

19



OFICINA ESPAÑOLA DE
PATENTES Y MARCAS

ESPAÑA



11 Número de publicación: **2 966 028**

51 Int. Cl.:

C12N 15/10 (2006.01)
C12Q 1/6806 (2008.01)
C12Q 1/6816 (2008.01)
C12Q 1/6844 (2008.01)
C12Q 1/6813 (2008.01)
C12Q 1/6869 (2008.01)

12

TRADUCCIÓN DE PATENTE EUROPEA

T3

- 86 Fecha de presentación y número de la solicitud internacional: **04.06.2019 PCT/US2019/035422**
87 Fecha y número de publicación internacional: **12.12.2019 WO19236599**
96 Fecha de presentación y número de la solicitud europea: **04.06.2019 E 19814056 (8)**
97 Fecha y número de publicación de la concesión europea: **13.09.2023 EP 3810774**

54 Título: **Bibliotecas de transcriptomas unicelulares de alto rendimiento y métodos de preparación y de uso**

30 Prioridad:

04.06.2018 US 201862680259 P
21.03.2019 US 201962821678 P

45 Fecha de publicación y mención en BOPI de la traducción de la patente:
18.04.2024

73 Titular/es:

ILLUMINA, INC. (50.0%)
5200 Illumina Way
San Diego, CA 92122, US y
UNIVERSITY OF WASHINGTON (50.0%)

72 Inventor/es:

STEEMERS, FRANK J.;
SHENDURE, JAY;
CAO, JUNYUE;
TOME, JACOB y
GASPERINI, MOLLY

74 Agente/Representante:

ELZABURU, S.L.P

ES 2 966 028 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

Bibliotecas de transcriptomas unicelulares de alto rendimiento y métodos de preparación y de uso

Campo

- 5 Las realizaciones de la presente descripción se refieren a la secuenciación de ácidos nucleicos. En particular, las realizaciones de los métodos y composiciones proporcionadas en la presente memoria se refieren a la producción de bibliotecas de transcriptomas unicelulares indexadas y la obtención de datos de secuencia a partir de las mismas.

Antecedentes

- 10 Las células transitan a través de estados funcional y molecularmente distintos durante diversos procesos, como durante el desarrollo de un organismo multicelular y en respuesta a diferentes condiciones, como la exposición a un agente terapéutico. Caracterizar la ruta de transición del estado celular, o destino celular, es útil para comprender las vías que incluyen el desarrollo y la respuesta molecular de las células a entornos cambiantes. Por ejemplo, se pueden identificar reguladores de defectos del desarrollo y se puede lograr una mejor comprensión de cómo los agentes terapéuticos afectan a las células.

- 15 La indexación combinatoria de células individuales ("sci-") es un marco metodológico que emplea códigos de barras de grupo dividido para marcar de forma única el contenido de ácido nucleico de un gran número de células o núcleos individuales. Sin embargo, las técnicas genómicas unicelulares actuales carecen del rendimiento y la resolución necesarios para obtener una visión global de los estados moleculares y las trayectorias de un número de tipos de células que se diversifican y expanden rápidamente y que suelen estar presentes durante el desarrollo de un organismo multicelular. Las técnicas genómicas unicelulares actuales sólo capturan una instantánea del estado de una célula, por lo que no pueden proporcionar información sobre la dinámica de transición celular regulada por factores intrínsecos (p. ej., el programa intrínseco del ciclo celular de una célula) y extrínsecos (p. ej., la respuesta de una célula a un estímulo externo como un agente terapéutico).

- 20 El documento US 2003/082584 A1 y Rapolas Zillionis et al.: "Single-cell barcoding and sequencing using droplet microfluidics", NATURE PROTOCOLS, 2016, