, a menudo aproximadamente 10 veces o más, y a veces aproximadamente 100 veces o más, 1000 veces o más, 10 000 veces o más, 100 000 veces o más o 1 000 000 de veces o más.

El término “sustancialmente” como se utiliza en la presente memoria, significa al menos el 75 %, al menos el 80 %, al menos el 90 % o al menos el 95 %, a menos que se indique explícitamente lo contrario. En algunas realizaciones, dos o más ácidos nucleicos son sustancialmente idénticos. Dos o más ácidos nucleicos, o partes de los mismos que son sustancialmente idénticos, se refieren a una secuencia de nucleótidos de los dos o más ácidos nucleicos (por ejemplo, dos o más oligonucleótidos, oligonucleótidos de captura, anclajes, sondas de amplificación, códigos de barras y similares) que comparten al menos el 75 %, al menos el 80 %, al menos el 85 %, al menos el 90 %, al menos el 95 %, al menos el 99 % o el 100 % de porcentaje de identidad. El término “idéntico en porcentaje” o “porcentaje de identidad” se refiere a la identidad de secuencia entre dos secuencias de aminoácidos. La identidad se puede determinar comparando una posición en cada secuencia que puede alinearse con fines de comparación. Cuando una posición equivalente en las secuencias comparadas está ocupada por el mismo nucleótido, entonces las secuencias de ácido nucleico son idénticas en esa posición. La expresión como porcentaje de identidad se refiere a una función del número de nucleótidos idénticos, o un derivado o variante del mismo, en posiciones correspondientes (por ejemplo, como se define mediante alineación) compartida por las secuencias comparadas. Se pueden usar diversos algoritmos y/o programas de alineación para determinar el porcentaje de identidad, cuyos ejemplos no limitantes incluyen FASTA, BLAST o ENTREZ. FASTA y BLAST están disponibles como parte del paquete de análisis de secuencias GCG (Universidad de Wisconsin, Madison, WI) y se pueden usar, por ejemplo, en su configuración predeterminada. ENTREZ está disponible a través del National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD.

El término construcción “sustancialmente idéntica” se refiere a un número de construcciones o componentes que difieren de una construcción de referencia, por ejemplo, la construcción “primera” o una construcción adicional específica, solamente en la secuencia de un identificador molecular único opcional o su ausencia de la construcción. En una realización, cada una de una construcción sustancialmente idéntica comparte la misma diana, ligando, sonda de amplificación, código de barras y anclaje como lo hace la construcción de referencia (primera o adicional). En otra realización, cada una de las construcciones sustancialmente idénticas comparte la misma diana, ligando, código de barras y anclaje que la construcción de referencia (primera o adicional). En una realización, una construcción sustancialmente idéntica a la “primera construcción” difiere de la “primera” construcción de referencia en la secuencia y/o presencia del UMI. En otra realización, la construcción adicional sustancialmente idéntica difiere de la construcción adicional de referencia en el UMI y la sonda de amplificación.

El término “unión” o “unir” como se utiliza en la presente memoria para describir la interacción entre los componentes de las construcciones responde a uniones covalentes o una variedad de tipos de unión no covalentes. Otras químicas de unión útiles en el ensamblaje de las construcciones descritas en la presente memoria incluyen, aunque no de forma limitativa, tiol-maleimida, tiol-haloacetato, amina-NHS, amina-isotiocianato, azida-alquino (CuAAC), tetrazol-cicloocteno (IEDDA) (véanse, por ejemplo, la referencia 24 y otras referencias en dicho documento). En una realización, cada construcción polimérica está unida al ligando mediante un enlace covalente irreversible. En otra realización, cada construcción polimérica está unida al ligando mediante un enlace covalente escindible, por ejemplo un enlace disulfuro o un enlazador fotoescindible.

Como se utiliza en la presente memoria, “diana” se refiere a cualquier molécula biológica o química natural o sintética. En una realización, la diana se refiere a cualquier molécula biológica o química expresada en la superficie de una célula. En algunas realizaciones, una diana se refiere a cualquier molécula biológica o química en la superficie o en el interior de un exosoma, un núcleo, un orgánulo celular, un virus o una bacteria. En ciertas realizaciones, una diana es una proteína de la superficie celular. En algunas realizaciones, una diana es una célula. En algunas realizaciones, una diana es un núcleo, exosoma, bacteria o fago. En otra realización, la diana se refiere a cualquier molécula biológica o química expresada intracelularmente. En otra realización, la diana se refiere a cualquier molécula biológica o química que se produce en una genoteca, panel o mezcla de dianas de origen natural, sintético, diseñado recombinantemente o aislada. En otra realización, la diana se refiere a cualquier molécula biológica o química que se produce en una muestra biológica. Los términos correspondientes “primera diana” y cada “diana adicional” (por ejemplo, una segunda, tercera, cuarta diana o similares) se refieren a diferentes dianas. Las dianas primera y adicional se pueden seleccionar independientemente de un péptido, una proteína, un anticuerpo o fragmento de anticuerpo, un aficuerpo, una secuencia de ácido ribonucleico o una secuencia de ácido desoxirribonucleico, un aptámero, un lípido, un polisacárido, una lectina o una molécula quimérica formada por varios de los anteriores o diferentes dianas. En los ejemplos siguientes, las dianas son antígenos o epítomos de la superficie celular.

En algunas realizaciones, una muestra es una muestra biológica. Como se utiliza en la presente memoria, una “muestra biológica” como se utiliza en los métodos descritos en la presente memoria se refiere a una muestra de origen natural o bien una muestra diseñada deliberadamente o sintetizada o una genoteca que contiene una o más dianas seleccionadas. En una realización, una muestra contiene una población de células o fragmentos celulares, que incluyen sin limitación componentes de membrana celular, exosomas y componentes subcelulares. Las células pueden ser una población homogénea de células, tales como células aisladas de un tipo particular, o una mezcla de diferentes tipos de células, tal como procedentes de un fluido biológico o tejido de un sujeto humano o mamífero o un sujeto de otra especie. Otras muestras adicionales para usar en los métodos y con las composiciones incluyen, sin limitación, muestras de sangre que incluyen suero, plasma, sangre completa y sangre periférica, saliva, orina, secreciones vaginales o del cuello de útero, líquido amniótico, fluido placentario, líquido cefalorraquídeo o fluidos serosos, secreciones mucosas (por ejemplo, bucal, vaginal o rectal). Otras muestras adicionales incluyen una muestra biológica

derivada de sangre o derivada de una biopsia de tejido o un lisado celular (es decir, una mezcla derivada de tejido y/o células). Otro tejido adecuado incluye el cabello, las uñas y similares. Otras muestras adicionales incluyen genotecas de anticuerpos, fragmentos de anticuerpos y miméticos de anticuerpos como aficuerpos. Dichas muestras pueden estar adicionalmente diluidas con solución salina, tampón o un diluyente fisiológicamente aceptable. Alternativamente, dichas muestras se concentran por medios convencionales. Otras muestras adicionales pueden ser sintéticas o colecciones diseñadas de moléculas químicas, proteínas, anticuerpos o cualesquiera otras de las dianas descritas en la presente memoria. A menudo, una muestra se obtiene a partir de, o se deriva de, una fuente, sujeto o paciente específico. En algunas realizaciones, una muestra se obtiene a menudo a partir de, se deriva de, o está asociada a un experimento, lote, ciclo o repetición específico. En consecuencia, en ciertas realizaciones, cada una de una pluralidad de muestras (por ejemplo, muestras derivadas de diferentes fuentes, diferentes sujetos o diferentes ciclos, por ejemplo) se puede identificar y/o diferenciar mediante el uso de un método o composición descritos en la presente memoria. En algunas realizaciones, una muestra se detecta, se rastrea, se etiqueta y/o se identifica por un método de *hashtagging* que se describe en la presente memoria. En algunas realizaciones, la presencia, cantidad o ausencia de una muestra se determina por un método de *hashtagging* que se describe en la presente memoria. En ciertas realizaciones, una diana (por ejemplo, una célula, núcleo, proteína, etc.) que se deriva de una muestra específica, o fuente, se detecta, se rastrea, se etiqueta y/o se identifica por un método de *hashtagging* que se describe en la presente memoria.

En ciertas realizaciones, una muestra comprende uno o más orgánulos, mitocondrias, exosomas, liposomas, vesículas sintéticas o naturales, microvesículas, ectosomas, núcleos, bacterias, virus, fagos, perlas, partículas, micropartículas, nanopartículas, macromoléculas, lípidos o membranas sintéticas o naturales, membranas de fosfolípidos, esferas de membrana, similares, o combinaciones de los mismos. En algunas realizaciones, una muestra comprende una o más células. En algunas realizaciones, una muestra comprende uno o más núcleos. Una o más dianas pueden estar presentes en, o en la superficie de, o unidas covalentemente o no covalentemente a, un orgánulo, mitocondria, exosoma, liposoma, una vesícula sintética o natural, una microvesícula, una macrovesícula, un ectosoma, un núcleo, una bacteria, un virus, un fago, una perla, una partícula, una micropartícula, una nanopartícula, una macromolécula, una membrana lipídica sintética o natural, una membrana de fosfolípidos o una membrana o esfera lipídica.

El “ligando” usado en estas composiciones y métodos se refiere a cualquier molécula biológica o química natural o sintética que se usa para unirse específicamente a una única diana identificada. La unión puede ser covalente o no covalente, es decir, conjugada o por cualquier medio conocido que tenga en cuenta la naturaleza del ligando y su diana respectiva. Las expresiones “primer ligando” y “ligando adicional” se refieren a ligandos que se unen a diferentes dianas o diferentes porciones de una diana. Por ejemplo, múltiples “primeros ligandos” se unen a la misma diana en el mismo sitio. Múltiples ligandos adicionales se unen a una diana diferente del primer ligando y diferente de cualquier ligando adicional. Un ligando (por ejemplo, un primer ligando, y ligandos adicionales, por ejemplo, un segundo, tercer, cuarto y quinto ligandos, etc.) se pueden seleccionar independientemente de un péptido, una proteína, un anticuerpo o fragmento de anticuerpo (por ejemplo, una porción de unión a antígeno de un anticuerpo), un mimético de anticuerpo, un aficuerpo, una secuencia de ácido ribonucleico o desoxirribonucleico, un aptámero, un lípido, un polisacárido, una lectina o una molécula quimérica formada por varios de los anteriores o por diferentes ligandos. Ejemplos adicionales no limitativos de un ligando incluyen un Fab, Fab', F(ab')₂, fragmento Fv, Fv monocatenario (scFv), diacuerpo (dAb), sincuerpo, nanocuerpos, BITE, SMIP, DARPin, DNL, duocalinas, adnectinas, finómeros, dominios AlbuDabs de Kunitz, DART, DVD-Ig, cuerpos Covx, pepticuerpos, scFv-Ig, SVD-Ig, dAb-Ig, Knob-in-Holes, triomAb, similares o combinaciones de los mismos. En algunas realizaciones, un ligando es una proteína recombinante o natural. En ciertas realizaciones, un ligando es un anticuerpo monoclonal o policlonal, o fragmento del mismo. En una realización, el ligando o ligandos de las construcciones también pueden marcarse directamente con uno o más marcadores detectables, tales como fluoróforos (véanse los marcadores que se analizan a continuación) que se pueden medir por métodos independientes de los métodos de medición o detección de la construcción polimérica descrita de cualquier otra manera en la presente memoria.

Un fragmento de anticuerpo o fragmento de unión a antígeno de un anticuerpo se refiere a una porción de un anticuerpo que se une específicamente a una diana y puede incluir un Fab, Fab', F(ab')₂, fragmento Fv, Fv monocatenario (scFv), scFv-Ig y otros fragmentos o porciones de un anticuerpo que se puede unir específicamente a una diana.

Como se utiliza en la presente memoria, el término “marcador detectable” significa un reactivo, resto o compuesto que puede proporcionar una señal detectable, dependiendo del formato de ensayo empleado. Un marcador puede asociarse con la construcción en su conjunto, o solamente con el ligando, o con la construcción polimérica o una porción funcional de la misma. Alternativamente, se pueden usar diferentes marcadores para cada componente de la construcción. Dichos marcadores pueden, tanto en solitario como concertados con otras composiciones o compuestos, proporcionar una señal detectable. En una realización, los marcadores son deseablemente interactivos para producir una señal detectable. De forma más deseable, el marcador es detectable visualmente, por ejemplo, colorimétricamente. Una variedad de sistemas enzimáticos funcionan para revelar una señal colorimétrica en un ensayo, por ejemplo, la glucosa oxidasa (que usa glucosa como sustrato) libera peróxido como producto que en presencia de peroxidasa y un donante de hidrógeno tal como tetrametilbencidina (TMB) produce TMB oxidada que aparece de color azul. Otros ejemplos incluyen la peroxidasa de rábano picante (HRP) o la fosfatasa alcalina (AP) y la hexoquinasa junto con la glucosa-6-fosfato deshidrogenasa que reacciona con ATP, glucosa y NAD⁺ para producir,

entre otros productos, NADH que se detecta como un aumento de la absorbancia a una longitud de onda de 340 nm. Otros sistemas de marcadores adicionales que se pueden utilizar en los métodos y construcciones descritos se pueden detectar por otros medios, por ejemplo, micropartículas de látex coloreadas (Bangs Laboratories, Indiana) a las que se incorpora un tinte en lugar de enzimas para proporcionar una señal visual indicativa de la presencia del ligando o construcción marcado en ensayos aplicables. Otros marcadores adicionales incluyen compuestos fluorescentes, fluoróforos, compuestos o elementos radiactivos. En una realización, un fluorocromo fluorescente detectable, por ejemplo, isotiocianato de fluoresceína (FITC), ficoeritrina (PE), alofocianina (APC), corifosfina-O (CPO) o colorantes en tándem, PE-cianina-5 o PE-cianina-7 (PC5 o PC7)), PE-Texas Red (ECD), PE-cianina-5.5, rodamina, PerCP y colorantes Alexa. Se pueden usar combinaciones de dichos marcadores, tales como Rojo Texas y rodamina, FITC + PE, FITC + PECy5 y PE + PECy7, entre otros dependiendo del método de ensayo. La selección y/o generación de marcadores adecuados para su uso en el marcado del ligando y/o cualquier componente de la construcción polimérica está dentro de la experiencia de la técnica, provista en esta memoria descriptiva.

Otros componentes de las composiciones y métodos descritos en la presente memoria también pueden marcarse de forma detectable. De manera adicional o alternativa al marcado del ligando, la una o más construcciones poliméricas se pueden marcar con uno o más marcadores detectables, tales como fluoróforos y otros marcadores definidos a continuación. La detección de estos marcadores se realiza por métodos independientes de los métodos descritos en la presente memoria para la medición de la construcción polimérica o sus componentes. De manera adicional o alternativa, el ligando y la una o más construcciones poliméricas pueden marcarse de manera que, cuando se ensamblan en la construcción final, se puede detectar el ensamblaje correcto, tal como para la producción de la construcción final. De manera adicional o alternativa, en los métodos descritos a continuación, el polímero de captura se puede marcar con uno o más marcadores detectables. De manera adicional o alternativa, se pueden usar marcadores detectables en los métodos descritos a continuación, para proporcionar indicaciones de unión correcta. Por ejemplo, el sustrato al que se inmoviliza el polímero de captura se puede marcar con uno o más marcadores detectables. De manera adicional o alternativa, se pueden usar uno o más marcadores detectables para mostrar la unión correcta entre el polímero de captura y la construcción polimérica. En otra realización, se puede marcar la unión correcta entre el polímero de captura y el sustrato. De manera adicional o alternativa, la asociación correcta entre la construcción polimérica y el sustrato al que se inmoviliza el polímero de captura se puede marcar con uno o más marcadores detectables. Además, tales marcadores se pueden usar para indicar la asociación correcta entre el ligando y el polímero de captura. De manera adicional o alternativa, tales marcadores se pueden usar para indicar la asociación entre el ligando y el sustrato al que se inmoviliza el polímero de captura. Se contemplan otros usos adicionales de los marcadores detectables en estos métodos y composiciones.

Como se utiliza en la presente memoria, un “anticuerpo o fragmento” es un anticuerpo monoclonal, un anticuerpo sintético, un anticuerpo recombinante, un anticuerpo quimérico, un anticuerpo humanizado, un anticuerpo humano, un anticuerpo injertado con CDR, una construcción de unión multispecífica que se puede unir a dos o más dianas, un anticuerpo específico dual, un anticuerpo biespecífico o un anticuerpo multispecífico, o un anticuerpo madurado por afinidad, un anticuerpo monocatenario o un fragmento scFv, un diacuerpo, una cadena individual que comprende scFv complementarios (scFv en tándem) o scFv en tándem biespecíficos, una construcción Fv, un Fv unido por disulfuro, una construcción Fab, una construcción Fab', una construcción F(ab')₂, una construcción de Fc, una construcción monovalente o bivalente de la cual se han eliminado los dominios no esenciales para la función del anticuerpo monoclonal, una molécula monocaténaria que contiene un dominio de unión a antígeno V_L o V_H y uno o dos dominios “efectores” constantes opcionalmente conectados por dominios enlazadores, un anticuerpo univalente que carece de una región bisagra, un anticuerpo de dominio único, una proteína de unión de inmunoglobulina de dominio variable dual (DVD-Ig) o un nanocuerpo. También se incluyen en esta definición miméticos de anticuerpos tales como aficuerpos, es decir, una clase de proteínas de afinidad diseñadas, generalmente proteínas pequeñas de dominio único (~6,5 kDa) que pueden aislarse con alta afinidad y especificidad para cualquier diana proteica dada.

El “enlazador” se refiere a cualquier resto usado para unir o asociar el ligando a la porción de construcción polimérica/secuencia de oligonucleótidos de las construcciones. Por lo tanto, en una realización, el enlazador es un enlace covalente. En otra realización, el enlazador es un enlace no covalente. En otra realización, el enlazador está compuesto de al menos uno a aproximadamente 25 átomos. Por lo tanto, en diversas realizaciones, el enlazador está formado por al menos 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24 o 25 átomos. En otra realización más, el enlazador tiene al menos de uno a aproximadamente 60 ácidos nucleicos. Por lo tanto, en diversas realizaciones, el enlazador está formado por una secuencia de al menos 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, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, hasta 60 ácidos nucleicos. En otra realización adicional, el enlazador se refiere a de al menos uno a aproximadamente 30 aminoácidos. Por lo tanto, en diversas realizaciones, el enlazador está formado por una secuencia de al menos 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, hasta aproximadamente 30 aminoácidos. En otras realizaciones más, el enlazador puede ser un compuesto más grande o dos o más compuestos asociados de manera covalente o no covalente. En otras realizaciones más, el enlazador puede ser una combinación de los enlazadores definidos en la presente memoria. Los enlazadores usados en las construcciones de las composiciones y métodos son, en una realización, escindibles. Los enlazadores usados en las construcciones de las composiciones y métodos son, en una realización, no escindibles. Sin limitación, en una realización, el enlazador es un enlazador escindible, por ejemplo, un enlace disulfuro o un enlace fotoescindible. En los ejemplos siguientes, el enlazador ilustrado comprende un complejo de

biotina unido a la secuencia de oligonucleótidos de la construcción mediante un enlace disulfuro, con estreptavidina fusionada al ligando. En otra realización, la biotina se une al ligando y la estreptavidina se fusiona con la secuencia de oligonucleótidos de la construcción. Aunque los ejemplos muestran el enlazador ilustrado unido al extremo 5' del oligonucleótido de la construcción, en otras realizaciones, el enlazador puede estar covalentemente unido o conjugado de forma que no sea covalente a cualquier parte de la secuencia de oligonucleótido de la construcción. En otra realización adicional, cuando el ligando es un anticuerpo recombinante o sintético, el enlazador puede diseñarse dentro de la secuencia de anticuerpo para facilitar el acoplamiento 1:1 a la construcción polimérica, simplificando así la fabricación del ligando, de la construcción y/o de la construcción polimérica. Por ejemplo, un enlazador Halotag® se puede diseñar dentro del ligando seleccionado (por ejemplo, un anticuerpo) o en la construcción polimérica o componente con dicho fin. De manera adicional o alternativa, el ligando se une a la construcción polimérica tras su producción en la misma célula. Véanse, por ejemplo, los protocolos Halotag® descritos en el Flexi® Vector Systems Technical Manual (TM254 - revisado el 5/17), copyright 2017 de Promega Corporation; y en Janssen D.B., "Evolving haloalkaline dehalogenase", *Curr. Opin. Chem. Biol.*, 2004, 8:150-159.

La "construcción polimérica" o "secuencia de oligonucleótidos de la construcción" es la parte de la construcción que está asociada con el ligando. Como se ha indicado anteriormente, esta asociación puede ser covalente, no covalente o mediante cualquier conjugación adecuada y emplear cualquier enlazador adecuado. La construcción polimérica está formada por una serie de elementos poliméricos funcionales, por ejemplo, secuencias de ácido nucleico u otros polímeros como se ha definido anteriormente, teniendo cada uno una función como se define en la presente memoria. El ligando se puede unir a la secuencia de oligonucleótidos de la construcción en su extremo 5' o en cualquier otra porción, siempre que la unión o conjugación no impida las funciones de los componentes de la secuencia de oligonucleótidos de la construcción. Según se ha descrito anteriormente, estos componentes son, para cada secuencia de oligonucleótidos de la construcción "primera" o "adicional", una sonda de amplificación; un código de barras, un UMI opcional y un anclaje. En general, la construcción polimérica puede tener cualquier longitud que permita las longitudes de sus componentes funcionales. En una realización, la construcción polimérica tiene entre 20 y 100 componentes monoméricos, por ejemplo, bases de ácido nucleico, de longitud. En algunas realizaciones, la secuencia de oligonucleótidos de la construcción tiene al menos 20, 30, 40, 50, 60, 70, 80, 90 o más de 100 componentes monoméricos, por ejemplo, bases de ácido nucleico, de longitud. En otras realizaciones, el oligonucleótido de la construcción tiene de 200 a aproximadamente 400 componentes monoméricos, por ejemplo, nucleótidos, de longitud. En una realización, la construcción polimérica generalmente se compone de ácidos desoxirribonucleicos (ADN). En una realización, el oligonucleótido de la construcción es una secuencia de ADN. En otras realizaciones, el oligonucleótido de la construcción, o partes del mismo, comprende bases de ADN modificadas. La modificación de las bases de ADN es conocida en la técnica y puede incluir bases modificadas químicamente que incluyen marcadores. En otras realizaciones, el oligonucleótido de la construcción, o partes del mismo, comprende secuencias de ácido ribonucleico (ARN) o bases de ribonucleótidos modificadas. La modificación de las bases de ARN es conocida en la técnica y puede incluir bases modificadas químicamente que incluyen marcadores. En otras realizaciones más, diferentes partes de la secuencia de oligonucleótidos de la construcción pueden comprender ADN y ARN, bases modificadas o conexiones de polímeros modificados (que incluyen, aunque no de forma limitativa, los PNA y LNA). Para ver una descripción de modificaciones realizadas en oligonucleótidos, véanse los proveedores comerciales, por ejemplo, el sitio web estadounidense de Integrated DNA Technologies; Custom Oligonucleotide Modifications Guide, Sigma-Aldrich, www.sigmaaldrich.com/technical-documents/articles/biology/custom-dna-oligos-modifications.html, así como Modified Oligonucleotides, TriLink, www.trilinkbiotech.com/oligo/modifiedoligos.asp. Como se ha descrito anteriormente, en otras realizaciones más, la construcción polimérica está compuesta por poliamidas, PNA, etc.

Como se utiliza en la presente memoria, la expresión "sonda de amplificación" se refiere a un componente funcional de la secuencia de oligonucleótidos de la construcción que en sí misma es una secuencia de oligonucleótidos o polinucleótidos que proporciona un sitio de hibridación para la amplificación de la secuencia de oligonucleótidos de la construcción. La sonda de amplificación puede estar formada por polímeros de ADN, ARN, PNA, bases modificadas o combinaciones de estas bases, o poliamidas, etc. En una realización, la sonda de amplificación tiene aproximadamente 10 de dichos componentes monoméricos, por ejemplo, bases de nucleótidos, de longitud. En otras realizaciones, la sonda de amplificación tiene al menos de aproximadamente 5 a 100 componentes monoméricos, por ejemplo, nucleótidos, de longitud. Por lo tanto, en diversas realizaciones, la sonda de amplificación está formada por una secuencia de al menos 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, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99 o hasta 100 componentes monoméricos, por ejemplo, ácidos nucleicos. En una realización, cuando está presente en secuencias de oligonucleótidos de la construcción primeras o adicionales, la sonda de amplificación puede ser igual o diferente, dependiendo de las técnicas destinadas a usarse para la amplificación. En ciertas realizaciones, la sonda de amplificación puede ser una secuencia genérica adecuada como sitio de hibridación para una variedad de tecnologías de amplificación. Las tecnologías de amplificación incluyen, aunque no de forma limitativa, sistemas de amplificación basados en ADN polimerasa, tales como reacción en cadena de la polimerasa (PCR), PCR en tiempo real, amplificación isotérmica mediada por bucle (LAMP, MALBAC), amplificación por desplazamiento de cadena (SDA), amplificación por desplazamiento múltiple (MDA), amplificación de polimerasa/recombinasa (RPA) y polimerización mediante cualquier número de ADN polimerasas (por ejemplo, ADN polimerasa T4, ADN polimerasa Sulfolobus, ADN polimerasa de Klenow, polimerasa Bst, polimerasa Phi29) y sistemas de amplificación basados en ARN polimerasa (tales como la amplificación mediante T7-ARN polimerasa, T3-

ARN polimerasa y SP6-ARN polimerasa), la amplificación basada en la secuencia de ácido nucleico (NASBA), la replicación de secuencia autosostenida (3SR), la amplificación por círculo rodante (RCA), la reacción en cadena de la ligasa (LCR), la amplificación dependiente de helicasa (HDA), el método de amplificación por ramificación y ARN-sec²³. Véase también, por ejemplo, la referencia 27.

El término “código de barras” o “código de barras de la construcción” describe un polímero definido, por ejemplo, un polinucleótido, que cuando es un elemento funcional de la construcción polimérica, es específico de un único ligando. Como se utiliza en los diversos métodos descritos en la presente memoria, el término Código de barras puede ser un “código de barras celular” o un “código de barras de sustrato”, que describe un polinucleótido definido específico para identificar una célula o sustrato en particular, por ejemplo, microperla de Drop-seq. En cualquiera de las realizaciones, el código de barras puede estar formado por una secuencia definida de ADN, ARN, bases modificadas o combinaciones de estas bases, así como cualquier otro polímero definido anteriormente. En una realización, el código de barras tiene de aproximadamente 2 a 4 componentes monoméricos, por ejemplo, bases de nucleótidos, de longitud. En otras realizaciones, el código de barras tiene de al menos aproximadamente 1 a 100 componentes monoméricos, por ejemplo, nucleótidos, de longitud. Por lo tanto, en diversas realizaciones, el código de barras está formado por una secuencia de al menos 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, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99 o hasta 100 componentes monoméricos, por ejemplo, ácidos nucleicos.

La expresión “Identificador molecular único” (UMI), también llamado de manera equivalente “Etiqueta molecular aleatoria” (RMT), es una secuencia aleatoria de componentes monoméricos de un polímero como se ha descrito anteriormente, por ejemplo, bases de nucleótidos que, cuando es un elemento funcional de la construcción polimérica, es específica de esa construcción polimérica. El UMI permite la identificación de duplicados de amplificación de la construcción polimérica/secuencia de oligonucleótidos de la construcción con la que está asociado. En la descripción de los métodos y composiciones en la presente memoria, uno o más UMI se pueden asociar con una única construcción polimérica/secuencia de oligonucleótidos de la construcción. El UMI puede estar situado en 5' o 3' respecto del código de barras de la composición. En otra realización, el UMI puede insertarse en la secuencia de oligonucleótidos de la construcción/construcción polimérica como parte de los métodos descritos. En una realización de los métodos descritos en la presente memoria, dependiendo de qué método de secuenciación de ARN se use, se añade un UMI durante el método. Sin embargo, no todos los métodos de ARN-sec utilizan los UMI. En el ejemplo de secuenciación de ARN en gotícula de células individuales descrito más adelante, se introduce otro UMI durante la retrotranscripción. Cada UMI es específico de su secuencia de oligonucleótidos de la construcción. Por lo tanto, cuando las composiciones o métodos comprenden múltiples “primeras construcciones”, cada primera construcción difiere únicamente por la secuencia de su UMI. Cada construcción adicional también tendrá su propio UMI, que no está presente en las construcciones adicionales duplicadas o en las construcciones adicionales que difieren entre sí por la diana, ligando, código de barras y especificidad de anclaje. De manera similar a como se utiliza en los diversos métodos descritos en la presente memoria, un UMI puede asociarse con un polímero, por ejemplo, una secuencia de oligonucleótidos o polinucleótidos, usado en un formato de ensayo particular, o bien con un polímero, por ejemplo, un oligonucleótido o polinucleótido, que está inmovilizado sobre un sustrato. Cada UMI de cada construcción polimérica, por ejemplo, oligonucleótido o polinucleótido, es diferente de cualquier otro UMI usado en las composiciones o métodos. En cualquiera de las realizaciones, el UMI está formado por una secuencia aleatoria de ADN, ARN, bases modificadas o combinaciones de estas bases u otros monómeros de los polímeros identificados anteriormente. En una realización, un UMI tiene aproximadamente 8 componentes monoméricos, por ejemplo, nucleótidos, de longitud. En otras realizaciones, cada UMI puede tener al menos de aproximadamente 1 a 100 componentes monoméricos, por ejemplo, nucleótidos, de longitud. Por lo tanto, en diversas realizaciones, el UMI está formado por una secuencia aleatoria de al menos 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, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99 o hasta 100 componentes monoméricos, por ejemplo, ácidos nucleicos.

Como se utiliza en la presente memoria, el término “Anclaje” se refiere a un polímero definido, por ejemplo, una secuencia de polinucleótidos u oligonucleótidos, que está diseñada para hibridarse con otra secuencia de oligonucleótidos, por ejemplo, un polímero de captura, un oligonucleótido de captura, un cebador y similares. En una realización de la construcción polimérica, el anclaje se diseña para generar una secuencia bicatenaria del oligonucleótido de la construcción. En algunas realizaciones, el anclaje se coloca en el extremo 3' de una secuencia de oligonucleótidos (por ejemplo, una secuencia de oligonucleótidos de la construcción). En otras realizaciones, el anclaje se coloca en el extremo 5' de una secuencia de oligonucleótidos de la construcción. En algunas realizaciones, cada anclaje es específico de su secuencia complementaria prevista. Por ejemplo, en ciertas realizaciones, el anclaje está configurado para hibridarse con el extremo 3' de un oligonucleótido de captura de manera que el extremo 3' del oligonucleótido de captura actúa como un cebador que puede generar una segunda hebra complementaria del oligonucleótido en presencia de una polimerasa. En ciertas realizaciones, cuando las composiciones o métodos comprenden múltiples “primeras construcciones”, cada primera construcción tiene la misma secuencia de anclaje. En una realización, cada anclaje adicional tiene una secuencia adicional diferente que se hibrida con una secuencia complementaria diferente. En otras realizaciones, cada anclaje adicional puede tener la misma secuencia de anclaje que la primera u otras construcciones, dependiendo de los pasos del método de ensayo. Cuando se usa en los diversos

métodos descritos en la presente memoria, un anclaje puede hibridarse con una secuencia complementaria libre o con una secuencia complementaria que está inmovilizada sobre un sustrato. En ciertas realizaciones, el anclaje puede estar formado por una secuencia de monómeros del polímero seleccionado, por ejemplo, ADN, ARN, bases modificadas o combinaciones de estas bases, PNA, poliamidas, etc. En una realización, un anclaje tiene de aproximadamente 3 a 15 componentes monoméricos, por ejemplo, nucleótidos, de longitud. En otras realizaciones, cada anclaje puede tener al menos de aproximadamente 3 a 100 componentes monoméricos, por ejemplo, nucleótidos, de longitud. En algunas realizaciones, un anclaje comprende de 3 a 100, de 3 a 50, de 3 a 30, de 5 a 30, de 10 a 20, de 5 a 20 o de 5 a 15 componentes monoméricos (por ejemplo, nucleótidos de longitud). En diversas realizaciones, el anclaje está formado por una secuencia de al menos 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, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99 o hasta 100 componentes monoméricos, por ejemplo, ácidos nucleicos. En algunas realizaciones, y como se muestra en los ejemplos, una secuencia de anclaje comprende o consiste en una secuencia de poli A. En ciertas realizaciones, una secuencia de poli A comprende una secuencia de ácido nucleico que comprende diez o más (por ejemplo, 10-40, 10-30 o 10-20) nucleótidos de adenosina consecutivos, derivados o variantes de un nucleótido de adenosina, análogos, o una combinación de los mismos. En otra realización, una secuencia de anclaje comprende o consiste en una secuencia de poli T. En otra realización, una secuencia de anclaje es una secuencia de poli G. En otra realización más, una secuencia de anclaje puede ser una secuencia aleatoria siempre que pueda hibridarse con su secuencia complementaria prevista (por ejemplo, un oligonucleótido de captura, cebador de amplificación o similares). Por ejemplo, en algunas realizaciones, un método descrito en la presente memoria puede utilizar una pluralidad de oligonucleótidos (por ejemplo, una pluralidad de construcciones que comprenden un ligando unido a un oligonucleótido), donde algunos o todos los oligonucleótidos comprenden un anclaje diferente (es decir, un anclaje que tiene una secuencia de ácido nucleico diferente o un anclaje que tiene una secuencia de ácido nucleico sustancialmente diferente). En algunas realizaciones, un método descrito en la presente memoria puede utilizar una pluralidad de oligonucleótidos (por ejemplo, una pluralidad de construcciones que comprenden un ligando unido a un oligonucleótido), donde algunos o todos los oligonucleótidos comprenden el mismo anclaje. En algunas realizaciones, un método descrito en la presente memoria puede utilizar una pluralidad de oligonucleótidos (por ejemplo, una pluralidad de construcciones que comprenden un ligando unido a un oligonucleótido), donde algunos o todos los oligonucleótidos comprenden un anclaje que es sustancialmente idéntico (por ejemplo, que comprende una secuencia de ácido nucleico que es sustancialmente idéntica). En algunas realizaciones, un método descrito en la presente memoria puede utilizar una pluralidad de oligonucleótidos (por ejemplo, una pluralidad de construcciones que comprenden un ligando unido a un oligonucleótido), donde algunos o todos los oligonucleótidos comprenden un anclaje que comprende una secuencia de poli A. En algunas realizaciones, la secuencia de poli A de una pluralidad de anclajes es sustancialmente idéntica. Como entenderá un experto en la técnica, las secuencias de poli A que son sustancialmente idénticas pueden diferir sustancialmente en longitud. En algunas realizaciones, una secuencia de poli A (por ejemplo, una secuencia de poli A de un anclaje) es un ácido nucleico configurado para hibridarse con una secuencia de poli T (por ejemplo, un oligonucleótido u oligonucleótido de captura que comprende una secuencia de poli T). Como entenderá un experto en la técnica, dependiendo de las condiciones de hibridación, una secuencia de poli A puede comprender uno, dos, tres o cuatro nucleótidos que no sean de poli A y seguir hibridándose eficazmente con una secuencia de poli T, proporcionando de este modo un complejo de poli A-poli T hibridado que comprende uno, dos, tres o más emparejamientos incorrectos. En consecuencia, en algunas realizaciones, una secuencia de poli A es una secuencia de ácido nucleico que comprende al menos 70 %, al menos 75 %, al menos 80 %, al menos 85 %, al menos 90 %, al menos 95 % o 100 % de nucleótidos de adenosina, análogos de adenosina, variantes de adenosina o una combinación de los mismos.

En algunas realizaciones, un oligonucleótido comprende una secuencia de poli T. En algunas realizaciones, un oligonucleótido de captura comprende una secuencia de poli T (por ejemplo, una secuencia de poli T en 3'). En algunas realizaciones, un método descrito en la presente memoria puede utilizar una pluralidad de oligonucleótidos (por ejemplo, una pluralidad de oligonucleótidos de captura), donde algunos o todos los oligonucleótidos comprenden una secuencia de poli T. En algunas realizaciones, una secuencia de poli T de una pluralidad de oligonucleótidos es sustancialmente idéntica. En algunas realizaciones, una pluralidad de oligonucleótidos de captura (por ejemplo, una pluralidad de oligonucleótidos de captura diferentes, por ejemplo, oligonucleótidos de captura específicos de perla diferentes) comprenden una secuencia de poli T que es sustancialmente idéntica. Como entenderá un experto en la técnica, las secuencias de poli T que son sustancialmente idénticas pueden diferir sustancialmente en longitud. En algunas realizaciones, una secuencia de poli T comprende de 3 a 100, de 3 a 50, de 3 a 30, de 5 a 30, de 10 a 20, de 5 a 20 o de 5 a 15 nucleótidos consecutivos (por ejemplo, nucleótidos de longitud). En ciertas realizaciones, una secuencia de poli T comprende una secuencia de ácido nucleico que comprende tres o más, diez o más, 3 a 100, 3 a 50, 3 a 30, 5 a 30, 10 a 20, 5 a 20 o 5 a 15 nucleótidos de timidina consecutivos, derivados o variantes de un nucleótido de timidina, similares o una combinación de los mismos. En algunas realizaciones, una secuencia de poli T (por ejemplo, una secuencia de poli T de un oligonucleótido de captura) es un ácido nucleico configurado para hibridarse con una secuencia de poli A. Como entenderá un experto en la técnica, dependiendo de las condiciones de hibridación, una secuencia de poli T puede comprender uno, dos, tres o cuatro nucleótidos que no sean de timidina y seguir hibridándose eficazmente con una secuencia de poli A, proporcionando de este modo un complejo de poli A-poli T hibridado que comprende uno, dos, tres o más emparejamientos incorrectos. En consecuencia, en algunas realizaciones, una secuencia de poli T es una secuencia de ácido nucleico que comprende al menos 70 %, al menos 75 %, al menos 80 %, al menos 85 %, al menos 90 %, al menos 95 % o 100 % de nucleótidos de timidina, análogos

de timidina, variantes de timidina o una combinación de los mismos. En algunas realizaciones, una secuencia de poli T comprende uno o más nucleótidos de uracilo o derivados de los mismos.

El “oligonucleótido de captura” u “oligo de captura” o “polímero de captura” es una secuencia polimérica, por ejemplo, un oligonucleótido, que comprende al menos una secuencia que es complementaria de a un anclaje. En algunas realizaciones, el polímero/oligo de captura no forma parte de las construcciones primeras o adicionales; en su lugar, es cualquier secuencia polimérica o de oligonucleótido que pertenece a un kit de construcción-purificación o un kit de secuenciación de ARNm. Como se utiliza en la presente memoria, la expresión “secuencia complementaria” se refiere a una secuencia con la que debe hibridarse una secuencia de anclaje (o de otro ácido nucleico, por ejemplo, un cebador u oligonucleótido de captura), a menudo dando como resultado un complejo bicatenario hibridado. En presencia de una polimerasa, un complejo hibridado a menudo puede extenderse en dirección 3' cuando está presente un molde de ácido nucleico. En consecuencia, en ciertas realizaciones, una secuencia complementaria de un anclaje puede hibridarse con una secuencia de anclaje, proporcionando así un cebador de amplificación y/o generar una secuencia bicatenaria. En ciertas realizaciones, la secuencia de polímero/oligonucleótido de captura puede contener secuencias que se pueden usar como sondas de amplificación y opcionalmente una o más secuencias del identificador molecular único y de código de barras. En los métodos descritos más adelante, la extensión del polímero/oligonucleótido de captura con su secuencia complementaria hibridada a la secuencia de anclaje copia el código de barras, el UMI y la sonda de amplificación desde las construcciones primera o adicional hasta el polímero/oligonucleótido de captura. En cualquiera de las realizaciones, el polímero/oligonucleótido de captura y su secuencia complementaria pueden estar formados por ADN, ARN, bases modificadas o combinaciones de estas bases o de cualquier otro componente polimérico como se ha definido anteriormente. Dependiendo de los pasos del ensayo involucradas y la diana prevista, es posible que la secuencia de captura no esté impedida o que esté “libre” en la muestra biológica. En una realización, el polímero/oligo de captura contiene una secuencia complementaria que es una secuencia de cebador diseñada para participar en la amplificación de la secuencia de oligonucleótidos de la construcción polimérica/secuencia de oligonucleótidos de la construcción. En otra realización, la secuencia de captura está inmovilizada sobre un sustrato. Similarmente a la secuencia de anclaje, cada secuencia de captura puede tener de al menos aproximadamente 3 a aproximadamente 100 unidades monoméricas, por ejemplo, nucleótidos, de longitud. Por lo tanto, en diversas realizaciones, la secuencia de captura o su secuencia complementaria está formada por una secuencia de al menos 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, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99 o hasta 100 unidades monoméricas, por ejemplo, ácidos nucleicos. En una realización, y como se muestra en los ejemplos, un oligo de captura contiene una secuencia de poli T de secuencia complementaria cuando la secuencia de anclaje es una secuencia de poli A. En otra realización, el oligo de captura contiene una secuencia de poli A. En otra realización más, la secuencia complementaria puede ser un polímero aleatorio, por ejemplo, una secuencia de oligonucleótidos, siempre que pueda hibridarse con su secuencia de anclaje prevista.

Las expresiones “*hashtaging* de células”, “*hashing* celular” o “*hashtaging*”, que se utilizan como sinónimos en la presente memoria, se refieren a un método novedoso para etiquetar una muestra, el contenido de una muestra, o dianas que se derivan de la misma muestra o fuente (por ejemplo, una muestra obtenida del mismo sujeto, del mismo paciente, del mismo lote, del mismo ciclo, etc.) donde la muestra debe mezclarse posteriormente con una pluralidad de muestras diferentes para la multiplexación. En algunas realizaciones, un método de *hashtaging* comprende poner en contacto las dianas de una muestra con una o más construcciones que comprenden un código de barras único que identifica la muestra. Cuando dos o más construcciones se usan en un método de *hashtaging* para etiquetar la misma muestra, a veces todas las construcciones comprenden el mismo código de barras. En algunas realizaciones, donde dos o más construcciones se usan en un método de *hashtaging* para etiquetar la misma muestra, todas las construcciones comprenden la misma sonda de amplificación o sondas de amplificación que son sustancialmente idénticas. Las dianas utilizadas para *hashtag* múltiples muestras pueden ser dianas iguales o diferentes. En algunas realizaciones, una primera muestra se etiqueta con una primera construcción, una segunda muestra se etiqueta con una segunda construcción y una tercera muestra se etiqueta con una tercera construcción, donde cada una de las construcciones está configurada para unirse específicamente a la misma diana, sin embargo, cada una de las construcciones primera, segunda y tercera comprende un código de barras distinguible que es sustancialmente diferente. Después de eliminar por lavado las construcciones no unidas, la muestra se puede combinar para análisis adicional usando un método descrito en la presente memoria. El *hashtaging* permite la detección, seguimiento y/o cuantificación posteriores de cada una de las muestras y dianas que se derivan de la misma muestra.

En algunas realizaciones, una primera construcción como se ha descrito anteriormente se usa para marcar todas las células de una muestra antes de agrupar múltiples muestras de células y antes de realizar otros métodos de secuenciación de scRNA o de CITE-seq usando otras de dichas construcciones que tienen diferentes secuencias de sonda de amplificación. Tras la retrotranscripción, las partes del oligonucleótido de las construcciones de *hashtag* celular se convierten en “*hashtags*” que permiten la identificación y asignación de cada célula comprendida en una mezcla heterogénea a su población original respectiva. Por lo tanto, la construcción célula-*hashtag* cumple con el objetivo de identificar todas las células de una muestra en particular. El ligando en la construcción célula-*hashtag* puede ser un grupo de anticuerpos contra proteínas ampliamente expresadas o un anticuerpo único contra dicha proteína o cualquier otro ligando de unión a células. Puesto que la secuencia de la sonda de amplificación del *hashtag*

celular es diferente de la que tiene la construcción primera o adicional usada en los métodos de CITE-seq, las células individuales de una muestra identificada se pueden rastrear a través de los métodos de CITE-seq, que se usan típicamente para identificar células de una muestra que expresan proteínas de superficie celular específicas de forma diferenciada.

Por el término “inmovilizado” se entiende que la secuencia de polímero/oligonucleótido de captura está unida a un sustrato sólido dando como resultado una reducción o pérdida de movilidad mediante adsorción física a través de una interacción entre cargas o bien una interacción hidrófoba, enlace covalente, interacción de estreptavidina-biotina o acoplamiento por afinidad. Véanse, por ejemplo, las referencias 28 y 29.

Por el término “sustrato” se entiende una micropartícula (perla), una micropartícula microfluídica (perla), un portaobjetos, una placa de múltiples pocillos o un chip. Los sustratos son convencionales y pueden ser vidrio, plástico o cualquier material convencional adecuado para los protocolos de ensayo o diagnóstico particulares. Véanse, por ejemplo, las referencias 1 y 31.

Los términos “un” o “una” se refieren a uno/a o más. Por ejemplo, se entiende que “un casete de expresión” representa uno o más de dichos casetes. De esta forma, los términos “un” (o “una”), “uno/a o más” y “al menos uno/a” pueden utilizarse indistintamente en la presente memoria.

Como se utiliza en la presente memoria, el término “aproximadamente” significa una variabilidad de más o menos 10 % de la referencia dada, salvo que se indique lo contrario.

Las palabras “comprenden”, “comprende” y “que comprende” deben interpretarse de manera inclusiva en lugar de exclusiva, es decir, para incluir otros componentes no especificados o pasos de proceso.

Las palabras “consiste”, “que consiste” y sus variantes, deben interpretarse de manera exclusiva, en lugar de inclusiva, es decir, para excluir componentes o pasos que no se mencionan específicamente.

Como se utiliza en la presente memoria, la expresión “que consiste esencialmente en” limita el alcance de una composición o método descrito a los materiales o pasos especificados y aquellos que no afectan materialmente las características básicas y novedosas del método o composición descrito o reivindicado. Siempre que en esta memoria descriptiva se describe un método o composición como “que comprende” ciertos pasos o características, también pretende abarcar el mismo método o composición que consiste esencialmente en esos pasos o características y que consiste en esas pasos o características.

Por simplicidad y facilidad de comprensión, a lo largo de esta memoria descriptiva, se proporcionan ciertos ejemplos específicos para enseñar la construcción, el uso y el funcionamiento de los diversos elementos de las composiciones y métodos descritos en la presente memoria. Dichos ejemplos específicos no pretenden limitar el alcance de esta descripción.

II. Las composiciones

Las composiciones usadas en los métodos descritos en la presente memoria comprenden una o más de las construcciones, primeras construcciones y construcciones adicionales y una variedad de selección de componentes de construcción como se ha descrito anteriormente. La selección de los componentes de la composición dependerá de la identidad de la diana buscada, los protocolos de secuenciación y amplificación de ARN empleados y el propósito del método de ensayo. En la sección de métodos, más adelante, los métodos ilustrados emplean metodologías Drop-seq; sin embargo, se pueden usar otros métodos. El método usado puede determinar la selección y las composiciones de los diversos componentes descritos anteriormente que forman la composición. Por lo tanto, la siguiente descripción de las composiciones no es exhaustiva, y un experto en la técnica puede diseñar muchas composiciones diferentes basadas en las enseñanzas proporcionadas en la presente memoria. La composición también puede contener las construcciones en un portador o excipiente adecuado. Los elementos de cada composición dependerán del formato de ensayo en el que se empleará.

En una realización, una composición comprende una “primera” construcción que comprende un “primer” ligando unido o conjugado a una construcción polimérica, por ejemplo, una secuencia de oligonucleótidos de la construcción, mediante un enlazador. En esta realización, la secuencia de oligonucleótidos de la construcción comprende a) una sonda de amplificación; b) un código de barras que identifica específicamente el primer ligando; c) un identificador molecular único opcional que se coloca adyacente al código de barras en su extremo 5' o 3'; y d) un anclaje (por ejemplo, de al menos 3 nucleótidos) para hibridarse a una secuencia del oligonucleótido de captura que comprende una secuencia complementaria del anclaje. En una realización, el primer ligando se une específicamente a una primera diana ubicada en o sobre la superficie de una célula, tal como un antígeno o epítipo de la superficie celular.

En otra realización, una composición comprende múltiples “primeras” construcciones sustancialmente idénticas, en donde cada primera construcción sustancialmente idéntica difiere de la “primera” construcción de referencia solo en la secuencia del identificador molecular único opcional o su ausencia de la construcción. Otra realización más de la

composición incluye al menos una construcción adicional, que comprende un ligando adicional unido o conjugado a una secuencia de oligonucleótidos de la construcción adicional mediante un enlazador, uniéndose específicamente el ligando adicional a una diana adicional ubicada en o sobre la superficie de una célula, y la secuencia de oligonucleótidos de la construcción adicional que comprende: a) una sonda de amplificación; b) un código de barras adicional que identifica específicamente el ligando adicional; c) un identificador molecular único adicional opcional que se coloca adyacente al código de barras adicional en su extremo 5' o 3', y d) un anclaje de al menos 3 nucleótidos para hibridarse a una secuencia complementaria. En una realización, la sonda de amplificación o el anclaje también difieren de los componentes correspondientes en cualquier otra construcción en la composición. Los componentes específicamente identificados como componentes "adicionales" difieren de los componentes correspondientes en cualquier otra construcción en la composición. En otra realización, una composición comprende múltiples construcciones "adicionales" sustancialmente idénticas, en donde cada construcción adicional sustancialmente idéntica difiere de la construcción "adicional" de referencia solo en la secuencia del identificador molecular único opcional o su ausencia de la construcción. El número de construcciones en una sola composición está limitado solo por el número de dianas que se desean identificar y/o cuantificar.

Como se describe en los ejemplos siguientes, en una composición específica, el ligando primero o adicional es un anticuerpo o fragmento de anticuerpo y la diana primera o adicional es un epítipo de superficie celular. En otra composición específica, el ligando primero o adicional es un anticuerpo o fragmento de anticuerpo y la primera o diana adicional es una proteína intracelular. Se puede preparar cualquier número de composiciones con diversas combinaciones de ligandos y dianas como se ha descrito anteriormente. Por ejemplo, una construcción de hashtag celular preferiblemente utiliza un ligando que se dirige a una proteína celular ampliamente expresada, basándose en las diferencias en el uso previsto de estas construcciones a diferencia de las construcciones de CITE-seq, como se describe en la presente memoria.

En otra composición, la primera construcción comprende un primer anticuerpo o fragmento del mismo unido o conjugado a una secuencia de oligonucleótidos de la construcción mediante un enlazador, uniéndose específicamente el primer anticuerpo o fragmento del mismo a una primera secuencia de epítipo ubicada en la superficie de una célula, y comprendiendo la secuencia de oligonucleótidos de la construcción: una sonda de amplificación; un código de barras que identifica específicamente el primer anticuerpo o fragmento; un identificador molecular único opcional que se coloca adyacente al código de barras en su extremo 5' o 3'; y una secuencia de anclaje de poli A de al menos 3 nucleótidos para hibridarse a una secuencia de poli T. Este tipo de composición es particularmente adecuado en donde la secuencia de poli T complementaria se inmoviliza sobre un sustrato, por ejemplo, un perla de microfluidos. Como se describe en los ejemplos siguientes, la construcción de esta composición contiene un enlazador que comprende biotina, que está unida al extremo 5' de la secuencia de oligonucleótidos de la construcción por un enlace disulfuro; y estreptavidina, que se fusiona con el anticuerpo o fragmento de anticuerpo. Puede diseñarse otra composición que contiene múltiples de estas primeras construcciones, que difieren solo en la secuencia del identificador molecular único opcional o su ausencia de la construcción.

En otra realización más, la composición contiene al menos una construcción adicional, que comprende al menos un anticuerpo adicional o fragmento del mismo que se une específicamente a un epítipo adicional ubicado en o sobre la superficie de una célula. El anticuerpo o fragmento adicional se conjuga con una secuencia de oligonucleótidos de la construcción adicional mediante un enlazador, en donde la secuencia de oligonucleótidos de la construcción adicional comprende, de 5' a 3': una sonda de amplificación; una secuencia de código de barras adicional que identifica específicamente el anticuerpo o fragmento adicional de cualquier otro anticuerpo o fragmento que reconoce un epítipo adicional, una secuencia de identificador molecular único adicional opcional que se coloca adyacente al código de barras en su extremo 5' o 3', y una secuencia de poli A de al menos 3 nucleótidos diseñada para hibridarse a una secuencia de poli T, en donde los componentes adicionales difieren de los componentes correspondientes en cualquier otra construcción. En otra realización, la sonda de amplificación o el anclaje difieren de los componentes correspondientes en cualquier otra construcción en la composición.

Otra composición específica ilustrativa contiene un mimético de anticuerpo como el primer ligando y la primera diana es una proteína expresada intracelularmente que está presente en una muestra biológica de tejido de biopsia. La primera construcción comprende el mimético de anticuerpo diseñado para unirse a la proteína diana unida covalentemente a una secuencia de oligonucleótidos de la construcción mediante un enlazador, por ejemplo, un enlazador de disulfuro. La secuencia de oligonucleótidos de la construcción comprende en orden 5' a 3': una sonda de amplificación; un código de barras que identifica específicamente el primer mimético de anticuerpo; un UMI se coloca adyacente al código de barras en su extremo 5'; y una secuencia de anclaje de poli A. La composición también contiene una o más primeras construcciones sustancialmente idénticas, donde cada primera construcción sustancialmente idéntica difiere de la "primera construcción" de referencia que contiene una secuencia diferente del UMI. En una realización, una construcción sustancialmente idéntica no contiene el UMI.

En otra realización ilustrativa más, la composición contiene dos construcciones adicionales. Cada construcción adicional comprende un mimético de anticuerpo diferente que se une específicamente a una proteína diferente presente en la muestra de tejido de biopsia. Cada una de las dos construcciones adicionales comprende el mimético de anticuerpo conjugado con su secuencia de oligonucleótido de la construcción adicional mediante un enlazador. Cada enlazador puede tener una química opcional como se ha enseñado anteriormente. En una de dichas

construcciones adicionales, la secuencia de oligonucleótidos de la construcción comprende de 3' a 5': una sonda de amplificación; una secuencia de código de barras que identifica específicamente el mimético de anticuerpo adicional de cualquier otro anticuerpo o fragmento que reconoce una diana de proteína diferente de las primeras construcciones, y una secuencia de UMI diferente adicional que se coloca adyacente al código de barras en su extremo 3', y una secuencia de poli A de al menos 5 nucleótidos diseñada para hibridarse a una secuencia de poli T. En otra realización, la segunda construcción adicional comprende de 5' a 3': una sonda de amplificación; una secuencia de código de barras que identifica específicamente un mimético de anticuerpo diferente del de las primeras construcciones y de la primera construcción adicional, y que reconoce una tercera proteína diana diferente de la primera construcción o primera construcción adicional. Esta segunda construcción adicional no contiene el UMI pero contiene una secuencia de poli A de al menos 3 nucleótidos diseñados para hibridarse a una secuencia de poli T. Estas dos construcciones adicionales tienen dianas, ligandos miméticos de anticuerpos, códigos de barras y UMI (si están presentes) que difieren de cada uno de los componentes correspondientes del otro y difieren de los componentes correspondientes en la "primera" construcción y cualesquiera "primeras" construcciones sustancialmente idénticas presentes en la composición. Debe entenderse además que las composiciones también pueden tener una o más construcciones adicionales sustancialmente idénticas, que difieren de la construcción adicional de referencia por el UMI, como se ha descrito anteriormente.

Muchos otros tipos de ligandos, dianas, muestras, UMI y códigos de barras como se ha descrito anteriormente se pueden usar para generar una amplia variedad de composiciones como se describe en la presente memoria.

También se proporcionan kits que contienen las composiciones. Dichos kits contendrán una o más construcciones primera o adicionales, uno o más conservantes, estabilizantes o tampones, y dichos reactivos de ensayo y amplificación adecuados dependiendo de los métodos de amplificación y análisis y protocolos en los que se usará la composición. Otros componentes adicionales de un kit incluyen reactivos opcionales para la escisión del enlazador, un tampón de lavado, una solución de bloqueo, un tampón de lisis y una solución de encapsulación, marcadores detectables, sustratos de inmovilización, sustratos opcionales para marcadores enzimáticos, así como otros artículos de laboratorio.

III. Métodos de uso de las composiciones

Las composiciones y kits descritos anteriormente se pueden usar en diversos entornos para la detección de diferentes dianas, empleando cualquier número de ensayos y métodos para la detección de dianas en general.

En una realización, un método para detectar una o más dianas en una muestra biológica usa las composiciones descritas en la presente memoria. El método incluye los pasos de poner en contacto la muestra biológica con una o más de las composiciones descritas anteriormente. En una realización, la muestra se pone en contacto con una composición que comprende una primera construcción que tiene un primer ligando unido o conjugado a una construcción polimérica, por ejemplo, una secuencia de oligonucleótidos de la construcción, mediante un enlazador. En una realización, el primer ligando se une específicamente a una primera diana ubicada en una célula o sobre la superficie de una célula, tal como un epítipo de la superficie celular. La secuencia de oligonucleótidos de la construcción comprende: una sonda de amplificación; un código de barras que identifica específicamente el primer ligando; un identificador molecular único opcional que se coloca adyacente al código de barras en su extremo 5' o 3'; y un anclaje para hibridarse a una secuencia complementaria para generar una secuencia de oligonucleótido bicatenaria. En otra realización, una muestra biológica se pone en contacto con una composición que comprende "primeras" construcciones sustancialmente idénticas, en donde cada primera construcción sustancialmente idéntica difiere de la "primera" construcción de referencia solo en la secuencia del UMI opcional o su ausencia de la construcción. Por lo tanto, la muestra biológica se pone en contacto con múltiples ligandos del mismo epítipo diana de la superficie celular.

En otra realización más del método, la muestra se pone en contacto con una primera construcción como se ha descrito anteriormente (o múltiplos de la misma); y una composición que comprende al menos una construcción adicional. El ligando adicional está unido o conjugado covalentemente con una secuencia de oligonucleótidos de la construcción adicional mediante un enlazador, uniéndose específicamente el ligando adicional a una diana adicional ubicada en una célula o en la superficie de una célula. Por lo tanto, la diana adicional es, en una realización, un epítipo de la superficie celular diferente. La secuencia de oligonucleótidos de la construcción adicional comprende: una sonda de amplificación; un código de barras adicional que identifica específicamente el ligando adicional; un identificador molecular único adicional opcional que se coloca adyacente al código de barras adicional en su extremo 5' o 3', y un anclaje de al menos 3 nucleótidos para hibridarse a una secuencia complementaria para generar una secuencia de oligonucleótido bicatenaria, en donde los componentes adicionales difieren de los componentes correspondientes en cualquier otra construcción en la composición. En otra realización más, la sonda de amplificación o el anclaje difieren de los componentes correspondientes en cualquier otra construcción en la composición. Debe entenderse que, en esta realización, puede diseñarse cualquier número de construcciones adicionales como se ha descrito anteriormente para unirse a muchos epítopos celulares según se desee, limitado solo por la elección y número de ligandos. Como se describe en la presente memoria, en otra realización, la composición puede contener una o más construcciones "adicionales" sustancialmente idénticas, en donde cada construcción adicional sustancialmente idéntica difiere de la construcción "adicional" de referencia solo en la secuencia del UMI opcional o su ausencia de la construcción.

En dicho método, después de la puesta en contacto y la unión que se produce entre las células de la muestra biológica y el primer ligando de la primera construcción y el número deseado de ligandos adicionales en la una o más construcciones adicionales, la muestra biológica se lava para eliminar las construcciones no unidas, de haberlas. Para cada construcción unida a su epítipo diana, la secuencia de anclaje se hibrida a continuación con su secuencia de oligo de captura complementaria correspondiente. Esto puede ocurrir mediante la adición de cebadores como secuencias de captura complementarias o una secuencia de oligo de captura complementaria inmovilizada sobre un sustrato, tal como una perla, un portaobjetos, una placa de múltiples pocillos o un chip. En ciertas realizaciones, el extremo 5' de la secuencia complementaria comprende además: una sonda de amplificación adicional; un código de barras adicional que identifica específicamente el sustrato al que está unido la secuencia del oligo de captura; y un identificador molecular único adicional opcional que se coloca adyacente al código de barras adicional en su extremo 5' o 3' que identifica cada secuencia de oligo de captura.

Una vez que el oligo de captura con su secuencia complementaria está presente en la muestra, puede producirse la generación de secuencias oligonucleotídicas bicatenarias. En ciertas realizaciones, el método también incluye insertar opcionalmente uno o más UMI en una posición adyacente al código de barras en su extremo 5' o 3' o en cualquier otra parte, siempre que la inserción no impida las funciones de los componentes de la secuencia de oligonucleótidos de la construcción antes o después de la hibridación al anclaje.

El método de detección, en una realización, incluye detectar las secuencias del código de barras de la construcción de cada una de las construcciones primera y adicional para identificar si la muestra biológica expresa o contiene la primera diana (por ejemplo, epítipo), las dianas adicionales (por ejemplo, uno o múltiples epítopos de superficie celular adicionales) o una combinación de la primera diana y dianas adicionales (por ejemplo, múltiples epítopos diferentes).

En otra realización más de este método de detección, el nivel de expresión de la primera diana o dianas adicionales en la muestra biológica se determina detectando la cantidad de los correspondientes códigos de barras de la construcción. En una realización, la detección se realiza mediante normalización a la cantidad de uno cualquiera de los identificadores moleculares únicos o la cantidad media de dos o más de los identificadores moleculares únicos.

Diversas realizaciones de los métodos pueden incluir añadir a la muestra biológica la composición que contiene la primera construcción o construcciones solamente, o bien composiciones que contienen la construcción o construcciones adicionales simultánea o secuencialmente antes del paso de lavado. Otros pasos del método pueden incluir aislar la muestra biológica en células individuales o poblaciones de células antes del paso de puesta en contacto o después del paso de lavado. Otro paso implica extender el oligonucleótido de captura hibridado a la secuencia de anclaje para copiar un código de barras de la construcción, una sonda UMI y de amplificación en secuencias bicatenarias. Las secuencias oligonucleotídicas bicatenarias también se pueden generar después de la hibridación del anclaje con cebadores hibridados con las sondas de amplificación después de la hibridación de anclaje y/o la inserción de los UMI.

En algunas realizaciones, un oligonucleótido comprende uno o más códigos de barras. Puede usarse cualquier código de barras adecuado para una composición o método descrito en la presente memoria. Un código de barras a menudo comprende o consiste en una secuencia de ácido nucleico relativamente corta de, por ejemplo, 2 a 50, 2 a 30, 2 a 20, 2 a 15, 10-20, 4 a 15 o 2 a 5 nucleótidos consecutivos, donde la secuencia de nucleótidos de un código de barras es única para un ácido nucleico, un oligonucleótido, una población de oligonucleótidos (por ejemplo, una pluralidad de oligonucleótidos sustancialmente idénticos), una población de construcciones sustancialmente idénticas, una muestra, una fuente de muestra, un ligando, un lote, un ciclo o una combinación de los mismos. Por ejemplo, un código de barras que es único para un oligonucleótido unido a un primer ligando, se puede usar para identificar la presencia o la cantidad del primer ligando después de detectar la secuencia del código de barras (por ejemplo, mediante secuenciación u otro método adecuado). En consecuencia, se puede usar un código de barras que es único para un oligonucleótido unido a un primer ligando para identificar específicamente la presencia, cantidad o ausencia de un ligando (por ejemplo, un primer ligando, un segundo ligando o un ligando adicional) en una muestra multiplexada que comprende otros ligandos, otros ácidos nucleicos, otros oligonucleótidos y otros códigos de barras. Dicho código de barras puede denominarse código de barras "específico de ligando", o un código de barras que identifica específicamente un ligando (por ejemplo, un primer ligando o cualquier ligando específico). Similarmente, un código de barras que es único para un oligonucleótido, muestra, perla, lote o ciclo, se puede usar para identificar específicamente un oligonucleótido, muestra, perla, lote o ciclo particular en una muestra multiplexada que comprende una pluralidad de otros oligonucleótidos, muestras, perlas, lotes o ciclos. Un código de barras que se puede usar para identificar específicamente una muestra puede denominarse código de barras "específico de muestra" o código de barras que identifica específicamente una muestra. Un código de barras que se puede usar para identificar específicamente una perla puede denominarse código de barras "específico de perla" o código de barras que identifica específicamente una perla. Un código de barras que se puede usar para identificar específicamente un oligonucleótido o ácido nucleico puede denominarse código de barras "específico de oligonucleótido" o código de barras "específico de ácido nucleico", respectivamente. En algunas realizaciones, un oligonucleótido comprende un identificador molecular único (UMI). En ciertas realizaciones, un UMI comprende un código de barras único que identifica específicamente un oligonucleótido individual de todos los demás oligonucleótidos usados en una composición o método descrito en la presente memoria.

Los métodos y composiciones descritos en la presente memoria se pueden usar, en algunas realizaciones, para determinar la presencia, cantidad o ausencia de una muestra, diana, construcción u oligonucleótido. En algunas realizaciones, determinar la cantidad de una muestra, diana, construcción u oligonucleótido comprende determinar una cantidad absoluta, aproximada, media, promedio o relativa de una muestra, diana, construcción u oligonucleótido en un ensayo multiplexado. En consecuencia, en ciertas realizaciones, los métodos y composiciones descritos en la presente memoria se pueden usar para cuantificar cantidades de una muestra, diana, construcción u oligonucleótido en un ensayo multiplexado.

En algunas realizaciones, un oligonucleótido comprende una sonda de amplificación. Puede usarse cualquier sonda de amplificación adecuada para una composición o método descrito en la presente memoria. En algunas realizaciones, una sonda de amplificación comprende una longitud relativamente corta de aminoácidos consecutivos que se integran en un oligonucleótido o ácido nucleico descrito en la presente memoria. Un sonda de amplificación puede tener cualquier longitud adecuada. En algunas realizaciones, una sonda de amplificación tiene de 5 a 50, de 5 a 40, de 5 a 35, de 5 a 25 o de 5 a 15 nucleótidos de longitud. En ciertas realizaciones, se usa una sonda de amplificación para la captura y/o amplificación y/o secuenciación de un ácido nucleico. Un sonda de amplificación puede comprender cualquier secuencia de ácido nucleico adecuada para la unión al cebador, la captura, la extensión con una polimerasa y/o la amplificación con una polimerasa. En algunas realizaciones, una sonda de amplificación comprende un sitio de unión al cebador. En algunas realizaciones, un cebador comprende una secuencia de ácido nucleico sustancialmente idéntica a una sonda de amplificación. En ciertas realizaciones, un oligonucleótido comprende un ácido nucleico interpuesto flanqueado por una sonda de amplificación en 5' y 3', o complemento del mismo, en donde las sondas de amplificación flanqueantes facilitan la amplificación del ácido nucleico interpuesto.

Otra variación de este método implica escindir el ligando de la construcción antes o después de la hibridación del anclaje a una secuencia complementaria. Otra realización adicional implica lisar la célula, cuando se desee. En varias realizaciones, la técnica de lisis puede implicar la exposición de las células a detergentes, soluciones tampón de detergente, tales como tampón RIPA, tampones de lisis IP, soluciones de reactivo M-PER o B-PER (Pierce Chemical) y similares. Las construcciones de ligando-oligonucleótido se pueden usar con dianas que no sean anticuerpos de la superficie celular y ligandos que no sean anticuerpos, como se describe en la presente memoria.

Una realización adicional implica la permeabilización celular y un procedimiento de fijación opcional antes del paso de puesta en contacto o entre los pasos de puesta en contacto secuenciales con las construcciones primeras o adicionales. En varias realizaciones, la técnica de permeabilización puede implicar la exposición de las muestras biológicas a disolventes orgánicos (por ejemplo, aunque no de forma limitativa, a metanol y acetona), detergentes (tales como Saponin™, Triton X-100™ y Tween-20™), otro reactivo disponible para el experto en la técnica (tal como solución de sal de cinc³², conjunto de tampones de fijación intracelular y permeabilización eBioscience™; y kit de fijación celular y permeabilización celular FIX & PERM®) y cualquier combinación de los mismos. El paso de fijación es opcional antes o durante la permeabilización. Los expertos en la técnica conocen técnicas de fijación, por ejemplo, poner en contacto las muestras biológicas con una solución que contiene fijadores de reticulación (tales como formaldehído, glutaraldehído y otro aldehído), fijadores de precipitación (tales como metanol, etanol, acetona y ácido acético), agentes oxidantes (tales como tetróxido de osmio, dicromato de potasio, ácido crómico y permanganato de potasio), mercuriatos, picratos, fijador de efecto de protección contra disolventes orgánicos (HOPE) mediado por tampón Hepes-ácido glutámico, 2,4,6-trimetilpiridina, conjunto de tampones de fijación intracelular y permeabilización eBioscience™ y kit de fijación celular y permeabilización celular FIX & PERM® o cualquier combinación de los mismos. En otras realizaciones adicionales, el paso adicional de recuperar una cantidad y calidad suficientes de construcciones, ADN o ARN después de llevar a cabo la permeabilización está implicada. Por ejemplo, véase la ref. 33.

Además, estos métodos pueden emplear protocolos de detección, que incluyen sin limitación, protocolos para PCR, inmunoPCR¹⁵ y ensayo de ligadura de proximidad o extensión de proximidad¹⁶, PEA²⁶, RCA²⁵, protocolos de secuenciación e hibridación de fluorescencia.

En otra realización más, el método es un método de alto rendimiento. En una realización, las composiciones descritas en la presente memoria se usan en protocolos de alto rendimiento tales como los siguientes. Un método de alto rendimiento para detectar uno o más epítomos en una muestra biológica puede emplear cientos o miles de pocillos que contienen las mismas o diferentes muestras. El método comprende poner en contacto una muestra biológica con una composición que comprende una primera construcción que comprende un primer anticuerpo o fragmento del mismo que se une específicamente a un primer epítipo, el primer anticuerpo o fragmento unido o conjugado con una secuencia de oligonucleótidos de la construcción mediante un enlazador, en donde la secuencia de oligonucleótidos de la construcción comprende: una sonda de amplificación, una secuencia de código de barras que identifica específicamente el primer anticuerpo o fragmento de cualquier otro anticuerpo o fragmento que reconoce un epítipo diferente, una secuencia de identificador molecular único opcional que se coloca adyacente al código de barras en su extremo 5' o 3', y una secuencia de anclaje (por ejemplo, de al menos 3 nucleótidos) para hibridarse a una secuencia complementaria y generar una secuencia de oligonucleótidos bicatenaria.

En una realización similar, la composición comprende una o más construcciones sustancialmente idénticas, en donde cada primera construcción sustancialmente idéntica difiere solo en la secuencia del identificador molecular único

opcional o su ausencia de una construcción de referencia (por ejemplo, “primera” o “adicional”). En otra realización, la composición comprende al menos una construcción adicional, que comprende un anticuerpo adicional o fragmento del mismo unido o conjugado a una secuencia de oligonucleótidos de la construcción adicional mediante un enlazador. El anticuerpo o fragmento del mismo adicional se une específicamente a un epítipo adicional. La secuencia de oligonucleótidos de la construcción adicional comprende: la misma sonda de amplificación u otra diferente, un código de barras adicional que identifica específicamente el anticuerpo adicional o fragmento del mismo; un identificador molecular único adicional opcional que se coloca adyacente al código de barras adicional en su extremo 5' o 3', y el mismo anclaje o uno diferente, en donde la diana adicional y el anticuerpo adicional o fragmento de ligando adicional, el UMI opcional y componentes de código de barras adicionales difieren de los componentes correspondientes en cualquier otra construcción en la composición.

Los protocolos de alto rendimiento también implican lavar la muestra biológica para eliminar las construcciones no unidas; hibridar la secuencia o secuencias de oligonucleótidos de la construcción a través de sus respectivos anclajes a las secuencias complementarias correspondientes y generar la secuencia o secuencias oligonucleotídicas bicatenarias. Los UMI también pueden insertarse opcionalmente en una posición adyacente al código de barras en su extremo 5' o 3', antes o después de la hibridación del anclaje.

Después de eso, dichos métodos implican detectar la una o más secuencias del código de barras de la construcción para identificar si la muestra biológica (o muestras presentes en pocillos individuales) expresa o contiene la primera diana, las dianas adicionales o una combinación de la primera diana y dianas adicionales. Alternativamente, el nivel de expresión de la primera diana o de las dianas adicionales en la muestra biológica se produce detectando la cantidad de los correspondientes códigos de barras. En una realización, la dicha detección se realiza normalizando con respecto a la cantidad de un identificador molecular único o la cantidad media de dos o más identificadores moleculares únicos.

Los métodos de alto rendimiento también pueden incluir añadir una o más construcciones primera y adicionales a una muestra biológica simultánea o secuencialmente antes del paso de lavado. Los métodos también pueden incluir aislar la una o más muestras biológicas unidas a una o más de las construcciones primera o adicionales en células individuales o poblaciones de células después del lavado; y amplificar la secuencia de oligonucleótidos bicatenaria con cebadores hibridados a las sondas de amplificación. Cualquiera del resto de parámetros de las composiciones se puede incluir si se coordinan con los protocolos de ensayo utilizados en la detección.

Otra realización específica adicional del uso de las composiciones descritas en la presente memoria para un método de detección de dianas se analiza en los ejemplos siguientes. Las composiciones descritas en la presente memoria se diseñan y usan para superar las limitaciones de los métodos actualmente existentes para detectar y/o medir transcritos de ARN y proteínas en células individuales (es decir, tecnología de gotículas). El método denominado Indexación celular de transcriptoma y epítomos mediante secuenciación (*Cellular Indexing of Transcriptome and Epitopes by sequencing*, CITE-seq) usa las composiciones descritas en la presente memoria para caracterizar simultáneamente el transcriptoma y un número potencialmente ilimitado de marcadores de superficie celular de la misma célula con alto rendimiento. Combina un perfil de expresión de todo el genoma no sesgado con la medición de marcadores de proteínas específicos en miles de células individuales mediante el uso de microfluídica de gotículas. Las composiciones tienen utilidad, además de añadir una dimensión adicional a los datos del transcriptoma de células individuales. Este método proporciona una caracterización más detallada de las poblaciones celulares, pero también permite el estudio de la regulación génica posterior a la transcripción (y posterior a la traducción) en células individuales a una profundidad sin precedentes.

Como se describe en detalle a continuación, se empleó una suspensión de células mixtas humanas y de ratón y el protocolo Drop-seq con construcciones que comprenden anticuerpos monoclonales como ligandos de construcción unidos a los oligonucleótidos de la construcción que contenían las secuencias identificadoras de anticuerpos únicas (códigos de barras). La suspensión celular se marca con las construcciones de secuencia de ligando-oligonucleótido (en estos casos anticuerpos oligoetiquetados) y las células individuales se encapsulan posteriormente en gotículas acuosas de tamaño de nanolitro en un aparato microfluídico. En cada gotícula, las moléculas de anticuerpo y ADNc se indexan con el mismo código de barras único y se convierten en genotecas que se amplifican independientemente y se mezclan en proporciones apropiadas para secuenciación en el mismo carril. Como se recoge en los ejemplos siguientes, los inventores pudieron identificar inequívocamente las células humanas y de ratón basándose en sus proteínas de superficie celular específicas de especie e independientemente en su transcriptoma.

Los procesos celulares y patologías pueden entenderse a partir del perfilado transcriptómico y proteómico de células individuales con alto contenido de información llevando a cabo una CITE-seq en minigotas en diversos entornos de laboratorio. En una realización, este método CITE-seq es útil para caracterizar el sistema hematopoyético. CITE-seq permite la caracterización en profundidad de células individuales mediante la medición simultánea de niveles de expresión génica y proteínas de superficie celular, es muy escalable, solamente está limitada por el número de anticuerpos específicos que están disponibles y es compatible con otros sistemas de secuenciación de células individuales. Entre dichas plataformas conocidas de secuenciación de células individuales adecuadas para su integración con las composiciones y métodos descritos en la presente memoria está el método de Drop-Seq, que incluye, aunque no de forma limitativa, microfluidos, placas o micropocillos, el método Seq-Well^{TM35} y adaptaciones del protocolo básico, así como el método InDrop^{TM2} (1 Cell Bio). En otra realización, una plataforma de secuenciación

de células individuales adecuada para su integración con las composiciones y métodos descritos en la presente memoria es una solución 10x genómica de células individuales 3' (www.10xgenomics.com/single-cell/)³ o solución V(D)J de células individuales (www.10xgenomics.com/vdj/), analizada bien en un controlador Chromium o en un controlador para células individuales dedicado Chromium). Otros protocolos de secuenciación útiles para su combinación con CITE-seq como se describe en la presente memoria incluyen el método Wafergen iCell8^{TM3,38-40} (www.wafergen.com/products/icell8-single-cell-system); el método micropocillo-sec⁴¹, el método Fluidigm C1^{TM42-44} y productos equivalentes para células individuales. Otros protocolos de secuenciación conocidos útiles con las composiciones y métodos descritos en la presente memoria incluyen la plataforma de análisis de células individuales BD Resolve^{TM37} (derivada de Cyto-seq) y ddSeq⁶ (de Illumina® Bio-Rad® SureCélulaTM Kit WTA 3'Library Prep para el sistema ddSEQTM, 2017, Pub. n.º 1070-2016-014-B, Illumina Inc., Bio-Rad Laboratories, Inc.). En otra realización más, las composiciones y métodos descritos en la presente memoria son útiles con enfoques combinatorios basados en indexación (el método sci-RNA-seq^{TM20} o el método SPLiT-seq^{TM30}) y transcriptómica espacial, o bien enfoques de secuenciación resueltos espacialmente comparables³⁶. Los métodos y composiciones descritos en la presente memoria también se pueden usar como una capa de información adicional relativa a la clasificación de índice estándar (FACS) y los enfoques basados en secuencias de ARNm. En una realización, por ejemplo, los paneles de FACS estándar se complementan con otros anticuerpos etiquetados para CITE-seq detectables mediante secuenciación basada en placas. Otros protocolos de secuenciación adicionales se pueden combinar con las composiciones y métodos específicamente descritos en la presente memoria.

Se puede usar cualquier método de secuenciación de ácido nucleico adecuado para secuenciar los ácidos nucleicos descritos en la presente memoria y/o para detectar la presencia, ausencia o cantidad de los diversos ácidos nucleicos, construcciones, dianas, oligonucleótidos, productos de amplificación y códigos de barras descritos en la presente memoria.

Por lo tanto, un método de alto rendimiento para caracterizar una célula mediante la detección simultánea de uno o más epítomos ubicados en o sobre la célula y el transcriptoma implica poner en contacto una muestra biológica que contiene células con una o más de la composición como se ha descrito anteriormente. En una realización, una composición que comprende un primer anticuerpo o fragmento del mismo que se une específicamente a un primer epítopo ubicado en o sobre la superficie de una célula, conjugado el primer anticuerpo o fragmento a una secuencia de oligonucleótidos de la construcción mediante un enlazador, en donde la secuencia de oligonucleótidos de la construcción comprende: una sonda de amplificación; una secuencia de código de barras que identifica específicamente el primer anticuerpo o fragmento de cualquier otro anticuerpo o fragmento que reconoce un epítopo diferente, una secuencia de identificador molecular único opcional que se coloca adyacente al código de barras en su extremo 5' o 3', y una secuencia de poli A de al menos 3 nucleótidos diseñada para hibridarse a una secuencia de poli T inmovilizada en un perla de microfluidos. En otra realización, la composición comprende una o más "primeras" construcciones sustancialmente idénticas, en donde cada primera construcción sustancialmente idéntica difiere solo en la secuencia del identificador molecular único opcional o su ausencia de la "primera" construcción de referencia.

En otra realización más, la composición comprende además al menos una construcción adicional, que comprende un anticuerpo adicional o fragmento del mismo conjugado con una secuencia de oligonucleótidos de la construcción adicional mediante un enlazador, uniéndose el anticuerpo adicional o fragmento del mismo específicamente a un epítopo adicional, y comprendiendo la secuencia de oligonucleótidos de la construcción adicional de 5' a 3': la sonda de amplificación; un código de barras adicional que identifica específicamente el anticuerpo adicional o fragmento del mismo, un identificador molecular único adicional opcional que se coloca adyacente al código de barras adicional en su extremo 5' o 3' y el anclaje, en donde los componentes adicionales difieren de los componentes correspondientes en cualquier otra construcción en la composición. Las composiciones se pueden añadir a la muestra biológica simultánea o secuencialmente antes de un paso de lavado. En otra realización, la composición comprende una o más construcciones sustancialmente idénticas "adicionales", en donde cada construcción adicional sustancialmente idéntica difiere solo en la secuencia del identificador molecular único opcional o su ausencia de la construcción de referencia "adicional".

En dicho método, una sola célula individual unida a una o más construcciones se encapsula en una gotícula acuosa con una perla, en donde cada perla está conjugada a una construcción que comprende una secuencia única de código de barras de la célula que comprende una secuencia de poli T en 3'. La célula individual de cada gotícula se lisa, en donde los ARNm de la célula y del oligonucleótido de la construcción procedentes del anticuerpo o fragmento se hibridan con las secuencias de poli T de la perla. A partir de las secuencias hibridadas con la perla, se generan ADNc bicatenarios que contienen la secuencia de código de barras de la célula y los transcritos inversos del ARNm celular y del ADN bicatenario que contiene la secuencia de código de barras de la célula y la una o más secuencias de oligonucleótidos de la construcción. Este método también puede incluir un paso de insertar opcionalmente uno o más identificadores moleculares únicos en una posición adyacente al código de barras adicional en su extremo 5' o 3' antes o después del paso de hibridación.

Además, dicho método implica crear mediante amplificación una genoteca que contiene el ADNc del transcriptoma de la célula diana y el ADN que contiene la secuencia o secuencias de oligonucleótidos de la construcción. En una realización, las secuencias de código de barras de la construcción se detectan para identificar si la célula individual expresa el primer epítopo. En otra realización del método, el nivel de expresión del primer epítopo de la célula individual

se determina detectando la cantidad del código de barras de la construcción. En otra realización más del método, la detección se realiza mediante normalización a la cantidad de cualquiera de los identificadores moleculares únicos o la cantidad media de dos o más de los identificadores moleculares únicos. Sustancial y simultáneamente, el transcriptoma de la genoteca está asociado con la célula identificada por la unión e identificación de las construcciones primera y/o adicionales.

Dado el número de variaciones que se pueden generar en las construcciones usando las enseñanzas proporcionadas en la presente memoria, se pueden usar muchos otros métodos que emplean estas composiciones para una identificación rápida y compleja de dianas.

Para ayudar a definir el fenotipo de una célula, es esencial tener una comprensión de la presencia o ausencia de marcadores específicos de las proteínas superficiales y/o de las modificaciones posteriores a la traducción de dichos marcadores. Como se demuestra en los ejemplos siguientes, las composiciones y métodos descritos en la presente memoria proporcionan, en un aspecto, un método basado en secuenciación que combina la detección altamente multiplexada basada en ligando (por ejemplo, basada en anticuerpos) de marcadores proteicos bien establecidos junto con el perfil de transcriptoma no sesgado para miles de células individuales en paralelo. Específicamente, los ejemplos demuestran un método novedoso que puede perfilar muchas dianas, por ejemplo, marcadores celulares y transcriptomas de células individuales en miles de células en paralelo. Estas composiciones y métodos permiten el perfilado de células en el punto de recogida y el perfilado de células individuales a una escala y detalle grandes e inesperados.

En realizaciones adicionales, como paso adicional a cualquiera de los métodos descritos en la presente memoria, en primer lugar se puede realizar un paso de *hashtagging* celular, etiquetando cada célula dentro de una muestra a analizar con la misma “primera construcción” y agrupando a continuación dichas muestras *hashtagged* para su posterior análisis. El análisis adicional incluye el análisis mediante cualquiera de los métodos descritos en la presente memoria. Las partes del oligonucleótido de las construcciones de células *hashtag*, particularmente las secuencias de la sonda de amplificación, son diferentes de las usadas en los métodos analíticos “adicionales”, que permiten el *hashtagging* celular de las muestras sometidas a esos métodos. Este método de “*hashtagging*” realizado antes del agrupamiento de las muestras sometidas a análisis adicionales tiene varias ventajas. La multiplexación permite ahorros de costes y la capacidad de controlar los efectos por lotes, por ejemplo, el proceso tratado/no tratado al mismo tiempo. Las construcciones de células *hashtag* permiten una determinación inequívoca de la mayoría de los dobletes. Finalmente, la combinación de estas dos ventajas permite sobrecargar ampliamente los experimentos de secuenciación de scARN basados en gotículas (es decir, el uso de 20.000 células, en lugar de 4.000 células, por carril), lo que da como resultado un menor coste de los experimentos y una mayor información producida por los experimentos. Esta realización de *hashtagging* se puede usar para multiplexar muestras del mismo genotipo sin la necesidad de realizar el genotipado de las muestras.

En otras realizaciones adicionales, los métodos de *hashtagging* se pueden extender a la asignación de códigos de barras o para identificar núcleos celulares, así como otros componentes celulares.

Los siguientes ejemplos describen el método CITE-seq y un método de *hashtagging* celular como dos meras realizaciones de uso de las composiciones descritas en la presente memoria. Estos ejemplos deben interpretarse en el sentido de que abarcan todas y cada una de las variaciones que se evidencien como resultado de la enseñanza aquí proporcionada.

Ejemplos

Ejemplo 1: Diseño y validación de complejos de anticuerpo-oligo.

Se diseñaron anticuerpo-oligos con las siguientes características: una sonda de amplificación genérica (sonda de PCR) para la preparación de la genoteca de secuenciación avanzada, una secuencia única de código de barras específica de cada anticuerpo y un tramo de poli A en el extremo 3' (Figura 1A). Se generaron dos anticuerpo-oligos. Los anticuerpos dirigidos contra la integrina beta 1 (CD29) de ratón se unieron al código de barras del oligo 1 que contenía un puente disulfuro, una secuencia común (sonda de amplificación, sonda de PCR), un código de barras identificador de anticuerpo único (5'-ATGTCCT-3') y un UMI que contiene 4 nt seguido de una cola de poli A (Figura 2B, panel superior). Los anticuerpos dirigidos contra CD29 humano se unieron al código de barras del oligo 2 que contiene un puente disulfuro, una secuencia común (sonda de amplificación, sonda de PCR), un código de barras identificador de anticuerpo único (5'-GCCATTA-3') y un UMI que contiene 4 nt seguido de una cola de poli A (Figura 2B, panel inferior).

Para los experimentos presentados en los Ejemplos 1 a 7, los oligos se modificaron con biotina y un enlace disulfuro en el extremo 5' del oligo y se unieron a anticuerpos modificados con estreptavidina. El oligo podía liberarse del anticuerpo reduciendo el enlace disulfuro. Específicamente, se adoptó para los Ejemplos 1 a 7 una interacción de estreptavidina-biotina (SAV) usada comúnmente para unir los anticuerpos a los oligonucleótidos¹⁹. Se usó un kit comercial para anticuerpos marcados con estreptavidina (generalmente usado para el posterior marcado de fluoróforo para FACS). Los anticuerpos se unieron a oligos biotinilados (Figura 2A, Carril y Panel n.º 1). La reducción de un

enlace disulfuro en el extremo 5' del oligo liberó estos oligos de los anticuerpos (Figura 2A, Carril y Panel n.º 2). Otras químicas de unión son útiles, que incluyen aunque no de forma limitativa, tiol-maleimida, tiol-haloacetato, amina-NHS, amina-isotiocianato, azida-alquino (CuAAC), tetrazol-cicloocteno (IEDDA, usado en el Ejemplo 7,⁴⁵ (referencias 25, 45 y 46 y referencias en los mismos) y pueden ser enlaces covalentes escindibles o no escindibles.

Ejemplo 2: Métodos y materiales

Conjugación de anticuerpos a oligonucleótidos con códigos de barras de ADN.

Los anticuerpos monoclonales altamente específicos probados por citometría de flujo se conjugaron a oligonucleótidos que contenían secuencias identificadoras de anticuerpos únicas y una cola de poli A.

Se adoptó una interacción estreptavidina-biotina comúnmente usada para unir los oligos a los anticuerpos¹⁹. Los anticuerpos se marcaron con estreptavidina usando el Kit rápido de conjugación de anticuerpos con estreptavidina LYNX (Bio-Rad, EE. UU.) según las instrucciones del fabricante con modificaciones. Específicamente, 15 µg de anticuerpo se marcaron con 10 µg de estreptavidina. En esta relación, se conjugará un promedio de dos tetrámeros de estreptavidina por molécula de anticuerpo, lo que da como resultado un promedio de ocho sitios de unión para la biotina en cada anticuerpo. Se adquirieron oligonucleótidos de ADN con una modificación de amina en 5' a IDT (EE. UU.) y se biotinilaron usando química de NHS según las instrucciones del fabricante (EZ Biotina S-S NHS, Thermo Fisher Scientific, EE. UU.). El enlace disulfuro opcional permite la separación del oligo del anticuerpo con agentes reductores en algunas realizaciones. La separación del oligo del anticuerpo puede no ser necesaria en todas las aplicaciones. El exceso de Biotina-NHS se eliminó mediante filtración en gel (Micro Biospin 6, Bio-Rad) y precipitación con etanol. Los anticuerpos marcados con estreptavidina se incubaron con oligonucleótidos biotinilados en relación equimolar (suponiendo dos tetrámeros de estreptavidina por anticuerpo en promedio) durante la noche a 4 °C en PBS que contenía NaCl 0,5 M y Tween al 0,02 %. El oligo no unido se eliminó de los anticuerpos mediante el uso de filtros de centrifuga con un corte de MW a 50 KDa (Millipore, EE. UU.). La eliminación del exceso de oligo se verificó mediante electroforesis en gel de agarosa al 4 %. Los conjugados de anticuerpo-oligo se almacenaron en PBS complementado con azida de sodio (0,05 %) y BSA (1 µg/µl) a 4 °C.

Lista de anticuerpos usados para CITE-seq.

Los anticuerpos y los clones utilizados fueron CD3e (clon UCHT1, BioLegend, EE. UU.); CD19 (Clon HIB19, BioLegend, EE. UU.); CD4 (Clon RPA-T4, BioLegend, EE. UU.); CD8a (Clon RPA-T8, BioLegend, EE. UU.); CD56 (Clon MEM-188, BioLegend, EE. UU.); CD16 (Clon B73.1, BioLegend, EE. UU.); CD11c (Clon B-ly6, BD Pharmingen, EE. UU.); CCR7 (Clon 150603, R&D Systems, EE. UU.); CCR5 (Clon J418F1, BioLegend, EE. UU.); CD34 (Clon 581, BioLegend, EE. UU.); CD14 (Clon M5E2, BioLegend, EE. UU.); CD10 (Clon HI10a, BioLegend, EE. UU.); CD45RA (Clon HI100, BioLegend, EE. UU.); D29 (Clon MA1-19105, Thermo Fisher, EE. UU.); CD29 (Clon MA5-16707, Thermo Fisher, EE. UU.); CD2 (Clon RPA-2.10, BioLegend, EE. UU.); CD57 (Clon H-NK1, BioLegend, EE. UU.). Consulte la Ref. 46, pestaña suplementaria. 2.

Secuencias de anticuerpo-oligo.

Los inventores aprovecharon la actividad de la ADN polimerasa dependiente de ADN de las transcriptasas inversas comúnmente usadas⁵⁶ para convertir oligonucleótidos de ADN de CITE-seq en ADNc durante la retrotranscripción al mismo tiempo que los ARNm. La actividad de la ADN polimerasa dependiente de ADN de las transcriptasas inversas MMLV está bien establecida. Todos los protocolos SMART (mecanismo de conmutación en el extremo 5' del molde de ARN) de preparación de genoteca (por ejemplo, comercializado por Clontech) se basan en esta actividad. La enzima RT cambia al final del molde de ARN a un oligo de cambio de molde (TSO) para la síntesis de ADNc adicional. Los protocolos ARN-sec de células individuales (que incluyen 10x Genomics y Drop-seq) también dependen completamente de esta actividad para añadir una sonda de PCR al extremo 5' de los ADNc de longitud completa. La sonda de PCR se usa para la amplificación posterior. Dependiendo de la aplicación, la sonda de amplificación por PCR en los oligos de anticuerpo-código de barras debe cambiarse dependiendo de qué lectura de secuencia se usa para la lectura del ARN (por ejemplo, 10x Single Cell 3' v1 utiliza la lectura 1, mientras que Drop-seq y 10x Single Cell 3' v2 utilizan la lectura 2). Los diseños de prueba de concepto del oligonucleótido anticuerpo (humano y de ratón)-código de barras de los inventores incluyeron:

los UMI, que son redundantes para los protocolos Drop-seq y 10x debido a la adición del UMI al ADNc en la retrotranscripción. Los UMI del oligonucleótido conjugado con anticuerpo pueden ser útiles para otras iteraciones del método donde los UMI no forman parte del protocolo de preparación de genotecas scRNA-seq.

Mezcla de especies, Drop-seq (que contiene la sonda Nextera lectura 2).

BC6: /5AmMC12/GTCTCGTGGGCTCGGAGATGTGTATA
AGAGACAGGCCAATNNBAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAA SEQ ID NO: 1

BC12: /5AmMC12/GTCTCGTGGGCTCGGAGATGTGT
ATAAGAGACAGCTTGTANNBAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAA SEQ ID NO: 2

Mezcla de especies, 10x (versión 3' para células individuales, Nextera lectura 1

5

handle). BC6: /5AmMC12/TCGTCGGCAGCGTCAGATGT
GTATAAGAGACAGGCCAATNNBAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAA SEQ ID NO: 3

BC12: /5AmMC12/TCGTCGGCAGCGTCAGATGTGTATA
AGAGACAGCTTGTANNBAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAA SEQ ID NO: 4

10

Perfilado de CBMC: (oligos compatibles con Drop-seq y 10x v2, que contienen la sonda de lectura 2 de ARN pequeño TruSeq). v2_BC1: /5Am

MC12/CCTTGGCACCCGAGAATTCCAATCACGBAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA SEQ ID NO: 5

15

v2_BC2: /5AmMC12/CCTTGGCACCCGAGAATTCCAC
GATGTBAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAA SEQ ID NO: 6

v2_BC3: /5AmMC12/CCTTGGCACCCGAGAATTCCAT
TAGGCBAAAAAAAAAAAAAAAAAAAAAAAAAAAA SEQ ID NO: 7

20

v2_BC4: /5AmMC12/CCTTGGCACCCGAGAATTC
CATGACCABAAAAAAAAAAAAAAAAAAAAAAAA
AAAA SEQ ID NO: 8

v2_BC6: /5AmMC12/CCTTGGCACCCGAGAATTCCAGC
CAATBAAAAAAAAAAAAAAAAAAAAAAAAAAAA SEQ ID NO: 9

v2_BC9: /5AmMC12/CCTTGGCACCCGAGAATTCC
AGATCAGBAAAAAAAAAAAAAAAAAAAAAAAA
AAA SEQ ID NO: 10

25

v2_BC10: /5AmMC12/CCTTGGCACCCGAGAATTCC
ATAGCTTBAAAAAAAAAAAAAAAAAAAAAAAA
AAA SEQ ID NO: 11

v2_BC12: /5AmMC12/CCTTGGCACCCGAGAATT
CCACTTGTABAAAAAAAAAAAAAAAAAAAA
AAAAA SEQ ID NO: 12

30

v2_BC8: /5AmMC12/CCTTGGCACCCGAGAATTCC
 AACTTGABAAAAAAAAAAAAAAAAAAAAAAAAAAAA
 AAA SEQ ID NO: 13

v2_BC11: /5AmMC12/CCTTGGCACCCGAGAATTCC
 AGGCTACBAAAAAAAAAAAAAAAAAAAAAAAAAAAA
 AAA SEQ ID NO: 14

v2_BC13: /5AmMC12/CCTTGGCACCCGAGAATTCC
 AAGTCAABAAAAAAAAAAAAAAAAAAAAAAAAAAAA SEQ ID NO: 15

v2_BC14: /5AmMC12/CCTTGGCACCCGAGAATTCC
 AAGTTCCBAAAAAAAAAAAAAAAAAAAAAAAAAAAA SEQ ID NO: 16

v2_BC5: /5AmMC12/CCTTGGCACCCGAGAATTCC
 AACAGTGBAAAAAAAAAAAAAAAAAAAAAAAAAAAA SEQ ID
 NO: 17

'Tinción' de células con anticuerpos que tienen un código de barras de ADN para CITE-seq.

Se resuspendieron aproximadamente 500 000 células en PBS frío que contenía BSA al 2 % y Tween al 0,01 % y se filtraron a través de filtros celulares de 40 µm (Falcon, EE.UU.) para eliminar los posibles grupos y partículas grandes. A continuación, las células se incubaron durante 10 minutos con el bloque receptor de Fc (TruStain FcX, BioLegend, EE. UU.) para bloquear la unión de anticuerpos no específicos. Posteriormente, las células se incubaron con mezclas de anticuerpos con código de barras durante 30 min a 4 °C. Las concentraciones de anticuerpos fueron 1 µg por prueba, según lo recomendado por el fabricante (BioLegend, EE. UU.) para aplicaciones de citometría de flujo. Las células se lavaron 3 veces mediante resuspensión en PBS que contenía BSA al 2 % y Tween al 0,01 %, seguido de centrifugación (~480 g 5 min a 4 °C) e intercambio de sobrenadante. Después del lavado final, las células se resuspendieron a una concentración celular apropiada en PBS para aplicaciones Drop-seq1 o 10x Genomics3.

CITE-seq en la plataforma Drop-seq.

Drop-seq se realizó como se describe con modificaciones. Para el experimento mixto humano/ratón, las células se cargaron a una concentración de 400 células/µl para lograr una alta tasa de dobletes. Para los experimentos de PBMC, las células se cargaron a 150 células/µl. Se amplificó el ADNc durante diez ciclos y a continuación los productos se separaron por tamaño con perlas Ampure (Beckman Coulter, EE. UU) en fragmentos de <300 nt que contenían etiquetas derivadas de anticuerpo (ADT) y fragmentos de > 300 nt que contenían los ADNc derivados del ARNm celular. Se amplificaron los ADT durante diez ciclos más usando cebadores específicos que agregaban secuencias P5 y P7 para agrupamiento en celdas de flujo Illumina. Alternativamente, las etiquetas de anticuerpos se pueden amplificar directamente a partir de perlas de Drop-seq lavadas minuciosamente después de la amplificación de ARN-ADNc mediante el uso de cebadores específicos del anticuerpo-oligo y la perla Drop-seq-RT oligo. Los ADNc derivados del ARNm se convirtieron en genotecas de secuenciación mediante etiquetado como se describe¹. Después de la cuantificación, las genotecas se fusionaron a las concentraciones deseadas (10 % de un carril para ADT, genoteca de ADNc del 90 %). La secuenciación se realizó con un HiSeq 2500 Rapid Run y química v2 según las instrucciones del fabricante (Illumina, EE. UU).

CITE-seq en la plataforma 10x.

El ciclo de células individuales 10x se realizó según las instrucciones del fabricante (10x Genomics, EE. UU.) con modificaciones. Para el experimento mixto humano/ratón, (analizado en Single Cell 3' versión 1), ~17.000 células se cargaron para producir ~10.000 células con una tasa intermedia/alta de dobletes. Para el perfilado de CBMC (ciclo en Single Cell 3' versión 2), se cargaron ~7000 células para obtener un rendimiento de ~4000 células. Para el perfilado de CBMC, los inventores enriquecieron con células de ratón a baja frecuencia (~4 %). Esto permitió extraer los cortes de señal de anticuerpo a ruido y estimar las tasas de doblete verdaderas (4 %) en los experimentos de los inventores y comparar estas tasas con las estimaciones proporcionadas por el fabricante del equipo (~3,1 %) (véase más adelante). El ADNc durante diez ciclos y a continuación los productos se separaron por tamaño con perlas Ampure (Beckman Coulter, EE. UU) en fragmentos de <300 nt que contenían etiquetas derivadas de anticuerpo (ADT) y fragmentos de > 300 nt que contenían los ADNc derivados del ARNm celular. Se amplificaron los ADT durante diez ciclos más usando cebadores específicos que agregaban secuencias P5 y P7 para agrupamiento en celdas de flujo Illumina. Se generó una genoteca de secuenciación a partir de los ADNc derivados del ARN usando un enfoque basado en etiquetas

similar al usado en Drop-seq para los experimentos Single Cell 3' v1, o según las instrucciones del fabricante para los experimentos Single Cell 3' v2. Las genotecas de ADT y ADNc se fusionaron y secuenciaron como se ha descrito anteriormente.

5 Cultivo celular.

Las células HeLa (humanas), 4T1 (ratón) y 3T3 (ratón) se mantuvieron según los procedimientos estándar en medio de Eagle modificado por Dulbecco (Thermo Fisher, EE. UU.) suplementado con suero de feto de bovino al 10 % (FBS, Thermo Fisher, EE. UU.) a 37 °C con 5 % de CO₂. Para el experimento de mezcla de especies, las células HeLa y 4T1 se mezclaron en proporciones iguales y se incubaron con anticuerpos CITE-seq con código de barras de ADN como se ha descrito anteriormente. Para los enriquecimientos de ratón de baja frecuencia, -5 % de células 3T3 se mezclaron en el grupo de CBMC antes de realizar la CITE-seq.

Células mononucleares de la sangre.

Se aislaron células mononucleares de sangre del cordón umbilical (CBMC) (New York Blood Center) como se describe⁵⁷. Las células se mantuvieron en hielo durante y después del aislamiento. Se obtuvieron células mononucleares de sangre periférica de Allicells (EE. UU.).

20 Comparación entre citometría de flujo y CITE-seq.

Las células se tiñeron con una mezcla de anticuerpos marcados con fluoróforo (CD8a-FITC, BioLegend, EE. UU.) y anticuerpos oligoetiquetados CITE-seq del mismo clon de anticuerpo monoclonal (RPA-T8) dirigido a CD8a, a concentraciones recomendadas por el fabricante (1 ug por prueba, BioLegend, EE. UU.). Las células también se tiñeron con anticuerpo dirigido contra CD4-APC (RPA-T4, BioLegend, EE. UU.). Las células se clasificaron en conjuntos de diferentes niveles de expresión de CD8a usando el clasificador celular Sony SH800, que se hizo funcionar según las instrucciones del fabricante. A continuación, los grupos se dividieron en dos y se volvieron a analizar mediante citometría de flujo usando Sony SH800 o se procesaron para CITE-seq usando Drop-seq como se ha descrito anteriormente. Los datos de citometría de flujo se representaron mediante FlowJo v9 (EE.UU.).

30 Citometría de flujo de múltiples parámetros.

Las células se tiñeron con los siguientes anticuerpos de ratón contra Ig humana, que se adquirieron a BD Biosciences (EE. UU.). Los anticuerpos, clones y fluoróforos utilizados fueron CD3e (clon SK7) Hilyte 750 Aloficocianina (H7APC), CD4 (clon SK3) Brilliant Blue (BB) 630, CD8a (clon SK1) Ficoeritrina (PE), CD14 (clon M5E2) Brilliant Violet (BV) 750, CD19 (clon HIB19) BV570, CD11c (clon B-ly6) Cianina5 PE, CD2 (clon RPA-2.10) Brilliant Ultraviolet (BUV) 805, y CD57 (clon, NK-1) BB790. Después de lavar las células en PBS y fijarlas en paraformaldehído al 0,5 %, se adquirieron las muestras en un citómetro de flujo BD Symphony A5 y los datos se analizaron usando FlowJo v9 (EE.UU.).

40 Métodos informáticos.

Procesamiento y filtrado de datos de ARN de células individuales. Los datos sin procesar de Drop-seq se procesaron con el flujo estándar (Herramientas Drop-seq versión 1.12 del laboratorio McCarroll). Los datos 10x del experimento de mezcla de especies se procesaron mediante Cell Ranger 1.2 con los parámetros predeterminados y no se aplicó ningún filtrado adicional. Los datos 10x de los experimentos de CBMC (química v2) se procesaron con el mismo flujo que los datos de Drop-seq. Las lecturas se alinearon con la secuencia humana de referencia GRCh37/ hg19 (comparación de CD8a con FACS) o con una concatenación de hg19 y mm10 de ratón de referencia (experimento de mezcla de especies, CBMC).

Los datos Drop-seq del experimento de mezcla de especies se filtraron para contener solo células con al menos 500 UMI cartografiados para genes humanos o 500 UMI de cartografiado para genes de ratón. Para los datos de comparación de CD8a con FACS, se mantuvieron solo células con PCT_USABLE_BASES $\geq 0,5$ (fracción de bases que se cartografiaban con ARNm, esto forma parte de las métricas emitidas por la canalización de procesamiento predeterminada). Además, se eliminó cualquier célula con menos de 200 genes detectados y las células con un número total de UMI o genes (en log₁₀ después de añadir un pseudocuento) mayor de 3 s.d. por encima o por debajo de la media. Se usó la misma estrategia de filtrado para los datos de CBMC, siendo la única diferencia un umbral de gen de 500.

60 Procesamiento y filtrado de datos de ADT de células individuales.

Los códigos de barras de anticuerpos y células se extrajeron directamente de las lecturas en los archivos fastq. Dado que los códigos de barras de anticuerpos eran suficientemente diferentes en el experimento de mezcla de especies, también se contaron las secuencias con una distancia de Hamming inferior a 4. En la CBMC se contaron las secuencias con una distancia de Hamming inferior a 2. Las lecturas con la misma combinación de código de barras celular, molecular y de anticuerpo solo se contaron una vez. Solo se mantuvieron las células que pasaron los filtros

específicos de ARN y tenían un número mínimo de recuentos de ADT totales (recuentos mínimos utilizados: mezcla de especies, 10; Comparación FACS de CD8a, 1; CBMC, 50).

Normalización y agrupamiento de ARN de CBMC.

Después de la alineación de lectura y el filtrado celular, se asignó la especie a cada código de barras celular. Si más del 90 % de los recuentos de UMI provenían de genes humanos, el código de barras celular se consideró humano. Si fuera menos del 10 % de los recuentos de UMI, la especie asignada fue ratón. Los códigos de barras celulares con valores entre el 10 % y el 90 % de 'humano' se consideraron especies mixtas. La asignación resultante fue de 8005 lecturas humanas, 579 de ratón y 33 mixtas. Salvo que se indique lo contrario, el análisis se realizó solo en las células humanas y genes procedentes del genoma humano de referencia.

La matriz de recuentos de UMI se convirtió en una matriz de expresión normalizada logarítmica x con

$$x_{i,j} = \log \{ c_{i,j} \times 10,1000/m_j \}$$

donde $c_{i,j}$ es el recuento de moléculas del gen i en la célula j , y m_j es la suma de todos los recuentos de moléculas para la célula j . Después de normalizar, cada gen se escaló para tener una expresión media 0 y una varianza 1.

Se identificaron 556 genes de alta variabilidad ajustando una línea continua (LOESS, tramo = 0,33, grado = 2) a $\log_{10}(\text{var}(\text{UMI})/\text{media}(\text{UMI}))$ en función del $\log_{10}(\text{media}(\text{UMI}))$ manteniendo todos los genes con un residuo estandarizado por encima de 1 y una tasa de detección de al menos 1 %.

Para agrupar las células, se realizó una reducción de dimensionalidad seguida de una optimización de modularidad. Se llevó a cabo el análisis de componentes principales (PCA) usando la matriz de expresión de genes variables. Para determinar el número de dimensiones significativas, se observó el cambio porcentual en valores propios sucesivos. El último valor propio para caracterizar una reducción de al menos el 5 % constituyó el número significativo de dimensiones de los inventores (en este caso el número fue 13). Para la agrupación, se usó un algoritmo de optimización de modularidad que encuentra una estructura comunitaria en los datos⁵⁷. Los datos se representan en forma de una red ponderada siendo las células nodos y las similitudes de Jaccard al cuadrado como las ponderaciones del borde (basadas en la distancia euclidiana de PC significativos y un tamaño de vecindad de 40 (0,5 % de todas las células)). El algoritmo de agrupamiento, implementado en la función "cluster_lovain" del paquete igraph R, encuentra una partición de las células con alta densidad dentro de comunidades en comparación con entre comunidades. Para visualización 2D, se redujo aún más la dimensionalidad de los datos a 2 usando t-SNE^{58,34,59}.

Normalización y agrupamiento de etiquetas derivadas de anticuerpos CBMC. Dado que cada recuento de ADT en una célula dada puede interpretarse como parte de un todo (todos los recuentos de ADT asignados a esa célula), y solo hay 13 componentes en este experimento, se trataron este tipo de datos como datos de composición y se aplicó la transformación de relación logarítmica (CLR) centrada⁶¹. Explícitamente, se generó un nuevo vector ADT transformado por CLR y para cada célula donde

$$y = \text{clr}(x) = [\ln((x_1/g(x))), \ln((x_2/g(x))), \dots, \ln((x_5/g(x)))],$$

y x es el vector de recuentos de ADT (incluido un pseudorecuento para cada componente), y $g(x)$ es la media geométrica de x .

Se observó que los recuentos de ADT estaban en escalas algo distintas para los diferentes anticuerpos, que quizás estaba producido por diferencias en la especificidad del anticuerpo y/o la abundancia del epítipo. Para compensar los desplazamientos resultantes en la señal de ADT inicial no específica, los inventores examinaron la distribución de densidad de los recuentos ADT transformados por CLR en todos los anticuerpos, por separado para células humanas y de ratón. Para cada ADT, se determinó la media y la varianza de las células de ratón y los inventores definieron el punto de corte independiente de la especie (separar el estado "on" del estado "off" donde la proteína está presente) para que fuera una desviación estándar mayor que la media.

Para agrupar células según los recuentos ADT, se tomó el mismo enfoque general que para los datos de ARN, salvo que no se realizó una reducción de dimensionalidad. En cambio, los inventores restaron los cortes derivados de ratón de los recuentos de ADT transformados por CLR para cada anticuerpo. Los pesos entre células fueron similitudes Jaccard al cuadrado, basadas en la distancia euclidiana y el tamaño de vecindad de 0,5 % del número total de células.

Estimación de la tasa de dobletes usando enriquecimiento con células de ratón a baja frecuencia.

La adición de células de ratón a baja frecuencia permitió estimar las tasas de doblete verdaderas (4 %) en el experimento de perfilado de CBMC de los inventores, así como comparar dichas tasas con las estimaciones proporcionadas por el fabricante del equipo (-3,1 %). Para estimar la tasa de dobletes en los experimentos de los inventores, se modelizó el proceso de captura de células en gotículas como una distribución de Poisson con una tasa

de carga lambda y una fracción de células de ratón fija de 6,5 %. Se optimizó el valor de lambda para que los datos simulados coincidieran más estrechamente con la distribución de especies observada. El valor de lambda resultante fue 0,068, y la tasa de dobletes (fracción de gotículas con más de una célula de todas las gotículas que tenían al menos una célula) observada en las simulaciones fue del 4 %.

Ejemplo 3: Identificación de diferentes especies en la muestra mixta.

Los complejos de anticuerpo-oligo descritos en el Ejemplo 1 se incubaron con células usando condiciones establecidas para citometría de flujo, tal como la referencia 22. Las células se lavaron para eliminar los anticuerpos no unidos, a continuación, las células individuales se encapsularon en gotículas acuosas de tamaño de nanolitros en un aparato de microfluidos diseñado para realizar la Drop-seq¹ (Figura 1C). Después de la lisis celular (que se produjo inmediatamente en las gotículas cuando el tampón de lisis entró en contacto con las células), los ARNm celulares se hibridaron con perlas Drop-seq que contenían poli T (Figura 1B) a través de su cola de poli A (Figura 1C/6). Los oligos de los anticuerpos también se hibridaron con las perlas Drop-seq mediante su tramo de poli A en el extremo 3'. Una secuencia de código de barras única en la perla Drop-seq indexó el transcriptoma de cada célula coencapsulada. Después de romper la emulsión y eliminar el aceite, la retrotranscripción extendió el oligo del código de barras para crear el ADNc de la primera hebra a partir tanto de ARNm como de moldes de oligo derivados de anticuerpo. El ADNc y las etiquetas derivadas de anticuerpos se separaron por tamaño (Figura 2C) y se convirtieron independientemente en genotecas listas para Illumina. Los dos tipos de genoteca se secuenciaron conjuntamente. Debido a las ventajas de generar genotecas por separado, las proporciones relativas de las genotecas también se adaptan para garantizar que se obtiene la profundidad de secuenciación adecuada.

Tanto las genotecas de ADNc como las genotecas de etiquetas de oligo recuperadas de grupos de células individuales marcadas con anticuerpos se generaron y se separaron físicamente antes de preparar de forma independiente las genotecas para secuenciación como se muestra en la Figura 2C. Se realizó la secuenciación de las genotecas de etiquetas de oligo recuperadas. Todas las células humanas y de ratón en una suspensión de cientos de miles de células humanas y de ratón mezcladas se identificaron inequívocamente basándose en sus proteínas de superficie celular específicas de especie (Figura 2D) después de usar independientemente los datos de su transcriptoma para identificarlas (Figuras 2D y 2E). La gran mayoría de las gotículas identificadas como contenedores de células humanas (Figuras 2D y 2E, rodeadas por una línea continua) y células de ratón (Figuras 2D y 2E, rodeadas por una línea discontinua) según la alineación de transcriptoma también se identificaron por tener los mismos epítomos de especie unidos a la superficie usando el enfoque de etiquetado de oligos (Figura 2D).

En este experimento, se usó deliberadamente una elevada concentración celular para obtener altas tasas de dobletes (gotículas que contenían dos o más células), para correlacionar los datos de transcriptoma de especies mixtas con las señales de proteínas de especies mixtas procedentes de las gotículas individuales. Véanse, por ejemplo., las Figuras 2D y 2E, en las que los puntos que no están sobre el eje son señales mixtas. Las gotículas que contenían mezclas de células humanas y de ratón también tenían claramente lecturas de secuenciación de anticuerpos humanos y de ratón (Figura 2D, puntos no rodeados). Este resultado ilustró que pueden obtenerse señales de múltiples anticuerpos de una gotícula.

Este método se expande para medir simultáneamente un gran número de marcadores celulares basados en anticuerpos y transcriptoma establecidos en decenas de miles de células en paralelo.

Un experimento adicional produce mediciones cualitativas de la expresión de proteínas de la superficie celular junto con los datos de expresión de todo el transcriptoma. El experimento se lleva a cabo para confirmar que la señal del oligo del anticuerpo refleja la concentración de epítipo en la superficie celular. Se han identificado sesgos experimentales procedentes de las siguientes fuentes: 1) señal artificial que surge de la duplicación de la secuenciación de la genoteca por PCR, 2) reactividad cruzada y disponibilidad de suficientes especies de anticuerpos bien caracterizadas y 3) niveles variables de conjugación entre el oligo y los anticuerpos que produce una estimación imprecisa de la concentración de epítipo.

Los métodos para corregir los sesgos identificados anteriormente incluyen los siguientes. En primer lugar, los duplicados de PCR en conjuntos de datos de secuenciación se filtran usando identificadores moleculares únicos (UMI) integrados en el diseño del oligo de Drop-seq. En segundo lugar, para abordar la reactividad cruzada de anticuerpos, se evita el uso de anticuerpos con baja especificidad y solo se usan anticuerpos de citometría de flujo altamente optimizados y probados para los experimentos de estudio a escala de laboratorio. Los anticuerpos optimizados están disponibles de grandes consorcios tales como el Human Protein Atlas, que producen continuamente más anticuerpos para suplementar el grupo ya existente de miles de anticuerpos específicos⁷. En tercer lugar, usando la conjugación estreptavidina-biotina, y según la publicación del fabricante, se estimó que aproximadamente 4-12 moléculas de oligonucleótido se unen a cada molécula de anticuerpo.

Se prueban diferentes estrategias de conjugación de anticuerpo-oligo, como se ha identificado anteriormente en la definición de "unión", para marcar anticuerpos con un número definido de moléculas de oligonucleótido y obtener una medición más cuantitativa. Etiquetar una molécula de oligo para cada anticuerpo, junto con el uso de los UMI, hace que el método sea como mínimo tan cuantitativo como los enfoques basados en inmuno-PCR. Con este fin se realizan

evaluaciones para probar si las señales de una sola molécula pueden medirse de forma segura por encima del ruido en la cuantificación final de la genoteca. La concentración óptima de anticuerpos también se determina mediante experimentos de titulación con anticuerpos individuales como se realiza en citometría de flujo. Como patrón de referencia, se prueban los mismos anticuerpos monoclonales en un ciclo de citometría de flujo usando las mismas poblaciones celulares. Esto permite una determinación de la sensibilidad y la potencia cuantitativa de la medición de oligonucleótidos basada en secuenciación.

Ejemplo 4: Identificación de linajes de células mieloides y linfoides.

Los linajes de células mieloides y linfoides se han estudiado ampliamente mediante la expresión de marcadores de superficie celular en citometría de flujo y también se pueden identificar en función de sus perfiles de expresión génica. Numerosos anticuerpos de citometría de flujo bien establecidos y altamente específicos que reconocen marcadores globales de los linajes mieloides y linfoides y para subpoblaciones celulares específicas dentro de estos linajes se unen a los oligos descritos en el Ejemplo 1.

Los complejos de anticuerpo-oligo generados de este modo se incuban con células usando condiciones establecidas para citometría de flujo. Las células se lavan para eliminar los anticuerpos no unidos, a continuación, las células individuales se encapsulan en gotículas acuosas de tamaño de nanolitros en un aparato de microfluidos diseñado para realizar la Drop-seq¹ (Figura 1C). Después de la lisis celular (que se produce inmediatamente en las gotículas cuando el tampón de lisis entra en contacto con las células), los ARNm celulares se hibridan con perlas Drop-seq que contienen poli T (Figura 1B) a través de su cola de poli A. Los oligos del anticuerpo se hibridan con las perlas Drop-seq mediante su tramo de poli A en el extremo 3'. Una secuencia de código de barras única en la perla Drop-seq indexa el transcriptoma de cada célula coencapsulada. Después de romper la emulsión y eliminar el aceite, la retrotranscripción extiende el oligo del código de barras para crear el ADNc de la primera hebra a partir tanto de ARNm como de moldes de oligo derivados de anticuerpo. El ADNc y las etiquetas derivadas de anticuerpos se separan por tamaño, se convierten en genotecas listas para Illumina y se secuencian.

Se usan múltiples complejos de anticuerpo-oligo para probar la capacidad de multiplexación del método. En un experimento adicional, se prueban más de 100 complejos de anticuerpo-oligo.

Ejemplo 5: Identificación de células basándose en proteínas expresadas intracelularmente.

Se investigan procedimientos de fijación y permeabilización de células en condiciones suaves distintos a los usados para la tinción de anticuerpos intracelulares en ensayos FACS específicos de señalización¹⁸ para determinar si estos son compatibles con el ARN. Se generan complejos de anticuerpo-oligo que reconocen proteínas intracelulares como se describe en el Ejemplo 1. Además, el procedimiento establecido de permeabilización y fijación se lleva a cabo antes del paso de incubación del protocolo CITE-seq. Las células se identifican basándose en proteínas expresadas intracelularmente y los transcritos de ARNm. Este método no solo proporciona una caracterización más detallada de las poblaciones celulares, sino que también permite el estudio de la regulación génica posterior a la transcripción y posterior a la traducción en células individuales a una profundidad sin precedentes.

Ejemplo 6: Identificación de células.

Los métodos descritos en los Ejemplos 1 a 5 se adaptan a otras tecnologías de secuenciación de células individuales basados en gotículas o micropocillos como se ha descrito anteriormente. El tramo de poli A en el extremo 3' de los anticuerpo-oligos permite la captura en cualquier protocolo de secuenciación de ARNm basado en oligo-dT, tal como el descrito en Mortazavi y col²³. Se evalúan los parámetros específicos del análisis, y se valora la utilidad del método para instrumentos disponibles comercialmente (por ejemplo, 10x Genomics) y otras tecnologías que están en desarrollo¹¹.

Ejemplo 7: Agrupación celular mejorada y clasificación de células mononucleares de sangre del cordón umbilical

Se llevó a cabo un análisis CITE-seq sobre 8700 células mononucleares de la sangre marcadas con 10 construcciones de anticuerpo como se describe en la presente memoria que tiene los componentes incluidos en la Tabla 2. Se realizaron la tSNE (incrustación estocástica de vecinos distribuida en t)³⁴ y la agrupación usando análisis de correlación canónico, que integra mediciones de proteínas y ARN.

TABLA 2: Construcciones de CITE-seq

Núm.	Anticuerpo/Diana	Química de conjugación	Enlazador escindible	Sonda de amplificación	Código de barras	Anclaje
1	Anti-CD3 para diana CD3	SAV	Disulfuro	Secuencia compatible de ARN pequeño con Illumina® TruSeq® CCTGGCACCCGAG AATTCCA Id de sec. n.º: 18	ATCACG	30x A

TABLA 2: Construcciones de CITE-seq						
Núm.	Anticuerpo/Diana	Química de conjugación	Enlazador escindible	Sonda de amplificación	Código de barras	Anclaje
2	Anti-CD19 para diana CD19	SAV	Disulfuro	Igual que para el n.º 1	CGATGT	30x A
3	Anti-CD4 para diana CD4	SAV	Disulfuro	Igual que para el n.º 1	TGACCA	30x A
4	Anti-CD8a para diana CD8a	SAV	Disulfuro	Igual que para el n.º 1	GCCAAT	30x A
5	Anti-CD56 para diana CD56	SAV	Disulfuro	Igual que para el n.º 1	CTTGTA	30x A
6	Anti-CD16 para diana CD16	SAV	Disulfuro	Igual que para el n.º 1	TTAGGC	30x A
7	Anti-CD11c para diana CD11c	SAV	Disulfuro	Igual que para el n.º 1	TAGCTT	30x A
8	Anti-CD34 para diana CD34	SAV	Disulfuro	Igual que para el n.º 1	GGCTAC	30x A
9	Anti-CD14 para diana CD14	SAV	Disulfuro	Igual que para el n.º 1	AGTCAA	30x A
10	Anti-CD10 para diana CD10	SAV	Disulfuro	Igual que para el n.º 1	AGTTCC	30x A

Un resultado de este análisis es la membrana de transferencia puntual de la Figura 3A. Esta representación muestra que CITE-seq permite una mayor agrupación de células y clasificación de células mononucleares de sangre de cordón umbilical.

Otro análisis CITE-seq del mismo conjunto de datos de la Figura 3A se llevó a cabo usando solamente datos de ARN. Por ejemplo, células T CD8 y CD4 no se separaron en poblaciones diferenciadas. Estos resultados se muestran en la Figura 3B, en la que la gráfica de dispersión demuestra una resolución mejorada cuando se usan datos multimodales. Los símbolos de la figura son Mono (para monocitos), B para células B, T para células T, NK para células citolíticas naturales, DC para células dendríticas convencionales, pDC para plasmacitoide DC, Pre para precursores, y Ery para eritroblastos.

Finalmente, las representaciones biaxiales de los datos de anticuerpos CITE-seq para anticuerpos seleccionados, es decir, las construcciones núms. 1- 10 de la Tabla 2, se demuestran en la Figura 3C. Los datos son comparables a lo que se obtiene por citometría de flujo con la diferencia significativa de que el transcriptoma para cada célula individual (cada punto) dentro de la representación también está disponible cuando se usa la metodología CITE-seq y las construcciones. Por lo tanto, las células se pueden analizar y clasificar adicionalmente en función de sus datos de ARN, datos de proteínas o ambos.

Se generó otra serie de representaciones biaxiales mediante la multiplexación de 8700 células mononucleares de sangre marcadas con las 10 construcciones de anticuerpos de la Tabla 2 en un análisis CITE-seq como se ha descrito anteriormente. La Figura 4 muestra representaciones biaxiales de datos de anticuerpos CITE-seq para las 10 construcciones de anticuerpos de la Tabla 2. Los datos son comparables a lo que se obtiene por citometría de flujo con la diferencia significativa de que el transcriptoma para cada célula individual (cada punto) dentro de estas representaciones también está disponible cuando se usa CITE-seq. Por lo tanto, las células se pueden analizar y clasificar adicionalmente en función de sus datos de ARN, datos de proteínas o ambos.

Ejemplo 8: Cite-SEQ

Se describe la indexación celular de transcriptomas y epítomos mediante secuenciación (CITE-seq), un método en el que se usan anticuerpos etiquetados con oligonucleótidos para integrar las mediciones de proteínas celulares y transcriptomas en una lectura eficiente de células individuales. CITE-seq es compatible con los enfoques existentes de secuenciación de células individuales y se escala fácilmente con aumentos de rendimiento. El método CITE-seq combina la detección de marcadores de proteínas con alto multiplexado y el perfilado del transcriptoma no sesgado para miles de células individuales. El método se adapta fácilmente a dos aplicaciones scRNA-seq de alto rendimiento y muestra que el análisis de datos multimodal puede lograr una caracterización más detallada de fenotipos celulares que las mediciones de transcriptoma en solitario.

Los inventores diseñaron una lectura digital basada en secuenciación para los niveles de proteína mediante la conjugación de anticuerpos a oligonucleótidos (oligos) que se pueden capturar mediante cebadores oligo-dT (usados en la mayoría de preparaciones de genotecas scRNA-seq), contienen un código de barras para la identificación de

anticuerpos e incluyen una sonda para la amplificación por PCR. Una interacción estreptavidina-biotina comúnmente usada une el extremo 5' de los oligos a los anticuerpos. Los complejos de anticuerpo-oligo se incuban con suspensiones de células individuales en condiciones comparables a los protocolos de tinción por citometría de flujo; después de esta incubación, las células se lavan para eliminar los anticuerpos no unidos y se procesan para scRNA-seq. En el ejemplo de los inventores, se encapsularon células individuales en gotículas acuosas de tamaño de nanolitro en un aparato de microfluidos diseñado para realizar la Drop-seq¹. Después de la lisis celular en gotículas, los ARNm celulares y los oligos derivados de anticuerpos se hibridan a través de sus colas de poli A 3' a las perlas Drop-seq que contienen oligo-dT y se indexan mediante un código de barras celular compartido durante la retrotranscripción. Los ADNc amplificados y las etiquetas derivadas de anticuerpos (ADT) pueden separarse por tamaño y convertirse en genotecas de secuenciación Illumina independientemente. Es importante destacar que los dos tipos de genoteca se generan por separado, sus proporciones relativas se pueden ajustar en un solo carril agrupado para asegurar que se obtiene la profundidad de secuenciación requerida para cada genoteca. Véanse, por ejemplo, la referencia 46 y los datos publicados en línea.

Para evaluar la capacidad del método de los inventores para distinguir células individuales según la expresión de proteínas de superficie, los inventores diseñaron un experimento de “mezcla de especies” como prueba de concepto que aprovecha el marcador CD29 específico de especie de expresión elevada (Integrina beta-1). Una suspensión de células humanas (HeLa) y de ratón (4T1) se incubó con una mezcla de anticuerpos anti-ratón con código de barras y anti-CD29 humano. Después del lavado para eliminar los anticuerpos no unidos, se realizó la Drop-seq¹ para investigar la concordancia entre especies de origen de los transcritos y los ADT. Se usó deliberadamente una elevada concentración celular para obtener altas tasas de multipletes (gotículas que contenían dos o más células) para correlacionar los datos de transcriptoma de especies mixtas con las señales de ADT de especies mixtas procedentes de las gotículas individuales. La mayoría de las gotículas (97,2 %) que se identificaron como contenedoras de células humanas, de ratón o mixtas según el transcriptoma recibieron la misma clasificación de especies según los recuentos de ADT. Los recuentos de células basados en ARN o ADT están muy correlacionados entre ambos métodos y esto demuestra la baja tasa de abandono de las señales de ADT. Se realizó el mismo experimento usando un sistema disponible comercializado por 10x Genomics y se obtuvieron resultados comparables.

Los inventores deseaban caracterizar la naturaleza cuantitativa de la lectura de proteínas CITE-seq. La citometría de flujo es el patrón de oro para enumerar subconjuntos de células según diferencias cuantitativas en los marcadores de superficie^{47,48}. Por lo tanto, se pretendía igualar la sensibilidad de la detección de proteínas en CITE-seq con la de la citometría de flujo usando anticuerpos CITE-seq dirigidos contra marcadores de citometría de flujo comunes para identificar y discriminar subpoblaciones inmunitarias. Se realizaron experimentos de citometría de flujo multiparámetro y de CITE-seq usando el mismo conjunto de anticuerpos sobre alícuotas del mismo grupo de células mononucleares de sangre periférica. Usando los niveles de ADT, los inventores pudieron construir representaciones de agrupación “bixial” parecidas a la citometría y comparar estos datos cualitativa y cuantitativamente con datos de citometría de flujo. Los perfiles de distribución celular basados en la expresión de proteínas marcadoras asociadas con diversos subconjuntos de células T, células B, plasmacitoides, células dendríticas mieloides y monocitos fueron notablemente similares.

A continuación, los inventores se preguntaron si las diferencias cuantitativas en la expresión observadas en citometría de flujo pueden observarse en CITE-seq. Para esto, los inventores se centraron en el marcador CD8a, ya que sus niveles varían ampliamente entre poblaciones de células mononucleares de sangre de cordón umbilical (CBMC) con conjugados de anticuerpo CITE-seq y anticuerpos conjugados con fluoróforo, de modo que algunos epítomos CD8a de cada célula se marcarían con fluoróforos y algunos con el oligo. Las células se sometieron a clasificación celular activada por fluorescencia (FACS) en grupos separados según la fluorescencia de CD8a (muy alto (+++), alto (++), intermedio (+) y bajo (+/-)). A continuación, cada grupo se dividió y se volvió a analizar por separado mediante citometría de flujo y CITE-seq. Para cada grupo definido por FACS, se observaron niveles de expresión de CD8a relativos similares según ambos métodos. Se concluye que los niveles de ADT de CITE-seq son consistentes con la citometría de flujo como patrón oro y, por lo tanto, pueden permitir el inmunofenotipificado de alta resolución concertada con la transcriptómica.

El sistema inmunitario se ha perfilado ampliamente usando marcadores de superficie celular⁴⁷ y scRNA-seq^{3,9,49} y ambos métodos identifican de forma fiable los mismos tipos de células en proporciones consistentes. Una población de células inmunitarias complejas es, por lo tanto, un sistema ideal para validar la lectura multimodal de CITE-seq. Se preparó un panel CITE-seq de 13 anticuerpos monoclonales bien caracterizados que reconocen proteínas de superficie celular usadas rutinariamente como marcadores para la clasificación de células inmunitarias. Para estimar la unión de anticuerpos de fondo no específicos dentro de los experimentos, se desarrolló un control de “enriquecimiento” de bajo nivel. Una población de células murinas enriquecida rara debe distinguirse fácilmente transcriptómicamente, pero no debe reaccionar de forma cruzada con los anticuerpos dirigidos contra proteínas humanas; lo que permitiría definir los niveles de ADT de fondo directamente a partir de los datos. Por lo tanto, se agregaron los fibroblastos 3T3 de ratón (~4 %) a las CBMC de los inventores, se incubó el grupo de células con el panel de anticuerpos CITE-seq de los inventores y se aplicó el flujo de trabajo de células individuales de 10x Genomics a un total de 8,005 células. La agrupación no supervisada basa en el gráfico usando la expresión de ARN reveló tipos de células reconocibles que expresan genes marcadores seleccionados. Las células murinas se agruparon por separado (datos no mostrados) y mostraron recuentos de ADT bajos para cada marcador, y esto permitió establecer

una línea de base para la señal frente al ruido para delinear más claramente las poblaciones celulares positivas de las negativas. Con este paso de determinación de umbral, se identificaron tres conjugados de anticuerpo-oligo sin unión específica (es decir, sin umbral sobre el fondo de señalización) y se excluyeron de un análisis posterior.

Se detectó un fuerte enriquecimiento de ADT en las poblaciones inmunitarias correctas de CD3e dentro del grupo de células T; CD4 y CD8a en subpoblaciones de células T en gran parte no superpuestas; CD19 casi exclusivamente en células B; CD56, CD16 y CD8a en el grupo NK; y CD11c y CD14 en el grupo de monocitos y células dendríticas. También se identificó una población de células precursoras raras a menos del 2 % en sangre de cordón umbilical (células CD34+). Los recuentos de ADT por célula fueron mayores que los niveles de ARNm para los mismos genes y fueron menos propensos a los eventos de “abandono”. Consistente con esto, los inventores encontraron una correlación baja entre el ARNm y ADT en células individuales y una mayor correlación cuando se promedia la expresión dentro de las agrupaciones. Se usaron los niveles de ADT y la información de agrupamiento basada en transcriptoma para construir gráficos de clasificación “biaxial” de CITE-seq multimodal; esto reveló perfiles similares que están bien establecidos por citometría de flujo. Por ejemplo, los inventores pudieron resolver una fuerte anticorrelación de los niveles de ADT para CD4 y CD8a en las células T y las diferencias cuantitativas en la expresión del marcador entre subconjuntos: estas incluían las diferencias de expresión de CD8a entre las células NK y T o de CD4 entre monocitos y células T. Además, la agrupación basada en los niveles de ADT da como resultado una separación de tipo celular clara y consistente (véase la Ref. 46 para las Figuras de los datos no mostrados).

A continuación, los inventores se preguntaron si los datos multimodales de CITE-seq podrían mejorar la caracterización de los fenotipos de células inmunitarias en comparación con scRNA-seq en solitario. Se observó un gradiente opuesto de los niveles de ADT para CD56 y CD16 dentro de la agrupación de células NK derivada transcriptómicamente de los inventores, revelando potencialmente subconjuntos CD56bright y CD56dim^{50,51}; por lo tanto, los inventores subdividieron el grupo de células NK según los niveles ADT de CD56. Cuando se comparan los perfiles moleculares de estos grupos, se observaron cambios en proteína y ARN que eran altamente consistentes con la bibliografía^{50,51}. Se observó una complementariedad aparente entre los niveles de los ADT de CD16 y en menor medida de los de CD8a en comparación con los ADT de CD56 dentro de estos dos subconjuntos. Para 11 genes que previamente se han caracterizado por expresarse diferencialmente dentro de estos subtipos⁵⁰⁻⁵², se detectó una regulación positiva o una regulación negativa consistente con la bibliografía en diez casos, incluyendo para GZMB, GZMK y PRF1. Esto ilustra el potencial de los análisis integrados y multimodales para mejorar el descubrimiento y la descripción de fenotipos celulares, particularmente cuando se diferencian entre poblaciones celulares con diferencias transcriptómicas sutiles.

La capacidad de superponer mediciones moleculares adicionales sobre los datos de scRNA-seq representa una dirección muy interesante para la comunidad de investigación de células individuales. CITE-seq permite un análisis multimodal de células individuales a la escala proporcionada por enfoques de secuenciación de células individuales basadas en gotículas. Los inventores demostraron el valor del análisis multimodal para revelar fenotipos que no se podrían descubrir usando sRNA-seq en solitario, y también prevén el uso de CITE-seq para estudios de regulación génica postranscripcional para células individuales. A diferencia de la citometría de flujo y la de masa, la detección de anticuerpos codificados con oligo no está limitada por la colisión de señales; una secuencia de 10 nt puede codificar fácilmente más códigos de barras que las proteínas humanas, y esto permite la inmunofenotipificación a gran escala con paneles de decenas a cientos de anticuerpos. Además, los procedimientos de fijación y permeabilización de células en condiciones suaves usados en ensayos de citometría intracelular también deben ser compatibles con CITE-seq, y pueden expandir significativamente el número de marcadores útiles.

También se puede prever una versión modificada de CITE-seq en la que solo se analizan los ADT a una escala masivamente paralela sin capturar los ARNm celulares (citometría por secuenciación).

Finalmente, los inventores han demostrado que CITE-seq es completamente compatible con una plataforma de células individuales comercial (10x Genomics) y debe ser fácilmente adaptable a otras tecnologías de secuenciación de células individuales de alto rendimiento basadas en gotículas, micropocillos e indexación combinatoria^{2,54,55,20,30} con ningunas o pocas personalizaciones.

Ejemplo 9: Variaciones de CITE-SEQ

En un experimento, la lectura de CITE-seq se comparó con diferentes tecnologías de conjugación. Una tecnología usó el enlace biotina-estreptavidina (SAV) descrito previamente en los Ejemplos 1-8. Otro método para la conjugación de anticuerpo-oligo empleó conjugación covalente mediante química de iEDDA como se ha descrito anteriormente⁴⁵. La química de conjugación iEDDA usada es comparable a las químicas de conjugación ofrecidas en kits comerciales (Innova Biosciences, Thunderlink PLUS kit).

Se llevó a cabo un análisis CITE-seq sobre 4000 células mononucleares de sangre periférica (PBMC) marcadas con una de 6 construcciones de anticuerpo como se describe en la presente memoria que tiene los componentes incluidos en la Tabla 3. Se realizaron la tSNE (incrustación estocástica de vecinos distribuida en t)³⁴ y la agrupación usando análisis de correlación canónico, que integra mediciones de proteínas y ARN.

TABLA 3: Construcciones de CITE-seq					
Núm.	Anticuerpo/Diana	Química de conjugación	Sonda de amplificación	Código de barras	Anclaje
1	Anti-CD3 para diana CD3	SAV	Secuencia compatible de ARN pequeño con Illumina® TruSeq®	TTC GTG AGG T	30x A
			CCTTGGCACCCGAGAA TTCCA SEQ ID NO: 19	SEQ ID NO: 20	
2	Anti-CD3 para diana CD3	covalente	Igual que para el n.º 1	TCT TCG TCC A SEQ ID NO: 21	30x A
3	Anti-CD4 para diana CD4	SAV	Igual que para el n.º 1	ATG CTC TAC C SEQ ID NO: 22	30x A
4	Anti-CD4 para diana CD4	covalente	Igual que para el n.º 1	TAG ACA GCT G SEQ ID NO: 23	30x A
5	Anti-CD8 para diana CD8	SAV	Igual que para el n.º 1	ATG GAG GTA G SEQ ID NO: 24	30x A
6	Anti-CD8 para diana CD8	covalente	Igual que para el n.º 1	AGA TGA ACC C SEQ ID NO: 25	30x A

En las Figuras 5A y 5B adjuntas, los perfiles (histogramas) de diferentes poblaciones (por ejemplo, células NK, CD4, CD8) parecen comparables en la SAV y la conjugación directa.

5 Ejemplo 10: Variaciones en el uso de las construcciones de CITE-SEQ

- A pesar de los avances rápidos en la tecnología de secuenciación de células individuales, los efectos de lotes específicos de muestra, la detección de duplicaciones de células y el coste de generación de conjuntos de datos masivo siguen siendo desafíos pendientes. Aquí, los inventores introducen el “*hashing*” celular, donde los anticuerpos oligoetiquetados contra proteínas de superficie expresadas de forma ubicua se usan para marcar de forma única células de muestras distintas, que pueden agruparse posteriormente. Al secuenciar estas etiquetas junto con el transcriptoma celular, se puede asignar cada célula a su muestra de origen e identificar de manera sólida los dobletes que se originan en múltiples muestras. Los inventores demuestran su estrategia al agrupar ocho muestras de PBMC humanas en un único ciclo del sistema de 10x Chromium, reduciendo sustancialmente los costes por célula para la generación de genotecas. El “*hashing*” celular está inspirada por, y es complementario de, estrategias de multiplexación elegantes basadas en la variación genética⁷¹, que también se aprovechan para validar los resultados de los inventores. Por lo tanto, se prevé que el enfoque de los inventores ayude a generalizar los beneficios de la multiplexación de células individuales a diversas muestras y diseños experimentales.
- La genómica de células individuales ofrece enormes promesas para transformar el entendimiento de los procesos heterogéneos de los inventores y para reconstruir taxonomías no supervisadas de tipos celulares^{63,64}. A medida que los estudios han progresado al perfilado de tejidos humanos complejos^{65,66} e incluso organismos completos^{20,67}, existe una creciente apreciación de la necesidad de tecnologías masivamente paralelas y conjuntos de datos para descubrir estados raros y sutiles de las células¹⁻³. Si bien el coste por célula de la preparación de la genoteca ha caído, el perfilado rutinario de decenas a cientos de miles de células sigue siendo caro tanto para laboratorios individuales como para consorcios tales como el Human Cell Atlas⁶⁸. Los desafíos ampliamente relacionados también siguen, entre los que se incluyen la identificación sólida de señales de artefactos que aparecen de dobletes celulares o efectos de lotes dependientes de la tecnología⁶⁹. En particular, identificar de manera fiable los perfiles de expresión correspondientes a más de una célula (en la presente memoria como “multipletes”) sigue siendo un desafío no resuelto en el análisis

de ARN-seq de células individuales (scRNA-seq), y una solución robusta mejorará simultáneamente la calidad de los datos y permitirá un mayor rendimiento experimental. Si bien se espera que los multipletes generen genotecas de mayor complejidad en comparación con los singletes, la intensidad de esta señal no es suficiente para una identificación inequívoca⁶⁹. De manera similar, se ha demostrado que los efectos técnicos y “discontinuos” enmascaran la señal biológica en el análisis integrado de los experimentos de scRNA-seq⁷⁰, necesitando soluciones experimentales para mitigar estos desafíos.

Los desarrollos recientes han demostrado de manera evidente cómo la multiplexación de muestras puede superar simultáneamente múltiples desafíos^{71,72}. Por ejemplo, el algoritmo demuxlet¹¹ permite el agrupamiento de muestras con genotipos distintos en un único experimento de scARN-seq. Aquí, los polimorfismos genéticos específicos de la muestra sirven como huella digital de la muestra de origen y, por lo tanto, se pueden usar para asignar cada célula a un individuo después de la secuenciación. Este flujo de trabajo también permite la detección de multipletes que se originan en dos individuos, reduciendo los multipletes no identificables a una velocidad que es directamente proporcional al número de muestras multiplexadas. Si bien este enfoque elegante requiere muestras agrupadas procedentes de individuos genotipados previamente, en principio cualquier enfoque que asigne huellas digitales de muestra que puedan medirse junto con scRNA-seq permitiría una estrategia similar. Por ejemplo, la multiplexación de muestras se utiliza frecuentemente en la citometría de flujo y la de masa marcando muestras distintas con anticuerpos para la misma proteína de superficie expresada de forma ubicua, pero conjugada con diferentes fluoróforos o isótopos, respectivamente^{73,74}.

Recientemente se introdujo CITE-seq⁴⁶, donde se usan anticuerpos etiquetados con oligonucleótidos para convertir la detección de proteínas de la superficie celular en una lectura secuenciable junto con scRNA-seq. Los inventores razonaron que un conjunto definido de anticuerpos oligoetiquetados contra proteínas de superficie ubicuas podría etiquetar de forma única diferentes muestras experimentales. Esto permite agruparlos y usar la señal de anticuerpo con código de barras como huella digital para una demultiplexación fiable. Se hace referencia a este enfoque como “hashing” celular, ya que el conjunto de oligos de los inventores define una “tabla de consulta” para asignar cada célula multiplexada a su muestra original. Se demuestra este enfoque marcando y agrupando ocho muestras de PBMC humanas y analizándolas simultáneamente en un único ciclo de scRNA-seq basada en gotículas. Los *hashtags* de células permiten una multiplexación robusta de muestras, una identificación fiable de multipletes y la discriminación de células de baja calidad del ARN ambiental. Además de permitir la “supercarga” de las plataformas de scRNA-seq comerciales para reducir sustancialmente los costes, esta estrategia representa un enfoque generalizable para la identificación de dobletes y la multiplexación que puede adaptarse a cualquier muestra biológica o diseño experimental.

A. Demultiplexado habilitado para hashtags basado en la expresión de proteínas de superficie ubicuas

Se buscó extender estrategias de multiplexación basadas en anticuerpos^{73,74} a sRNA-seq usando una modificación del método *CITE-seq* de los inventores. Los inventores eligieron un conjunto de anticuerpos monoclonales dirigidos contra marcadores de superficie inmunitaria ubicua y altamente expresados (CD45, CD98, CD44 y CD11a), combinaron estos anticuerpos en ocho conjuntos idénticos (grupos de la A hasta la H) y posteriormente conjugaron cada grupo a un oligonucleótido de *hashtag* distinto (en lo sucesivo denominado HTO, Figura 6A). Los HTO contienen un código de barras único de 12 pb que se puede secuenciar junto con el transcriptoma celular, solo con modificaciones menores de los protocolos sRNA-seq estándar. Se utilizó una química de conjugación mejorada y simplificada en comparación con el enfoque anterior de los inventores mediante el uso de química clic IEDDA para unir covalentemente oligonucleótidos a anticuerpos⁴⁵.

La estrategia de los inventores se diseñó para permitir que CITE-seq y el “hashing” celular se realizaran simultáneamente, pero para generar genotecas de secuenciación separadas. Específicamente, los HTO contienen una sonda de amplificación diferente a las etiquetas derivadas de anticuerpos CIE-seq estándar (ADT) de los inventores. Esto permite que las genotecas HTO, ADT y scRNA-seq se amplifiquen y agrupen independientemente en las cantidades deseadas. En particular, los inventores han observado previamente una recuperación sólida de las señales de anticuerpos de epítomos altamente expresados debido a su número de copias extremadamente alto. Esto contrasta con los extensos niveles de “abandono” observados para los datos de scRNA-seq, y sugiere que se pueden recuperar fielmente los HTO de cada célula individual, permitiendo la asignación a muestra de origen con alta fidelidad.

Para hacer referencia a la estrategia de los inventores y demostrar su utilidad, se obtuvieron PBMC de ocho donantes humanos separados (denominados donantes desde la A hasta la H), y cada muestra se tiñó independientemente con uno de los grupos de anticuerpos conjugados con HTO de los inventores, mientras se realiza simultáneamente un experimento de titulación con un grupo de siete marcadores inmunofenotípicos para CITE-seq. Posteriormente, se agruparon todas las células en una proporción igual, junto con un número igual de células HEK-293T sin teñir (y células NIH-3T3 de ratón al 3 %) como controles negativos, y la agrupación se analizó en un solo carril del sistema 10x Genomics Chromium Single Cell 3' v2. Siguiendo el enfoque de Kang y col⁷¹, se “supercargó” el instrumento 10x Genomics, cargando células a una concentración significativamente mayor con un rendimiento esperado de 20.000 células individuales y 5.000 multipletes. Basándose en las estadísticas de Poisson, 4.365 multipletes deberían representar combinaciones de células de muestras distintas y potencialmente pueden desecharse, lo que conduce a una tasa de multipletes no resuelta del 3,1 %. En particular, lograr una tasa de multipletes similar sin multiplexar

produciría ~4000 singletes. Como el coste de los sistemas comerciales basados en gotículas es fijo por ciclo, la multiplexación permite, por lo tanto, perfilar un ~400 % más de células con el mismo coste.

Se realizó la división y retrotranscripción según los protocolos estándar, utilizando solo una estrategia de amplificación posterior algo modificada para generar genotecas de transcriptoma, HTO y ADT. Se agruparon y secuenciaron en un Illumina HiSeq2500 (dos células de flujo de ciclo rápido), buscando una contribución del 90 %:5 %:5 % de las tres genotecas en los datos de secuenciación. Además, se realizó el genotipado de las ocho muestras de PBMC y las células HEK-293T con la matriz Illumina Infinium CoreExome, lo que permite utilizar tanto HTO como genotipos de muestras (evaluados mediante demuxlet⁷¹) como enfoques de demultiplexación independientes.

Al examinar la expresión por pares de dos recuentos de HTO, los inventores observaron relaciones similares a las representaciones de “mezcla de especies” (Figura 6B), lo que sugiere una exclusividad mutua de la señal de HTO entre los singletes. Extendiéndose más allá del análisis por pares, los inventores desarrollaron un modelo estadístico sencillo para clasificar cada código de barras como “positivo” o “negativo” para cada HTO. Brevemente, se modelizó la señal de “fondo” de cada HTO independientemente como una distribución binomial negativa, estimando las células de fondo según los resultados de una agrupación inicial de k-medoides de todas las lecturas de HTO. Los códigos de barras con señales HTO por encima del cuantil del 99 % para esta distribución se marcaron como “positivos”, y los códigos de barras que eran “positivos” para más de un HTO se marcaron como multipletes. Se clasificaron todos los códigos de barras donde se detectaron al menos 200 UMI de ARN, independientemente de la señal de HTO. Las clasificaciones de los inventores (visualizadas como un mapa de calor en la Figura 6C) sugirieron una identificación clara de ocho poblaciones singlete, así como grupos multiplete. Los inventores también identificaron códigos de barras con una señal de fondo insignificante para cualquiera de los HTO (etiquetados como “negativos”), que consiste principalmente (87,5 %) de células HEK y de ratón. Se retiraron todas las células HEK y de ratón de los análisis posteriores, representando los códigos de barras restantes 13.964 singletes y 2.463 multipletes identificables, en línea con las expectativas. Las clasificaciones de los inventores también fueron completamente concordantes con una integración de tSNE, calculada mediante el uso de solo las ocho señales HTO, lo que permitió la visualización clara no solo de los 8 grupos de singletes (donantes desde la A hasta la H), sino también los 28 grupos pequeños que representan todas las posibles combinaciones de dobletes (Figura 6D). Además, se observó un cambio positivo claro en la distribución del UMI/código de barras de ARN para multipletes, como se esperaba (Figura 6E), mientras que los códigos de barras negativos restantes expresaron menos UMI y pueden representar reacciones fallidas o gotículas “vacías” que solo contienen ARN ambiental. Estos resultados sugieren ampliamente que los HTO asignaron correctamente cada código de barras a su muestra original y permitieron una detección sólida de multipletes de muestras cruzadas. Realizar la agrupación transcriptómica de los singletes clasificados permitió la detección clara de nueve subpoblaciones hematopoyéticas, que se intercalaron entre los ocho donantes (Figura 6F).

B. El demultiplexado basado en genotipo valida el “hashing” celular

A continuación, los inventores compararon sus clasificaciones basadas en HTO con las obtenidas por demuxlet⁷¹. En general, se observó una notable concordancia entre las técnicas, incluso cuando se considera la mezcla de muestra precisa en las llamadas dobletes (Figura 7A). Explorando zonas sin concordancia, se identificaron 1138 códigos de barras que se clasificaron según los niveles de HTO como singletes, pero se identificaron como “ambiguas” mediante demuxlet. En particular, la intensidad de la clasificación de HTO para estos códigos de barras discordantes (representados por el número de lecturas asignadas al HTO de mayor expresión) fue idéntica a los códigos de barras que se clasificaron como singletes por ambos enfoques (Figura 7B). Sin embargo, los códigos de barras discordantes tenían recuentos reducidos de UMI de ARN (Figura 7C). Se concluye que estos códigos de barras probablemente no podrían clasificarse genéticamente con la secuenciación poco profunda de los inventores, que está por debajo de la profundidad recomendada para usar demuxlet, pero probablemente representan células individuales verdaderas basadas en las clasificaciones de HTO.

Además, también se observaron 2547 códigos de barras que recibieron clasificaciones de singlete/doblete discordantes entre las dos técnicas (Figura 7D). Cabe señalar que esto refleja una minoría de códigos de barras (en comparación con 12.676 clasificaciones concordantes), y que en estos casos discordantes es difícil discernir cuál de estos métodos es correcto. Sin embargo, cuando los inventores examinaron las distribuciones de UMI de cada grupo de clasificación, se observó que solo los códigos de barras clasificados como dobletes por ambas técnicas exhibieron un cambio positivo en la complejidad transcriptómica (Figura 7D). Esto sugiere que estas llamadas discordantes están compuestas en gran medida por singletes verdaderos, pero representan positivos falsos conservadores de ambos métodos, quizás debido a la señal de ARN ambiental o HTO. Consistente con esta interpretación, cuando se restringió el análisis de los inventores a casos donde el demuxlet interpretó códigos de barras como dobletes con >95 % de probabilidad, se observó una caída del 71 % en el número de llamadas discordantes (Figura 7E).

C. El “hashing” celular permite la optimización eficiente de los paneles de anticuerpos CITE-seq

La estrategia de multiplexación de los inventores no solo permite la agrupación en donantes, sino también el perfilado simultáneo de múltiples condiciones experimentales. Esto es ampliamente aplicable para el perfilado simultáneo de diversas perturbaciones ambientales y genéticas, pero los inventores razonaron que también se podrían optimizar eficazmente los flujos de trabajo experimentales, tales como la titulación de las concentraciones de anticuerpos para

los experimentos CITE-seq. En la citometría de flujo, los anticuerpos se analizan individualmente de forma típica en grandes series de dilución para evaluar las relaciones señal a ruido e identificar concentraciones óptimas⁷⁵. Si bien dichos experimentos tendrían un coste extremadamente elevado si se ejecutaran en carriles individuales de 10x Genomics, los inventores razonaron que estos experimentos se podrían multiplexar conjuntamente usando “*hashing*” celular.

Por lo tanto, se incubaron PBMC de diferentes donantes con una serie de dilución de concentraciones de anticuerpos que varían en tres órdenes de magnitud. Las concentraciones de anticuerpos CITE-seq se escalaron entre las diferentes muestras para mantener la cantidad total de anticuerpo y oligo consistente en cada muestra. Después de la demultiplexación de muestra, se examinaron las distribuciones ADT para todas las concentraciones de cada anticuerpo (ejemplos en las Figuras 8A-8C), y la relación señal-ruido se evaluó calculando un índice de tinción similar a las métricas comúnmente usadas para la optimización de citometría de flujo (Figura 8D).

Todos los anticuerpos mostraron solo una señal de fondo en las condiciones de control negativo y una relación señal/ruido muy débil a 0,06 µg/prueba. Se observó que la relación señal/ruido para la mayoría de los anticuerpos comenzó a saturarse dentro del intervalo de concentración de 0,5 a 1 µg/prueba, comparable a las concentraciones recomendadas para la citometría de flujo (Figura 8D). Este experimento pretendía ser una prueba de concepto; un experimento de titulación ideal usaría células del mismo donante para todas las condiciones y un mayor intervalo de concentraciones, pero claramente demuestra cómo el “*hashing*” celular se puede usar para optimizar rápida y eficientemente los flujos de trabajo experimentales.

D. Los “hashtag” celulares permiten discriminar entre células de baja calidad y ARN ambiental

Los “*hashtags*” celulares de los inventores pueden discriminar entre células individuales y dobles según la expresión clara de un solo HTO, y a continuación los inventores se preguntaron si esta característica también podría distinguir las células de baja calidad del ARN ambiental. De ser cierto, esto permitiría reducir el “corte” de UMI de los inventores (previamente establecido en 200) y permitiría la posibilidad de que ciertos códigos de barras que representen el ARN ambiental puedan expresar más UMI que algunas células individuales verdaderas. La mayoría de los flujos de trabajo establecen cortes de UMI rigurosos para excluir todo el ARN ambiental, sesgando los resultados de scRNA-seq contra células con bajo contenido de ARN y probablemente sesgando estimaciones proporcionales de tipo de célula.

De hecho, al considerar 3473 códigos de barras que contienen 50-200 UMI, los inventores recuperaron 954 singletes adicionales según clasificaciones de HTO, con 2432 códigos de barras caracterizados como negativos. Se clasificó cada código de barras como uno de las nueve poblaciones hematopoyéticas de los inventores previamente determinadas (Figura 6F), y los resultados se visualizaron en una integración de tSNE transcriptómica, calculada independientemente para los grupos “singlete” y “negativo”. Para los singletes predichos, los códigos de barras proyectados para las poblaciones B, NK, T y mieloides, que se separaron consistentemente en el tSNE, lo que sugiere que estos códigos de barras representan células individuales verdaderas (Figura 8E). Por el contrario, los códigos de barras “negativos” no se separaron según su clasificación forzada, consistente con estos códigos de barras que reflejan las mezclas de ARN ambiental que pueden combinar múltiples subpoblaciones. Por lo tanto, se concluye que al proporcionar una lectura de identidad de muestra que es independiente del transcriptoma, el “*hashing*” celular puede ayudar a recuperar células de baja calidad que de otro modo pueden ser difíciles de distinguir del ARN ambiental (Figura 8F).

E. Métodos

Genotipado de PBMC: Se obtuvieron células mononucleares de sangre periférica de Allcells (EE. UU.). El ADN genómico se purificó usando el kit All-prep (Qiagen, EE. UU.) y se genotipó usando la matriz del exoma de Infinium core 24 (Illumina, EE. UU.), según las instrucciones del fabricante.

Cultivo de células: Las células HEK293T (humanas) y NIH-3T3 (ratón) se mantuvieron según los procedimientos estándar en medio de Eagle modificado por Dulbecco (Thermo Fisher, EE. UU.) suplementado con suero de feto de bovino al 10 % (Thermo Fisher, EE. UU.) a 37 °C con 5 % de CO₂.

Conjugados de anticuerpo-oligo: Conjugados de anticuerpo-oligo dirigidos contra CD8 [clon: RPA-T8], CD45RA [clon: HI100], CD4 [clon: RPA-T4], HLA-DR [clon: L243], CD3 [clon: UCHT1], CCR7 [clon: G043H7] y PD-1 [clon: EH12.2H7] fueron proporcionadas por BioLegend (EE.UU.) y contenían 1-2 oligos conjugados por anticuerpo en promedio.

Los anticuerpos utilizados para el “*hashing*” celular se obtuvieron como reactivos purificados no conjugados de BioLegend (CD45 [clon: HI30], CD98 [clon: MEM-108], CD44 [clon: BJ18] y CD11a [clon: HI111]) y se conjugaron de forma covalente e irreversible a HTO mediante química clic con iEDDA como se ha descrito anteriormente⁴⁵. En resumen, los anticuerpos se lavaron en solución salina tamponada con borato 1X (borato 50 mM, NaCl 150 mM pH 8,5) y se concentraron a 1 mg/ml usando un filtro de centrífuga Amicon Ultra 0,5 ml MWCO de 30 kDa (Millipore). Se disolvió éster metiltetrazina-PEG4-NHS (Click Chemistry Tools, EE. UU.) en DMSO seco y se añadió en un exceso de 30 veces al anticuerpo y se dejó reaccionar durante 30 minutos a temperatura ambiente. Los grupos NHS residuales se inactivaron mediante la adición de glicina y el marcador sin reaccionar se retiró mediante filtración en centrífuga.

Los HTO con amina en 5' se pidieron a Integrated DNA Technologies (EE. UU.) y se hicieron reaccionar con un exceso de 20 veces de trans-cicloocteno-PEG4-NHS (Click Chemistry Tools, EE. UU.) en solución salina tamponada con borato 1X suplementada con 20 % de DMSO durante 30 minutos. Los grupos NHS residuales se inactivaron mediante la adición de glicina y el marcador residual se eliminó mediante desalación (Bio-Rad Micro Bio-Spin P6). Los conjugados de anticuerpo-oligo se formaron mezclando el anticuerpo marcado apropiado y el HTO e incubando a temperatura ambiente durante al menos 1 hora. Los grupos metiltetrazina residuales en el anticuerpo se inactivaron mediante la adición de ácido trans-cicloocteno-PEG4 y el oligo sin reaccionar se eliminó mediante filtración con centrífuga usando un filtro Amicon Ultra 0,5 ml MWCO de 50 kDa (Millipore, EE. UU.).

- 10 Series de titulación de anticuerpos: Para probar la concentración óptima de conjugados de anticuerpo-oligo proporcionados por BioLegend (EE. UU.) para el experimento CITE-seq, se probaron 5 µg, 3 µg, 1 µg, 0,5 µg, 0,25 µg, 0,06 µg, y 0 µg para cada conjugado. Las titulaciones se escalonaron entre los diferentes lotes para mantener la concentración total de anticuerpos y oligos consistentes entre las condiciones (véase la Tabla 4 a continuación).

Tabla 4								
	Donante A	Donante B	Donante C	Donante D	Donante E	Donante F	Donante G	Donante H
CD8	5	3	1	0,5	0,25	0,06	0	0
PD.1	3	1	0,5	0,25	0,06	0	5	0
CCR7	1	0,5	0,25	0,06	0	5	3	0
CD3	0,5	0,25	0,06	0	5	3	1	0
HLA.DR	0,25	0,06	0	5	3	1	0,5	0
CD4	0,06	0	5	3	1	0,5	0,25	0
CD45RA	0	5	3	1	0,5	0,25	0,06	0

15 Formación de las muestras: Las PBMC de diferentes donantes se tiñeron independientemente con uno de los grupos de anticuerpos conjugados con HTO de los inventores y un grupo de 7 marcadores inmunofenotípicos para CITE-seq en diferentes cantidades (ver anteriormente). Las ocho muestras de PBMC se agruparon a una concentración igual, junto con HEK293T sin marcar y 3T3 de ratón como controles negativos y se cargaron en el instrumento 10x Chromium (véase la Tabla 5 a continuación).

Tabla 5					
Muestra en el grupo	Etiquetado con Hashtag n.º	Código de barras HTO	Id. de sec. n.º	n.º de células cargadas	Rendimiento celular est.
Donante de PBMC A	A	AGGACCATCCAA	26	~3,5 K	-2,5 K
Donante de PBMC B	B	ACATGTTACCGT	27	~3,5 K	-2,5 K
Donante de PBMC C	C	AGCTTACTATCC	28	~3,5 K	-2,5 K
Donante de PBMC D	D	TCGATAATGCGA	29	~3,5 K	-2,5 K
Donante de PBMC E	E	GAGGCTGAGCTA	30	~3,5 K	-2,5 K
Donante de PBMC F	F	GTGTGACGTATT	31	~3,5 K	-2,5 K
Donante de PBMC G	G	ACTGTCTAACGG	32	~3,5 K	-2,5 K
Donante de PBMC H	H	TATCACATCGGT	33	~3,5 K	-2,5 K
HEK293T	-	-		~3,5 K	-2,5 K
NIH-3T3	-	-		-0,8 K	-0,5 K

25 CITE-seq en el instrumento 10x Genomics: Las células se “tiñeron” con anticuerpos de *hashtaging* y anticuerpos CITE-seq como se describe para CITE-seq⁴⁶. Las células “teñidas” y lavadas se cargaron en flujo de trabajo de 10x Genomics single cell 3' v2 y se procesaron según las instrucciones del fabricante hasta el paso de amplificación de

ADNc (10x Genomics, EE. UU.). Se agregaron 2 pmol de oligonucleótidos de aditivos HTO y ADT en la PCR de amplificación de ADNc y el ADNc se amplificó según el protocolo 10x Single Cell 3'v2 (10x Genomics, EE. UU.). Después de la PCR, se usó 0,6 X SPRI para separar la fracción de ADNc grande derivada de los ARNm celulares (retenidos en las perlas) de la fracción que contenía ADT y *hashtag* (en el sobrenadante). La fracción de ADNc se procesó según el protocolo de 10x Genomics Single Cell 3' v2 para generar la genoteca de transcriptoma. Se añadió un volumen de reacción adicional 1,4X de perlas SPRI a la fracción ADT/*hashtag* para llevar la relación hasta 2,0X. Las perlas se lavaron con etanol al 80 %, se eluyeron en agua, y se realizó una ronda adicional de 2,0X SPRI para eliminar el exceso de oligonucleótidos monocatenarios de la amplificación de ADNc. Después de la elución final, se establecieron PCR independientes para generar la genoteca de ADT CITE-seq (cebadores SI-PCR y RPI-x), y la genoteca de *hashtag* (SI-PCR y D7xx_s). Se puede encontrar un protocolo detallado, actualizado regularmente punto por punto para CITE-seq, *hashtagging* celular y actualizaciones futuras en la dirección www.cie-seq.com.

Tabla 6		
Nombre del oligo	Secuencia	Id. de sec. n.º:
<i>hashtag</i> oligo: /5AmMC12/	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTn nnnnnnnnnnnnBAAAAAAAAAAAAAAAAAAAAA AAAAAAA	34
Aditivo HTO:	GTGACTGGAGTTCAGACGTGTGC*T*C	35
Aditivo ADT	CCTTGGCACCCGAGAATT*C*C	36
SI-PCR	AATGATACGGCGACCACCGAGATCTACACTCTTTC CCTACACGACGC*T*C	37
RPI-x	CAAGCAGAAGACGGCATACGAGATnnnnnnnnGTGA CTGGAGTTCCTTGGCACCCGAGAATTC*C*A	38
D7xx_s	CAAGCAGAAGACGGCATACGAGATnnnnnnnnGTGA CTGGAGTTCAGACGTGT*G*C	39
n: Secuencia de código de barras o índice y puede ser A, T, G o C		
B: T, G, C, no A		
*: Enlace de fosforotioato		

Procesamiento de datos de células individuales: Los archivos Fastq de las genotecas 10x con cuatro códigos de barras distintos se agruparon y se procesaron usando el proceso estándar de Drop-seq (Drop-seq herramientas v1.0, McCarroll Lab). Las lecturas se alinearon con la referencia concatenada de hg19-mm10, y se incluyeron los primeros 50.000 códigos de barras de células en la matriz de expresión digital sin procesar como salida de las herramientas de Drop-seq. Para la cuantificación de ADT y HTO, se implementó el flujo de cuantificación de etiquetas de los inventores desarrollado anteriormente⁴⁶ como una secuencia de comandos de python, disponible en <https://github.com/Hoohm/CITE-seq-Count>, y ejecutado con parámetros predeterminados (distancia máxima de hamming de 1).

Demultiplexación con datos de genotipado usando demuxlet: Primero se generó un archivo VCF que contenía el genotipo individual (GT) de la salida de matriz del exoma del núcleo de Infinium 24, usando las herramientas de línea de comandos PLINK (versión 1.07). Este archivo VCF (que contenía información de genotipo de los 8 donantes de PBMC, así como de también las células HEK), y el archivo de bam etiquetado del flujo de Drop-seq se usaron como entradas para el demuxlet⁷¹, con parámetros predeterminados.

Procesamiento de datos de ARN de células individuales: Se realizó la normalización y análisis posterior de los datos de ARN usando el paquete Seurat R (versión 2.1, Satija Lab) que permite el procesamiento integrado de conjuntos de datos de células individuales multimodales (ARN, ADT, HTO)^{78,79}. Se colapsó la matriz de expresión de ARN de especies conjuntas solamente para incluir los 100 genes de ratón con expresión más alta (junto con todos los genes humanos) mediante el uso de la función `CollapseSpeciesExpressionMatrix`.

Primero se consideró un conjunto de 22.119 códigos de barras donde de inventores detectaron al menos 200 UMI en los datos del transcriptoma. Dado que las células HEK y 3T3 no se marcaron con HTO, se identificaron estas células basándose en sus transcriptomas. Se realizó un preagrupamiento de baja resolución realizando PCA para los 500 genes de expresión más alta, seguido de agrupación de Louvain-Jaccard en una matriz de distancia basada en los primeros cinco componentes principales^{58,80,81}. Basándose en esta agrupación, se identificaron 248 células 3T3 y 3401 células HEK, representando el resto PBMC.

Como prueba separada de la identidad de HEK, se examinó el genotipo demuxlet para posibles células HEK. Se observaron 1.668 códigos de barras clasificados como HEK por el algoritmo demuxlet, pero cuyos transcriptomas se agruparon con PBMC. Estas células expresaron diez veces menos UMI en comparación con las células HEK clasificadas transcriptómicamente, y no expresaron transcritos específicos de HEK (es decir, NGFRAP1), ambos consistentes con una identidad de PBMC. Por lo tanto, se excluyeron estos códigos de barras de todos los análisis adicionales.

Clasificación de códigos de barras basados en los niveles de HTO: Los recuentos de HTO sin procesar se normalizaron mediante el uso de la transformación de la relación log (CLR) centrada, donde los recuentos se dividieron por la media geométrica de un HTO a través de las células, y se transformaron utilizando el logaritmo:

$$x_i' = \log \frac{x_i}{(\prod_{i=1}^n x_i)^{1/n}}$$

Aquí x_i denota el recuento para un HTO especificado en la célula i , n es el número total de células, $\log$ denota el logaritmo natural. El análisis por parejas de recuentos de HTO tanto normalizados como sin procesar (Figura 6B) reveló relaciones mutuamente exclusivas, aunque determinar los cortes exactos para las señales positivas y negativas requeriría más análisis. Los inventores razonaron que si se pudiera determinar una distribución de fondo para cada HTO basándose en células “negativas”, los valores atípicos de esta distribución representarían señales positivas.

Para ayudar en la identificación no supervisada de células “negativas”, se realizó un agrupamiento inicial de k -medoides para todas las células según los datos de HTO normalizados. Se estableció $k=9$, y se observó (como se esperaba) que ocho de las agrupaciones estaban muy enriquecidas para la expresión de un HTO particular, mientras que la novena agrupación estaba muy enriquecida para células con baja expresión de todos los HTO. Esto representa una solución inicial al problema de demultiplexación que sugiere poblaciones análogas de células “positivas” y “negativas” para el análisis estadístico.

Después de la agrupación, se realizó el siguiente procedimiento independientemente para cada uno de los 8 HTO. Se identificó el grupo de k -medoides con la expresión de HTO promedio más alta y se excluyeron estas células. A continuación se ajusta una distribución binomial negativa a los valores de HTO restantes, después de excluir además los valores más altos del 0,5 % como valores atípicos potenciales. Se calculó el cuantil $q = 0,99$ de la distribución ajustada y se umbralizó cada célula en el conjunto de datos basándose en este valor específico de HTO.

Se usó este procedimiento para determinar una “clasificación HTO” para cada código de barras. Los códigos de barras que eran positivos para un solo HTO se clasificaron como singletes. Los códigos de barras que eran positivos para dos o más HTO se clasificaron como dobletes y se asignaron los ID de muestra en función de sus dos HTO de mayor expresión. Los códigos de barras que fueron negativos para los ocho HTO se clasificaron como “negativos”.

Los inventores esperan que los códigos de barras clasificados como “singletes” representen células individuales, ya que se detectó un único HTO. Sin embargo, también podrían representar dobletes de una PBMC con una célula HEK o 3T3, ya que las últimas dos poblaciones no se marcaron y representan controles negativos. De hecho, cuando se analizó la “clasificación HTO” de células que se anotaron transcriptómicamente como células HEK o 3T3, se encontró que el 73,4 % se anotó como “negativo”, mientras que el 29,2 % se anotó como singletes, en completa concordancia con las relaciones esperadas en el experimento 10x “supercargado” de los inventores. Estas células aparecen en el mapa de calor de la Figura 6C, pero todas las células HEK y 3T3 se excluyeron del análisis posterior.

Para la visualización bidimensional de los niveles de HTO (Figura 1D), se usaron distancias euclidianas calculadas a partir de los datos HTO normalizados como entradas para tSNE. Las células se colorean basándose en su clasificación de HTO como se ha descrito anteriormente. Para la visualización y la agrupación basada en datos transcriptómicos (Figura 6F), se realizó primero una PCA en los 2000 genes más variables (según lo determinado por la relación varianza/media), y se usó la matriz de distancia definida por los primeros 11 componentes principales como entrada a la agrupación de tSNE y agrupación basada en gráfico en Seurat (Figura 6E). Se anotaron las nueve agrupaciones basadas en marcadores canónicos para poblaciones hematopoyéticas conocidas.

Comparación con demuxlet: Las clasificaciones de demuxlet se marcaron como singletes (SNG), dobletes (DBL) o ambiguos (AMB) según la columna BEST del archivo de salida *.best. En la Figura 7E se ha trazado la probabilidad posterior de una asignación doblete, según la columna PRB.DBL del mismo archivo.

Cálculo del índice de tinción para las titulaciones de anticuerpo: Para evaluar la eficiencia de tinción óptima para los experimentos CITE-seq, se consideraron los niveles de ADT para las células en un intervalo de concentraciones de anticuerpos, multiplexados en una serie de titulación. Los niveles de ADT se normalizaron mediante el uso de una transformación CLR de recuentos sin procesar, usando un enfoque idéntico para la normalización de los niveles de HTO como se ha descrito anteriormente.

Después de la normalización, se calculó un índice de tinción basado en enfoques estándar en citometría de flujo, que examinan la diferencia entre medianas de picos positivos y negativos, dividida por la dispersión (es decir, dos veces la desviación absoluta media) del pico negativo.

5

$$SI = \frac{Pos_{0.5} - Neg_{0.5}}{* \text{mad} (Neg)}$$

Para evitar la clasificación manual de picos positivos y negativos, se implementó un procedimiento automatizado que puede escalar a múltiples anticuerpos y concentraciones. Para aproximar el pico negativo, se aprovecharon células de control sin tefir (Donante H). Para aproximar el pico positivo, se agruparon los datos ADT en cada experimento de titulación (desde el Donante A hasta el Donante G). Para realizar la agrupación, se calculó una matriz de distancia euclidiana entre las células basándose en los niveles de ADT normalizados, y esto se usó como la entrada de la función FindClusters en Seurat con parámetros predeterminados. Se examinó los resultados para identificar el grupo con la señal de ADT enriquecida al máximo, y se refiere a la distribución de los niveles de ADT dentro de esta agrupación como el pico positivo.

15

Discriminación de células de baja calidad del ARN ambiente: Se realizó la clasificación de HTO de códigos de barras de baja calidad (que expresan entre 50 y 200 UMI), usando los umbrales de HTO determinados previamente. Para cada código de barras, se clasificó su expresión como una de las nueve poblaciones hematopoyéticas previamente determinadas de los inventores usando bosques aleatorios, como se implementó en el paquete de intervalo en R27. Primero se entrenó un clasificador en las 13.757 PBMC, usando los 2000 genes más variables como entrada y sus identidades de agrupamiento como marcadores de entrenamiento. A continuación, se aplicó este clasificador a cada uno de los códigos de barras de baja calidad. Se observa que este clasificador garantiza devolver un resultado para cada código de barras.

25

Este proceso descrito en el Ejemplo 10 se usó para enfoques basados en gotículas, pero también es aplicable a enfoques basados en micropocillos.

30

El *hashtaging* combinado para grupos divididos se puede usar para aumentar el número de códigos de barras y, por lo tanto, aumentar la capacidad de detección de doblete. El enfoque *hashtaging* es inherente en enfoques de código de barras in situ (SPLIT-seq, sci-ARNseq) si la primera ronda de código de barras define diferentes condiciones o muestras. En contraste con el demuxlet, este enfoque se puede usar para multiplexar muestras del mismo genotipo. No es necesario realizar el genotipado en la muestra. Este proceso puede extenderse a núcleos de código de barras.

35

40

Aquí se introduce un nuevo método para la multiplexación de scRNA-seq, donde las células se etiquetan con “*hashtag*” específicas de muestra para desmultiplexación posterior y la detección de dobletes. El enfoque de los inventores es complementario a estrategias pioneras de multiplexación genética, teniendo cada una ventajas únicas. La multiplexación genética no utiliza códigos de barras exógenos y, por lo tanto, no requiere alteraciones en los flujos de trabajo existentes antes o después de la agrupación de muestras. Por el contrario, el “*hashtaging*” celular requiere incubación con anticuerpos contra proteínas superficiales expresadas de forma ubicua, pero puede multiplexar muestras con el mismo genotipo. Ambos métodos aumentan ligeramente los costes de secuenciación posterior, debido a una mayor profundidad o longitud de lectura necesaria para identificar los SNP (enfoques genéticos) o la secuenciación de genotecas de HTO (“*hashing*” celular; aproximadamente el 5 % de los costes de secuenciación del transcriptoma). Se cree que los investigadores se beneficiarán de ambos enfoques, lo que permite la multiplexación para una amplia gama de diseños experimentales. En particular, se prevé que el método de los inventores será más útil cuando se procesen muestras genéticamente idénticas sometidas a diversas perturbaciones (o condiciones/optimizaciones experimentales, como en el experimento de titulación de los inventores), o para reducir la tasa de doblete cuando se analizan células de una sola muestra.

45

50

55

60

Al permitir la identificación sólida de multipletes de células, tanto el “*hashing*” celular como la multiplexación genética permiten la “supercarga” de las plataformas de scRNA-seq. Esto se demuestra en el contexto del sistema de 10x Genomics Chromium, pero este beneficio se aplica a cualquier tecnología de células individuales que se basa en la carga de Poisson para el aislamiento de las células. Por lo tanto, los ahorros de coste por célula para la preparación de la genoteca pueden ser significativos, acercándose a un orden de magnitud a medida que aumenta el número de muestras multiplexadas. En particular, el “*hashing*” celular permite incluso que una muestra individual se multiplexe en alto grado, ya que las células pueden dividirse en un número arbitrario de conjuntos. Como se describe claramente en Kang y col⁷¹, los ahorros en la preparación de la genoteca están parcialmente desplazados por las lecturas que se originan en multipletes, que deben secuenciarse y descartarse. Aun así, como los costes de secuenciación siguen disminuyendo, y los diseños experimentales buscan minimizar los efectos de lote impulsados por tecnología, la multiplexación debería facilitar la generación de grandes conjuntos de datos de scRNA-seq y CITE-seq. La detección informativa de los multipletes basados en datos transcriptómicos también sigue siendo un desafío importante para el campo, por ejemplo, para identificar dobletes que se originan en dos células dentro de la misma muestra.

En el estudio actual de los inventores, se ha usado un grupo de anticuerpos dirigidos contra proteínas de la superficie de linfocitos expresados en alto grado y ubicuamente como vehículo para los HTO de los inventores. Esta estrategia dirigida a mitigar la posibilidad de que la variación aleatoria o de tipo celular en la expresión de cualquier marcador introduciría sesgo en la recuperación de HTO. Más adelante, se espera que un grupo más universal de anticuerpos dirigidos contra marcadores expresados de forma ubicua se use como un reactivo de “*hashing*” celular universal para estudios más allá del sistema hematopoyético. Con el interés creciente en la secuenciación de núcleos individuales⁷⁶, un conjunto adicional de reactivos de “*hashing*” dirigidos contra proteínas nucleares generalizaría aún más este enfoque. Más allá de las interacciones anticuerpo/epítipo, célula o núcleo, que incluyen otras interacciones proteína:proteína, aptámeros⁷⁷ o conjugación química directa de oligos a células o núcleos. Estas mejoras permitirán además que las estrategias de multiplexación se generalicen a otros tipos de experimentos independientemente de la especie, el tejido o la tecnología.

La siguiente información de la Tabla 7 se proporciona para secuencias que contienen texto libre bajo el identificador numérico<223>.

Tabla 7 (Listado de secuencias sin texto)	
Id. de sec. n.º:	Texto libre bajo <223>
1	<221> misc_feature
	<222> (41)..(42)
	<223> n es a o g o c o t/u
2	<221> misc_feature
	<222> (41)..(42)
	<223> n es a o g o c o t/u
3	<221> misc_feature
	<222> (40)..(41)
	<223> n es a o g o c o t/u
4	<221> misc_feature
	<222> (40)..(41)
	<223> n es a o g o c o t/u
5	<223> Secuencia de oligonucleótidos sintética para el perfilado de CBMC
6	<223> Secuencia de oligonucleótidos sintética para el perfilado de CMBC
7	<223> Secuencia de oligonucleótidos sintética para el perfilado de CMBC
8	<223> Secuencia de oligonucleótidos sintética para el perfilado de CMBC
9	<223> Secuencia de oligonucleótidos sintética para el perfilado de CMBC
10	<223> Secuencia de oligonucleótidos sintética para el perfilado de CBMC
11	<223> Secuencia de oligonucleótidos sintética para el perfilado de CMBC
12	<223> Secuencia de oligonucleótidos sintética para el perfilado de CMBC
13	<223> Secuencia de oligonucleótidos sintética para el perfilado de CMBC
14	<223> Secuencia de oligonucleótidos sintética para el perfilado de CMBC
15	<223> Secuencia de oligonucleótidos sintética para el perfilado de CBMC
16	<223> Secuencia de oligonucleótidos sintética para el perfilado de CMBC
17	<223> Secuencia de oligonucleótidos sintética para el perfilado de CMBC
18	<223>Sonda de amplificación de la secuencia de oligonucleótidos sintética
19	<223>Sonda de amplificación de la secuencia de oligonucleótidos sintética
20	<223>Secuencia de código de barras sintética
21	<223>Secuencia de código de barras sintética
22	<223>Secuencia de código de barras sintética
23	<223>Secuencia de código de barras sintética

Tabla 7 (Listado de secuencias sin texto)	
Id. de sec. n.º:	Texto libre bajo <223>
24	<223>Secuencia de código de barras sintética
25	<223>Secuencia de código de barras sintética
26	<223>Secuencia de código de barras sintética
27	<223>Secuencia de código de barras sintética
28	<223>Secuencia de código de barras sintética
29	<223>Secuencia de código de barras sintética
30	<223>Secuencia de código de barras sintética
31	<223>Secuencia de código de barras sintética
32	<223>Secuencia de código de barras sintética
33	<223>Secuencia de código de barras sintética
34	<223>Secuencia de oligonucleótido sintética hashtag
35	<223>Secuencia de oligonucleótidos sintética de un aditivo hashtag
	<221> misc_feature
	<222> (23)..(24)
	<223>Bases modificadas por la presencia de enlace fosforotioato
36	<223>Secuencia de oligonucleótido de aditivo ADT sintética
	<220>
	<221> misc_feature
	<222> (18)..(19)
	<223>Bases modificadas por enlace fosforotioato
37	<223>Secuencia de oligonucleótido de PCR sintética
	<220>
	<221> misc_feature
	<222> (47)..(48)
	<223>Bases modificadas por enlace fosforotioato
38	<223>Secuencia de oligonucleótido sintética
	<220>
	<221> misc_feature
	<222> (25)..(32)
	<223> n es A, T, G, o C de una secuencia de código de barras o índice
	<220>
	<221> misc_feature
	<222> (63)..(64)
	<223>Bases modificadas por enlace fosforotioato
39	<223>Secuencias sintéticas que forman una genoteca de cebadores hashtag
	<220>
	<221> misc_feature
	<222> (25)..(32)
	<223> n es A o C o T o G que forman una secuencia de código de barras o índice
	<220>

Tabla 7 (Listado de secuencias sin texto)	
Id. de sec. n.º:	Texto libre bajo <223>
	<221> misc_feature
	<222> (53)..(54)
	<223>Bases modificadas por enlace fosforotioato

Referencias

1. Macosko, E. Z. y col. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. CELL 161, 1202-1214 (2015).
2. Klein, A. M. y col. Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells. CELL 161, 1187-1201 (2015).
3. Zheng, G. X. Y. y col. Massively parallel digital transcriptional profiling of single cells. bioRxiv 1-46 (Cold Spring Harbor Labs Journals, 2016). doi:10.1101/065912; also, Nat. Commun. 8, 1-12 (2017); doi: 10.1038/ncomms14049 (2017).
4. Schwanhauser, B. y col. Global quantification of mammalian gene expression control. Nature 473, 337- 342 (2011).
5. Grün, D. y col. Conservation of mRNA and Protein Expression during Development of C. elegans. Cell Reports 6, 565-577 (2014).
6. Stoeckius, M. y col. Global characterization of the oocyte-to-embryo transition in Caenorhabditis elegans uncovers a novel mRNA clearance mechanism. The EMBO Journal 33, 1751-1766 (2014).
7. Pontén, F. y col. A global view of protein expression in human cells, tissues, and organs. Mol Syst Biol 5, 337 (2009).
8. Paul, F. y col. Transcriptional Heterogeneity and Lineage Commitment in Myeloid Progenitors. CELL 163, 1663-1677 (2015).
9. Wilson, N. K. y col. Combined Single-Cell Functional and Gene Expression Analysis Resolves Heterogeneity within Stem Cell Populations. CELL STEM CELL 16, 712-724 (2015).
10. Ståhlberg, A. y col. Quantitative PCR analysis of DNA, RNAs, and proteins in the same single cell. Clinical Chemistry 58, 1682-1691 (2012).
11. Genshaft, A. S. y col. Multiplexed, targeted profiling of single-cell proteomes and transcriptomes in a single reaction. Genome Biol. 17:188 (2016). doi:10.1186/s13059-016-1045-6
12. Albayrak, C. y col. Digital Quantification of Proteins and mRNA in Single Mammalian Cells. Molecular Cell 61, 914-924 (2016).
13. Darmanis, S. y col. Simultaneous Multiplexed Measurement of RNA and Proteins in Single Cells. CellReports 14, 380-389 (2016).
14. Frei, A. P. y col. Highly multiplexed simultaneous detection of RNAs and proteins in single cells. Nature Methods 13, 269-275 (2016).
15. Sano, T., y col. Immuno-PCR: very sensitive antigen detection by means of specific antibody-DNA conjugates. SCIENCE-NEW YORK THEN ... (1992).
16. Gullberg, M. y col. A sense of closeness: protein detection by proximity ligation. Current Opinion in Biotechnology 14, 82-86 (2003).
17. Chattopadhyay, P. K. & Roederer, M. Cytometry: Today's technology and tomorrow's horizons. Methods 57, 251-258 (2012).
18. Bendall, S. C. & Nolan, G. P. From single cells to deep phenotypes in cancer. Nat Biotechnol 1-9 (2012). doi:10.1038/nbt.2283

19. Adler, M., y col. Sensitivity by combination: Immuno-PCR and related technologies. *Analyst* 133, 702-18 (2008).
20. Cao, Junyue, y col. Comprehensive single cell transcriptional profiling of a multicellular organism by combinatorial indexing. *Sci.*, 357(6352):661-667 (2017).
21. Bendall, S. C. & Nolan, G. P. From single cells to deep phenotypes in cancer. *Nat Biotechnol* 1-9 (2012).
22. Baumgarth, N., Roederer, M. A practical approach to multicolor flow cytometry for immunophenotyping. *J Immunol Methods* 243, 77-97 (2000).
23. Mortazavi y col. Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nature Methods* 5, 621 - 628 (2008).
24. Hermanson, G.T. *Bioconjugation Techniques*. 2nd Edition. Academic Press, San Diego, CA (2008).
25. Lizardi, P.M., y col. Mutation detection and single-molecule counting using isothermal rolling-circle amplification. *Nat Genet.* 1998;19:225-232.
26. Assarsson, E., y col. Homogenous 96-plex PEA immunoassay exhibiting high sensitivity, specificity, and excellent scalability. *PLOS ONE*. 2014;9:e95192.
27. Fakruddin, MD, y col. "Nucleic acid amplification: Alternative methods of polymerase chain reaction." *Journal of Pharmacy and Bioallied Sciences* 5.4 (2013): 245.
28. Nimse, SB y col. Immobilization techniques for microarray: challenges and applications. *Sensors* 14.12 (2014): 22208-22229.
29. Heise, C. and Bier, FF. Immobilization of DNA on microarrays. *Immobilization of DNA on Chips II*. Springer Berlin Heidelberg, 2005. 1-25.
30. Rosenberg, Alexander B., y col. Scaling single cell transcriptomics through split pool barcoding. *bioRxiv* (2017): 105163
31. Li, Zhenhua, y col. DNA nanostructure-based universal microarray platform for high-efficiency multiplex bioanalysis in biofluids. *ACS applied materials & interfaces* 6(20) (2014): 17944-17953
32. Zhao, Hong, y col. Cell fixation in zinc salt solution is compatible with DNA damage response detection by phospho-specific antibodies. *Cytometry Part A* 79.6 (2011): 470-476.
33. Iglesias-Ussel, Maria, Luigi Marchionni, and Fabio Romerio. Isolation of microarray-quality RNA from primary human cells after intracellular immunostaining and fluorescence-activated cell sorting. *Journal of Immunological Methods* 391.1 (2013): 22-30.
34. L.J.P. van der Maaten and G.E. Hinton. Visualizing High-Dimensional Data Using t-SNE. *Journal of Machine Learning Research* 9 (Nov 2008):2579-2605
35. Gierahn TM, y col, Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput. *Nat. Methods*, 2017 Apr, 14(4):395-398 (epub 2017 Feb. 13)
36. Crosetto, Nicola, Magda Bienko, and Alexander Van Oudenaarden. Spatially resolved transcriptomics and beyond. *Nature Reviews Genetics* 16.1 (2015): 57-66
37. Leah Cannon, Single Cell Analysis: A Mini-Report, lifesciencenetwork.com/blogs/leah-cannon/2017/03/21/single-cell-analysis-a-mini-report, 2017 March
38. Zhang, Kai, y col. Single-cell isolation by a modular single-cell pipette for RNA-sequencing. *Lab on a Chip* 16.24 (2016): 4742-4748;
39. Poulin, Jean-Francois, y col. Disentangling neural cell diversity using single-cell transcriptomics. *Nature neuroscience* 19.9 (2016): 1131-1141
40. Picelli, Simone. Single-cell RNA-sequencing: The future of genome biology is now. *RNA biology* (2016): 1-14)
41. Lai, Shujing, y col. Mapping Human Hematopoietic Hierarchy At Single Cell Resolution By Microwell-seq. *bioRxiv* (2017): 127217

42. Xin, Yurong, y col. Use of the Fluidigm C1 platform for RNA sequencing of single mouse pancreatic islet cells. *Proceedings of the National Academy of Sciences* (2016): 201602306
- 5 43. Islam, Saiful, y col. Quantitative single-cell RNA-seq with unique molecular identifiers. *Nature methods* 11.2 (2014): 163-166
44. Wu, Angela R., y col. Quantitative assessment of single-cell RNA-sequencing methods. *Nature methods* 11.1 (2014): 41-46
- 10 45. van Buggenum, MAGL y col., A covalent and cleavable antibody-DNA conjugation strategy for sensitive protein detection via immuno-PCR, *Sci. Reports*, 6:22675, DOI: 10.1038/srep22675
- 15 46. Stoeckius M, y col., Simultaneous epitope and transcriptome measurement in single cells 31 July 2017, *Nature Methods* 9, 2579-10 (2017). DOI:10.1038/NMeth.4380
47. Murphy, K., Travers, P. & Walport, M. *Janeway's Immunobiology* 7th edn (Garland Publishing, 2008).
- 20 48. Robinson, J.P. & Roederer, M., *Flow Cytometry Strikes Gold*, *Science* 350, 739-740 (2015).
49. Fan, H.C., Fu, G.K. & Fodor, S.P.A., Combinatorial labeling of single cells for gene expression cytometry, *Science* 347, 1258367 (2015).
- 25 50. Poli, A. y col., CD56bright natural killer (NK) cells: an important NK cell subset, *Immunology* 126, 458-465 (2009).
51. Ferlazzo, G. & Münz, C. J., NK Cell Compartments and Their Activation by Dendritic Cells, *Immunol.* 172, 1333-1339 (2004).
- 30 52. Wendt, K. y col., Gene and protein characteristics reflect functional diversity of CD56dim and CD56bright NK cells., *J. Leukoc. Biol.* 80, 1529-1541 (2006).
53. Shahi, P., Kim, S.C., Haliburton, J.R., Gartner, Z.J. & Abate, A.R., Abseq: Ultrahigh-throughput single cell protein profiling with droplet microfluidic barcoding, *Sci. Rep.* 7, 44447 (2017).
- 35 54. Yuan, J. & Sims, P.A. An Automated Microwell Platform for Large-Scale Single Cell RNA-Seq. *Sci. Rep.* 6, 33883 (2016).
- 55 55. Gierahn, T.M. y col. Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput, *Nat. Methods* 14, 395-398 (2017).
56. Baranauskas, A. y col. Generation and characterization of new highly thermostable and processive M-MuLV reverse transcriptase variants, *Protein Eng. Des. Sel.* 25, 657-668 (2012).
- 45 57. Breton, G., Lee, J., Liu, K. & Nussenzweig, M.C. Defining human dendritic cell progenitors by multiparametric flow cytometry, *Nat. Protoc.* 10, 1407-1422 (2015).
58. Blondel, V.D., y col. Fast unfolding of communities in large networks, *J. Stat. Mech.* 2008, P10008 (2008).
- 50 59. van der Maaten, L. *J. Mach. Learn. Res.* 15, 1-21 (2014).
60. Stoeckius, M. & Smibert, Cite-seq, *Protocol Exchange* <http://dx.doi.org/10.1038/protex.2017.068> (31 July 2017).
- 55 61. Aitchison, J., Measures of location of compositional data sets., *Math. Geol.* 21(7): 787-790 (1989).
62. Kang, H.M. y col., Multiplexing droplet-based single cell RNA-sequencing using natural genetic barcodes, *bioRxiv* 118778; doi: <https://doi.org/10.1101/118778>
- 60 63. Stubbington, M. J. T., Rozenblatt-Rosen, O., Regev, A. & Teichmann, S. A. Single-cell transcriptomics to explore the immune system in health and disease. *Science* 358, 58-63 (2017).
64. Tanay, A. & Regev, A. Scaling single-cell genomics from phenomenology to mechanism. *Nature* 541, 331-338 (2017).
- 65

65. Villani, A.-C. y col. Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science* 356, (2017).
- 5 66. Velten, L. y col. Human haematopoietic stem cell lineage commitment is a continuous process. *Nature Cell Biology* 19, 271-281 (2017).
67. Karaïskos, N. y col. The *Drosophila* embryo at single-cell transcriptome resolution. *Science* 8, ean3235-14 (2017).
- 10 68. Regev, A. y col. Science Forum: The Human Cell Atlas. *eLife* 6, e27041 (2017).
69. Stegle, O., Teichmann, S. A. & Marioni, J. C. Computational and analytical challenges in single-cell transcriptomics. *Nature Publishing Group* 16, 133-145 (2015).
- 15 70. Hicks, S. C., y col. Missing data and technical variability in single-cell RNA-sequencing experiments. *Biostatistics* (2017). doi:10.1093/biostatistics/kxx053
71. Kang, H. M. y col. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nature Biotechnology* (2017). doi:10.1038/nbt.4042
- 20 72. Tung, P.-Y. y col. Batch effects and the effective design of single-cell gene expression studies. *Scientific Reports* 7, 39921 (2017).
73. Krutzik, P. O. & Nolan, G. P. Fluorescent cell barcoding in flow cytometry allows highthroughput drug screening and signaling profiling. *Nat Meth* 3, 361-368 (2006).
- 25 74. Lai, L., Ong, R., Li, J. & Albani, S. A CD45-based barcoding approach to multiplex masscytometry (CyTOF). *Cytometry* 87, 369-374 (2015).
75. Hulspas, R. Titration of fluorochrome-conjugated antibodies for labeling cell surface markers on live cells. *Curr Protoc Cytom* Chapter 6, Unit 6.29 (2010).
- 30 76. Lake, B. B. y col. A comparative strategy for single-nucleus and single-cell transcriptomes confirms accuracy in predicted cell-type expression from nuclear RNA. *Scientific Reports* 1-8 (2017). doi:10.1038/s41598-017-04426-w
- 35 77. Delley, C. L., liu, L., Sarhan, M. F. & Abate, A. R. Combined aptamer and transcriptome sequencing of single cells. *bioRxiv* 1-10 (2017). doi:10.1101/228338
78. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology* 33, 495-502 (2015).
- 40 79. Butler, A. & Satija, R. Integrated analysis of single cell transcriptomic data across conditions, technologies, and species. *bioRxiv* (2017). doi:10.1101/164889
80. Levine, J. H. y col. Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. *Cell* 162, 184-197 (2015).
- 45 81. Shekhar, K. y col. Comprehensive Classification of Retinal Bipolar Neurons by Single- Cell Transcriptomics. *Cell* 166, 1308-1323.e30 (2016).
- 50 82. Wright, M. N. & Ziegler, A. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C and R. *Journal of Statistical Software* 77, (2017).

REIVINDICACIONES

1. Un método de alto rendimiento para caracterizar una célula mediante la detección simultánea de uno o más epítomos ubicados en o sobre la célula y el transcriptoma, comprendiendo el método

A)poner en contacto una muestra biológica que contiene células con

(i)una composición que comprende una primera construcción que comprende un primer anticuerpo o fragmento del mismo que se une específicamente a un primer epítomo ubicado en o sobre la superficie de una célula, conjugado dicho primer anticuerpo o fragmento a una primera construcción polimérica mediante un enlazador, en donde la primera construcción polimérica comprende una sonda de amplificación; una secuencia de código de barras que identifica específicamente dicho primer anticuerpo o fragmento de cualquier otro anticuerpo o fragmento que reconoce un epítomo diferente, una secuencia de identificador molecular único (UMI) opcional que se coloca adyacente al extremo 5' o 3' de la secuencia de código de barras y una secuencia de anclaje de poli A diseñada para hibridarse a una secuencia de oligonucleótido de captura que comprende una secuencia de poli T inmovilizada en una perla de microfluidos; o

(ii)una composición de (i) que comprende al menos una construcción adicional, que comprende un anticuerpo adicional o fragmento del mismo conjugado a una construcción polimérica adicional mediante un enlazador, uniéndose específicamente dicho anticuerpo adicional o fragmento del mismo a un epítomo adicional, y comprendiendo dicha construcción polimérica adicional: la sonda de amplificación de (i); una secuencia de código de barras adicional que identifica específicamente dicho anticuerpo adicional o fragmento del mismo; un UMI adicional opcional que se coloca adyacente al extremo 5' o 3' de la secuencia de código de barras adicional, y el dicho anclaje de (i), en donde dicho anticuerpo o fragmento adicional, código de barras adicional, UMI adicional y epítomo adicional difieren de los componentes correspondientes en cualquier otra construcción de la composición;

B)encapsular una única célula individual unida a una o más construcciones en una gotícula acuosa con dicha perla de microfluidos, en donde la perla está conjugada a secuencias del oligonucleótido de captura, comprendiendo cada secuencia del oligonucleótido de captura una secuencia de código de barras de perla única, un UMI opcional y una secuencia de poli T en 3'; y

C)lisar la célula individual en la gotícula, en donde los ARNm de la célula y las construcciones poliméricas liberados de los anticuerpos o fragmentos se hibridan con secuencias de poli T de las secuencias del oligonucleótido de captura; y generar a partir de las secuencias hibridadas con la perla de microfluidos (a) ADNc bicatenarios que contienen la secuencia de código de barras de la perla y los transcritos inversos del ARNm celular y (b) ADN bicatenario que contiene la secuencia de código de barras de la perla y la primera construcción polimérica.

2. El método según la reivindicación 1, que comprende además crear mediante amplificación una genoteca que comprende los ADNc de (a) y el ADN que contiene la construcción polimérica de (b); y detectar las secuencias.

3. El método según la reivindicación 2, en donde el paso de detección comprende detectar las secuencias de código de barras de la construcción para identificar si la célula individual expresa el primer epítomo.

4. El método según la reivindicación 2, en donde el paso de detección comprende determinar el nivel de expresión del primer epítomo o epítomo adicional en la célula individual mediante la detección de la cantidad de una secuencia de código de barras de la construcción normalizada por la cantidad de cualesquiera de los UMI o la cantidad media de dos o más UMI.

5. El método según la reivindicación 2 o 3, que comprende además asociar el transcriptoma o componentes del transcriptoma de la genoteca con la célula en la que se identificó el primer epítomo.

6. El método según una cualquiera de las reivindicaciones 1 a 5, en donde el paso de puesta en contacto comprende además añadir el uno o más de (i), y (ii), a la muestra biológica simultánea o secuencialmente.

7. El método según la reivindicación 1, en donde la muestra biológica es una mezcla de células obtenidas a partir de la combinación de muestras de células en donde, antes de la combinación, cada muestra se pone en contacto con una construcción específica de la muestra, comprendiendo cada construcción específica de la muestra una secuencia de código de barras específica de muestra y un anticuerpo o fragmento del mismo que se une específicamente a un epítomo ubicado en o sobre la superficie de una célula, conjugado dicho anticuerpo o fragmento con una construcción polimérica mediante un enlazador, en donde la construcción polimérica comprende una sonda de amplificación; una secuencia de identificador molecular único (UMI) opcional que se coloca adyacente al extremo 5' o 3' de la secuencia de código de barras específico de

muestra; y una secuencia de anclaje de poli A diseñada para hibridarse a una secuencia del oligonucleótido de captura que comprende una secuencia de poli T inmovilizada en la perla de microfluidos, y realizar una secuenciación de células individuales y asignar la célula a una muestra de origen basándose en la secuencia de código de barras específica de muestra.

8. El método según la reivindicación 7, en donde la secuenciación comprende el análisis de 20.000 células de la mezcla de células obtenidas por combinación de muestras de células.

9. El método según la reivindicación 1, en donde la muestra biológica se pone en contacto con más de 100 construcciones diferentes.

10. Un método de alto rendimiento para caracterizar una célula mediante la detección simultánea de uno o más epítomos ubicados en o sobre la célula y el transcriptoma, comprendiendo el método

A)obtener una mezcla de células obtenidas a partir de la combinación de muestras de células en donde, antes de la combinación, cada muestra se pone en contacto con una construcción específica de la muestra, comprendiendo cada construcción específica de la muestra una secuencia de código de barras específica de muestra y un anticuerpo o fragmento del mismo que se une específicamente a un epítipo ubicado en o sobre la superficie de una célula, conjugado dicho anticuerpo o fragmento con una construcción polimérica mediante un enlazador, en donde la construcción polimérica comprende una sonda de amplificación; una secuencia de identificador molecular único (UMI) opcional que se coloca adyacente al extremo 5' o 3' de la secuencia de código de barras específica de muestra; y una secuencia de anclaje de poli A diseñada para hibridarse a una secuencia del oligonucleótido de captura que comprende una secuencia de poli T inmovilizada en una perla de microfluidos,

B)poner en contacto la mezcla de células con:

(i)una composición que comprende una primera construcción que comprende un primer anticuerpo o fragmento del mismo que se une específicamente a un primer epítipo ubicado en o sobre la superficie de una célula, conjugado dicho primer anticuerpo o fragmento a una primera construcción polimérica mediante un enlazador, en donde la primera construcción polimérica comprende una sonda de amplificación; una secuencia de código de barras que identifica específicamente dicho primer anticuerpo o fragmento de cualquier otro anticuerpo o fragmento que reconoce un epítipo diferente, una secuencia de identificador molecular único (UMI) opcional que se coloca adyacente al extremo 5' o 3' de la secuencia de código de barras y una secuencia de anclaje de poli A diseñada para hibridarse a una secuencia de oligonucleótido de captura que comprende una secuencia de poli T inmovilizada en la perla de microfluidos; o

(ii)una composición de (i) que comprende al menos una construcción adicional, que comprende un anticuerpo adicional o fragmento del mismo conjugado a una construcción polimérica adicional mediante un enlazador, uniéndose específicamente dicho anticuerpo adicional o fragmento del mismo a un epítipo adicional, y comprendiendo dicha construcción polimérica adicional: la sonda de amplificación de (i); una secuencia de código de barras adicional que identifica específicamente dicho anticuerpo adicional o fragmento del mismo; un UMI adicional opcional que se coloca adyacente al extremo 5' o 3' de la secuencia de código de barras adicional, y el dicho anclaje de (i), en donde dicho anticuerpo o fragmento adicional, secuencia de código de barras adicional, UMI adicional y epítipo adicional difieren de los componentes correspondientes en cualquier otra construcción de la composición;

C)encapsular una única célula individual unida a una o más construcciones en una gotícula acuosa con la perla de microfluidos, en donde la perla está conjugada a secuencias del oligonucleótido de captura, comprendiendo cada secuencia del oligonucleótido de captura una secuencia de código de barras de perla única, un UMI opcional y una secuencia de poli T en 3';

D)lisar la célula individual en la gotícula, en donde los ARNm de la célula y las construcciones poliméricas liberados de los anticuerpos o fragmentos se hibridan con secuencias de poli T de las secuencias del oligonucleótido de captura; y generar a partir de las secuencias hibridadas con la perla de microfluidos (a) ADNc bicatenarios que contienen la secuencia de código de barras de la perla y transcritos inversos del ARNm celular y (b) ADN bicatenario que contiene la secuencia de código de barras de la perla y las secuencias de la construcción polimérica; y

E)realizar una secuenciación de células individuales y asignar la célula a una muestra de origen basándose en la secuencia de código de barras específica de la muestra única.

Fig. 2A

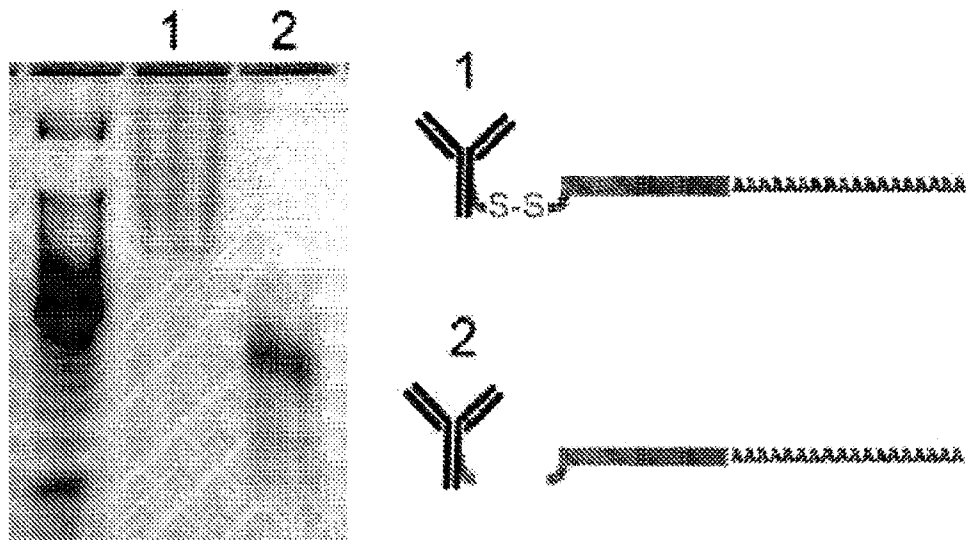


Fig. 2B

Sonda para proteínas de superficie celular ubicuas:
Integrina beta-1 (CD29)

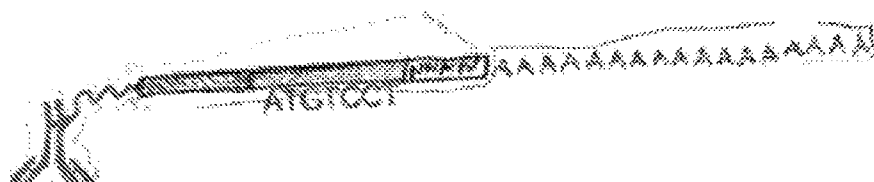
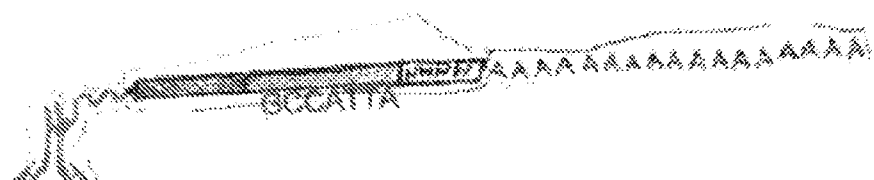
- ~  Oligo contra CD29 de ratón - Código de barras 1
- ~  Oligo contra CD29 humano - Código de barras 2

Fig. 2C

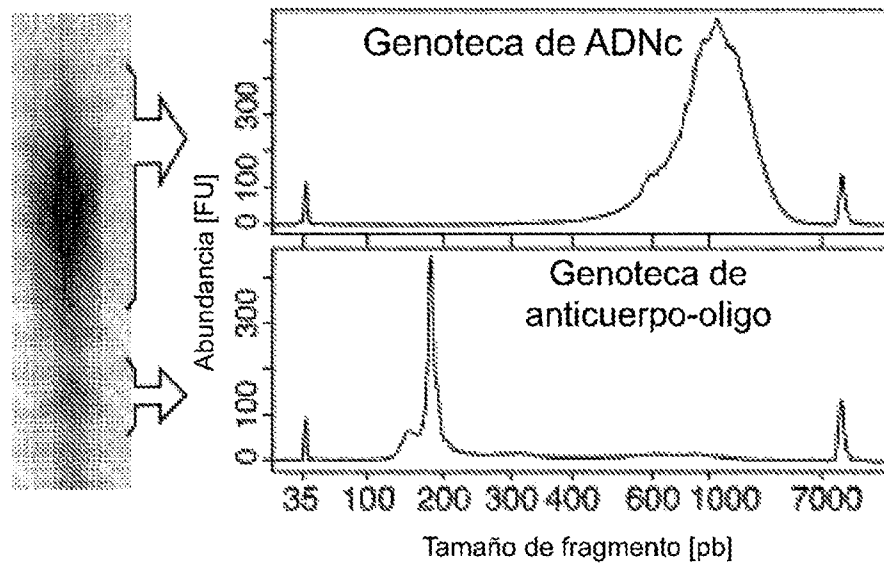


Fig. 2D

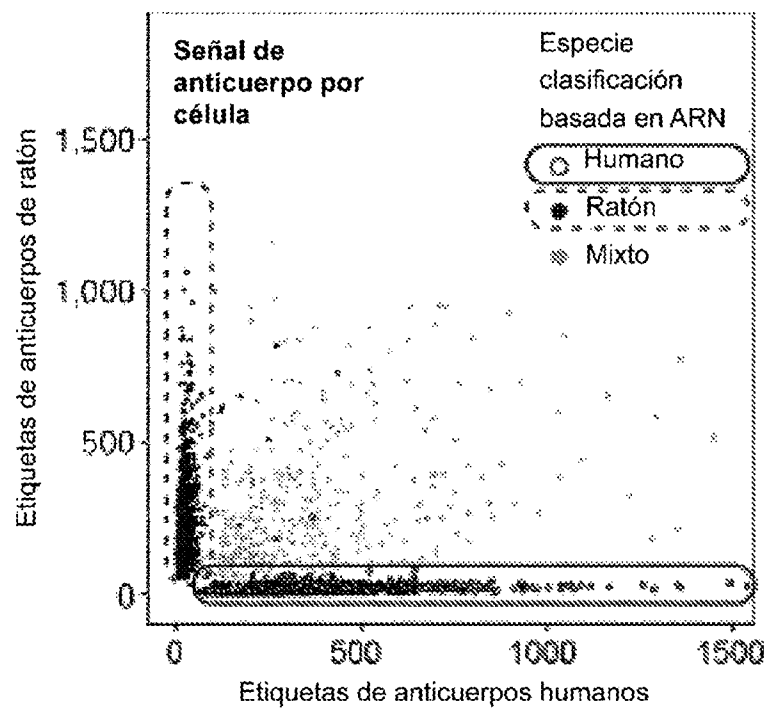


Fig. 2E

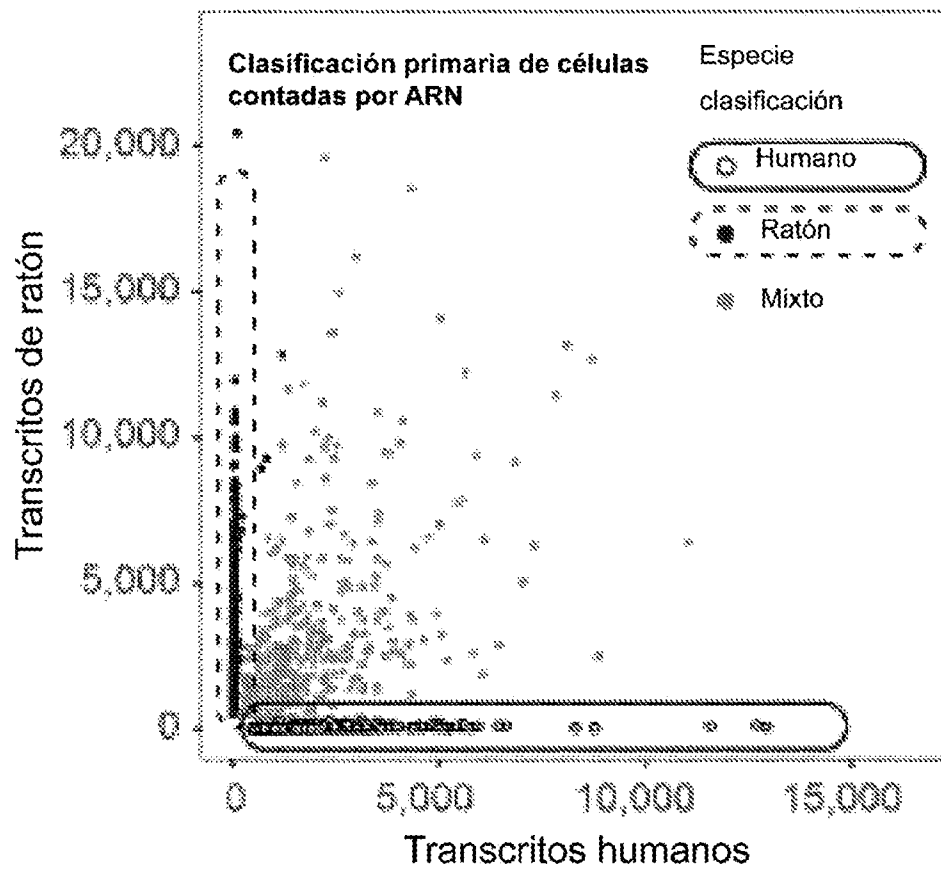


Fig. 3A

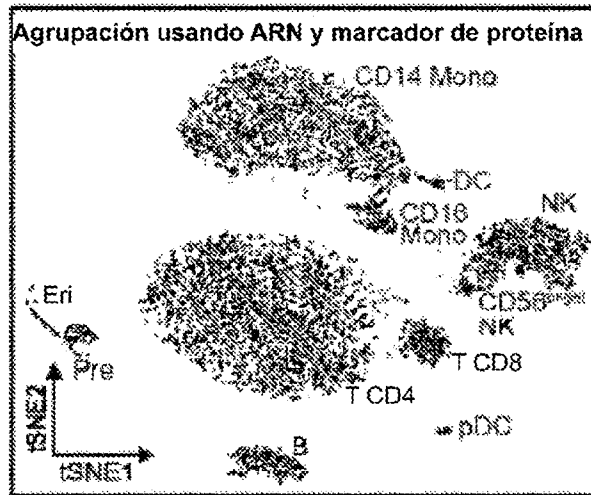


Fig. 3B

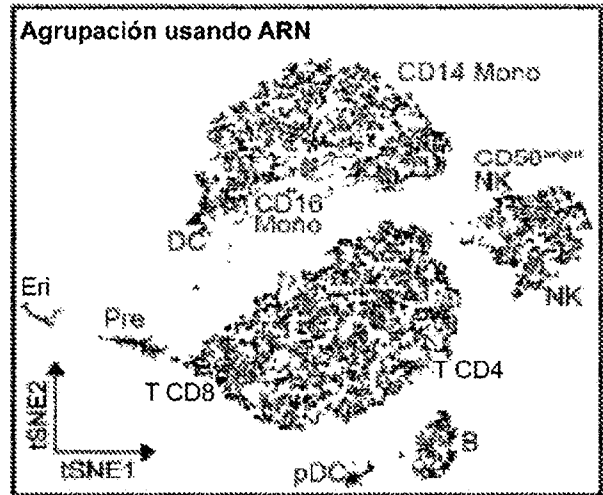


Fig. 3C

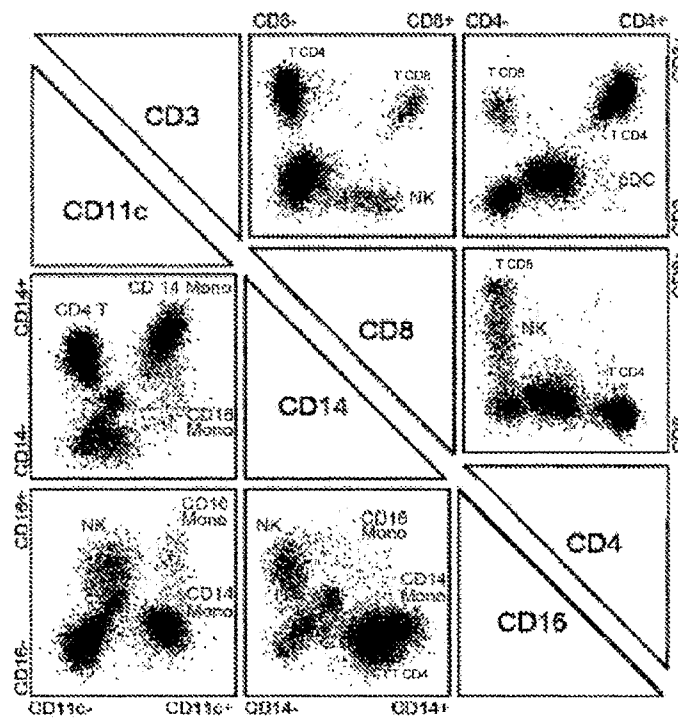


Fig. 4

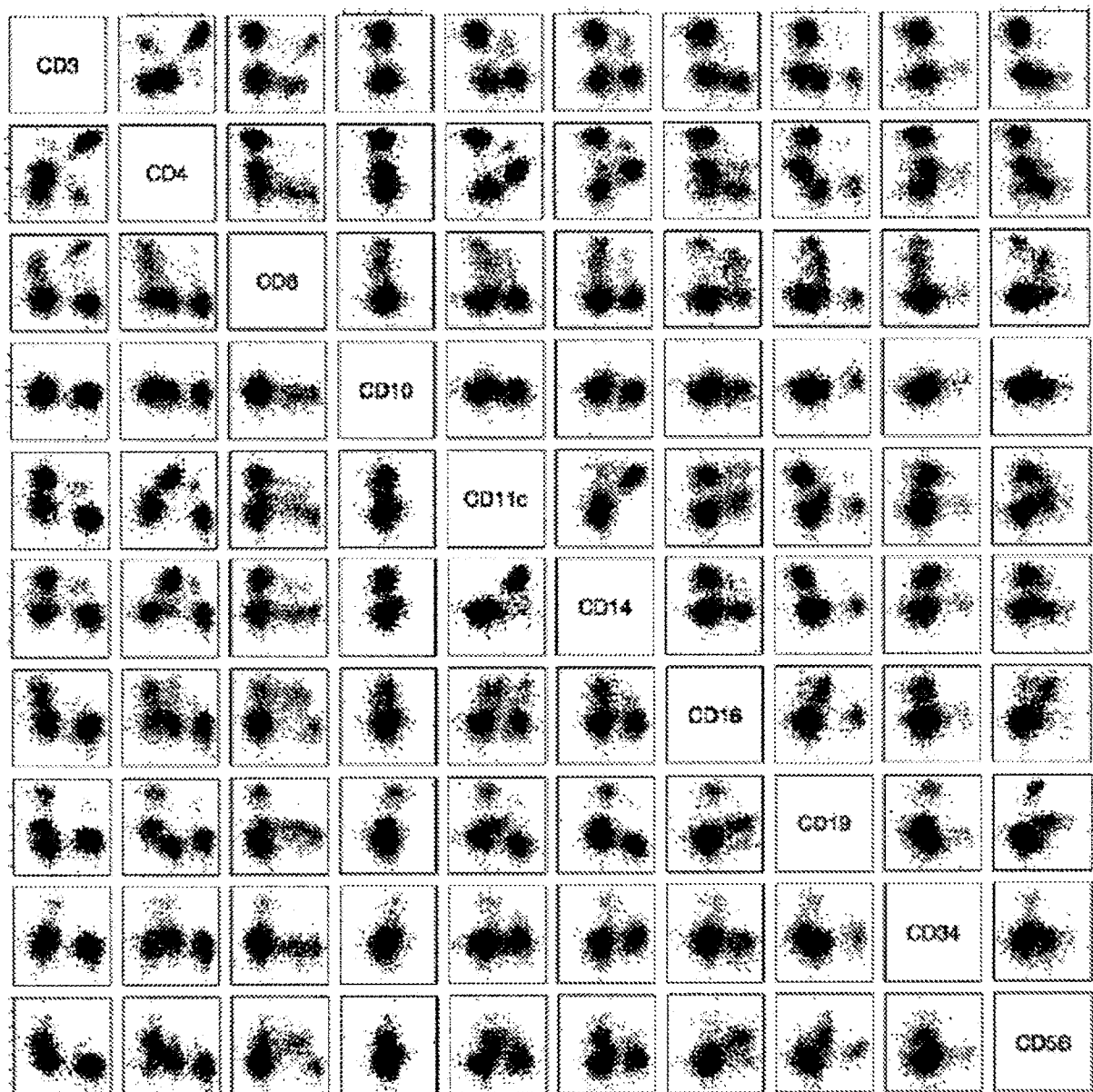


Fig. 5A

Agrupación de RNBA de ~4000 PBMC:

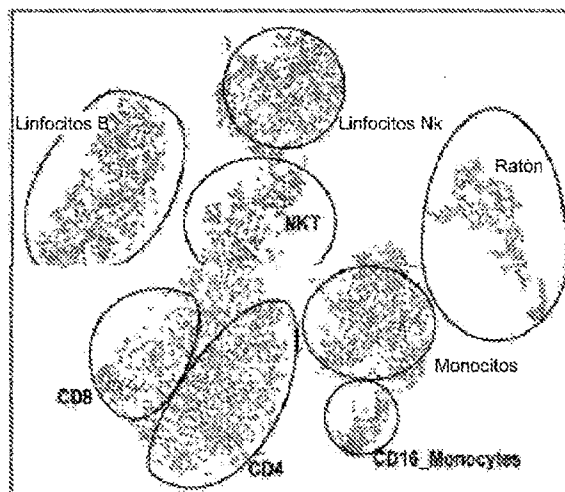
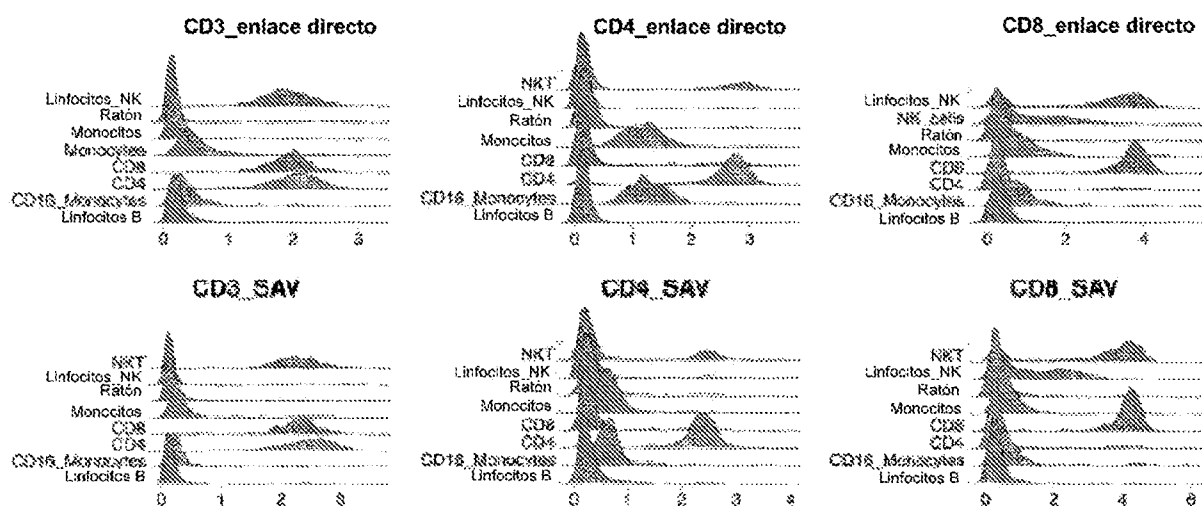


Fig. 5B

Niveles de ADT en las diferentes agrupaciones y condiciones:



Recuentos de ADT transformados por CLR

SAV = enlace estreptavidina-biotina

Enlace directo = enlace covalente mediante la química de iEDDA

Fig. 6A

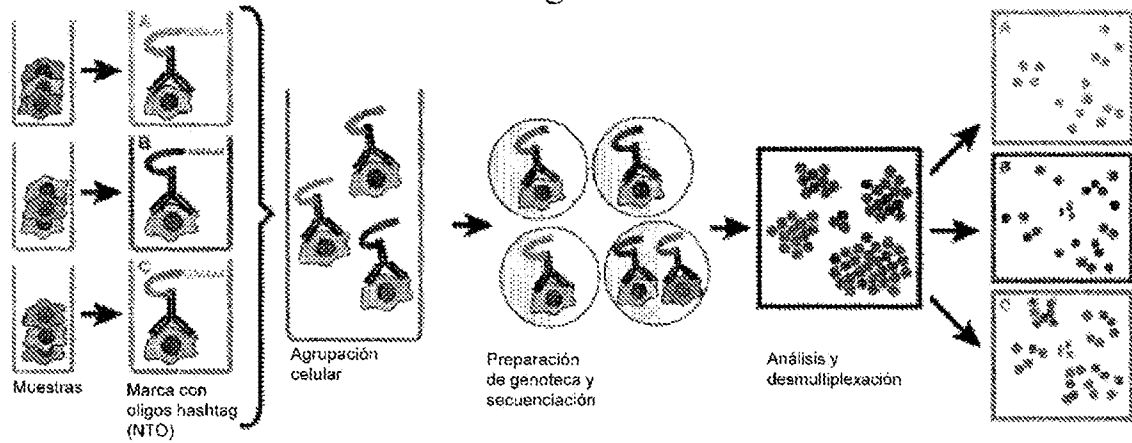


Fig. 6B

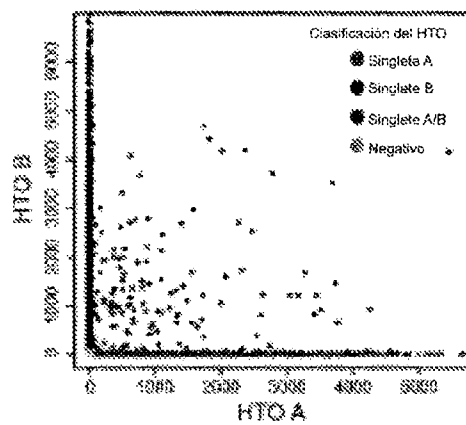


Fig. 6D

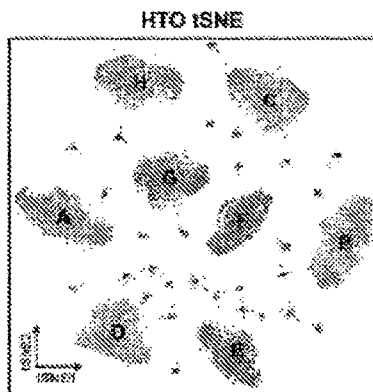


Fig. 6C

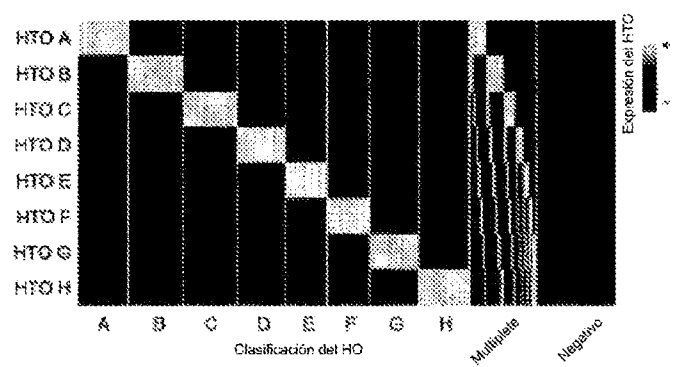


Fig. 6E

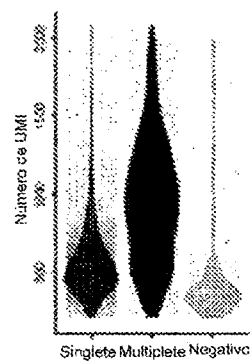


Fig. 6F

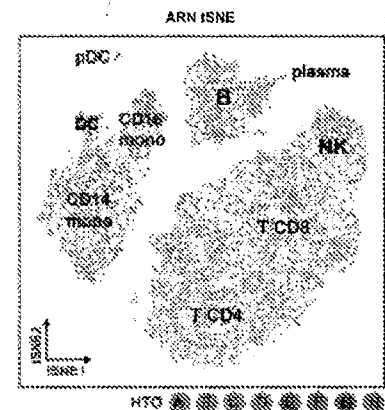


Fig. 7A

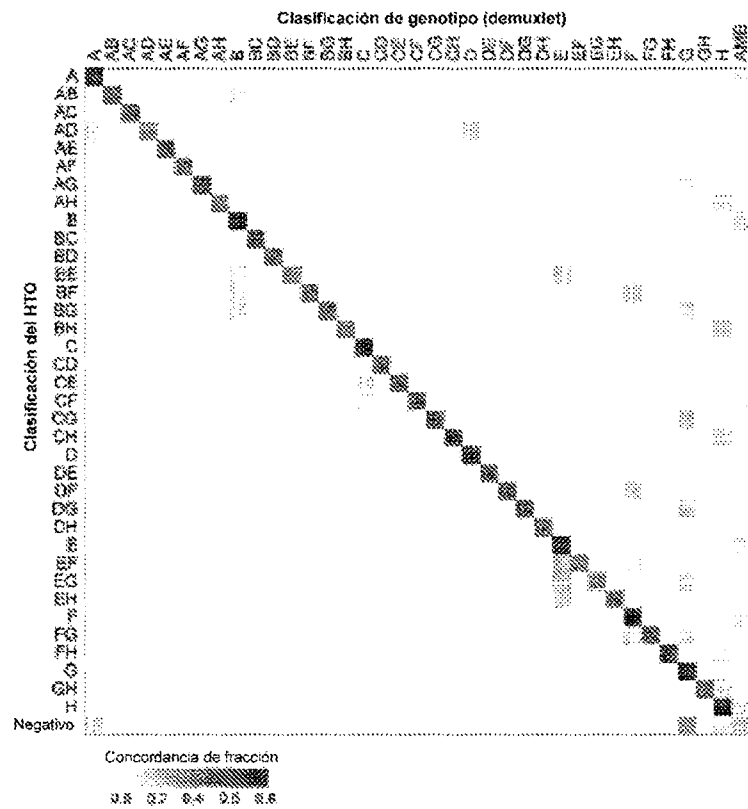


Fig. 7B

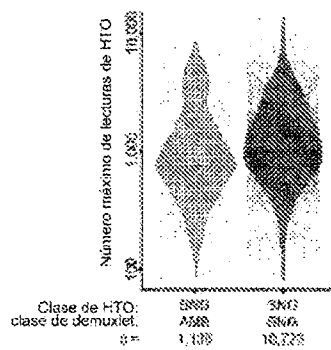


Fig. 7C

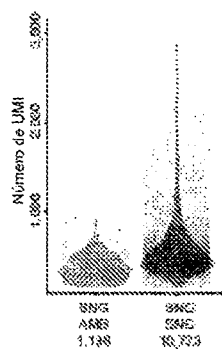


Fig. 7D

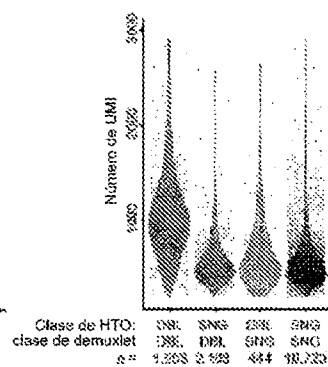


Fig. 7E

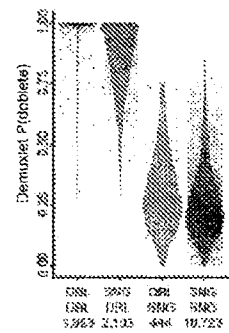


Fig. 8A

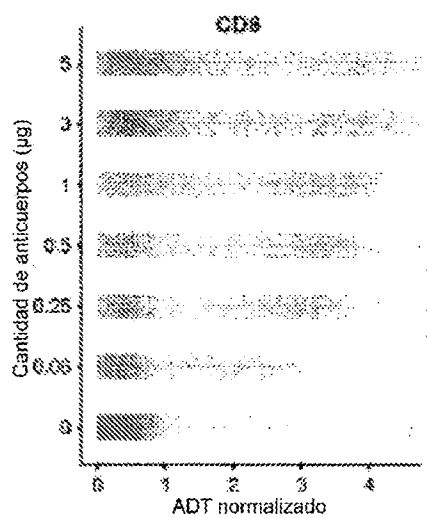


Fig. 8B

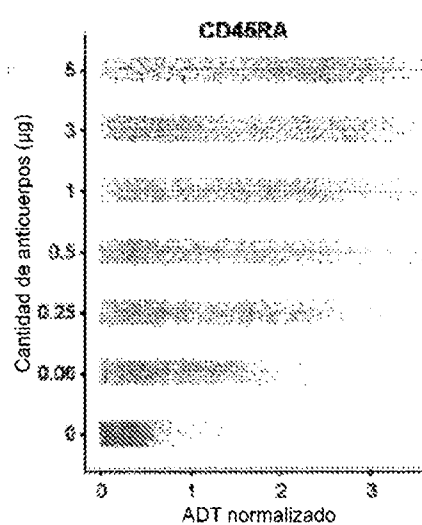


FIG. 8C

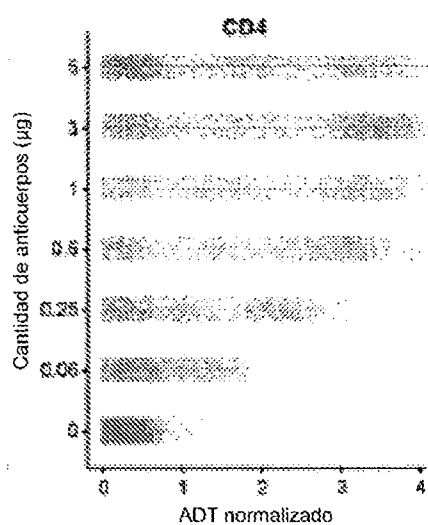


FIG. 8D

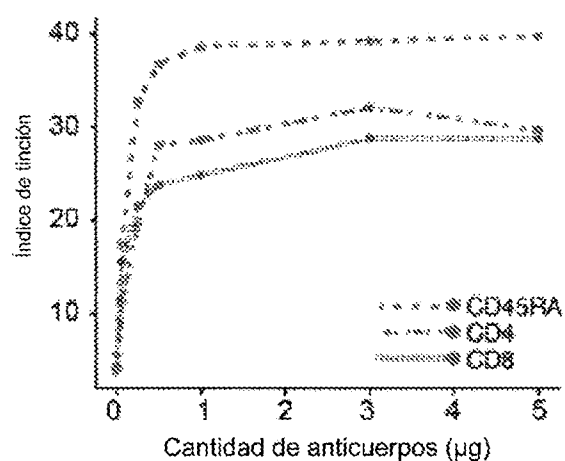


Fig. 8E

Células de bajo UMI predichas

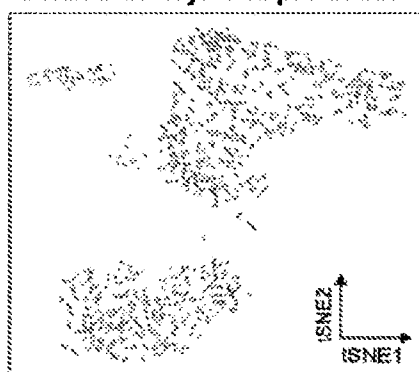
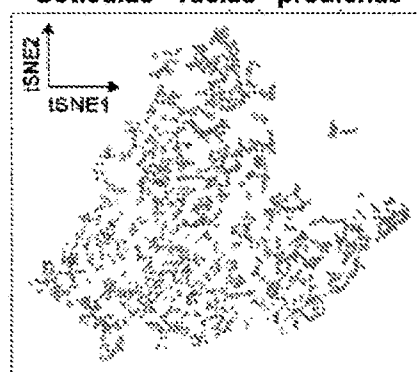


Fig. 8F

Gotículas "vacías" predichas



● T CD4 ● T CD8 ● CD14 Mono ● B ● NK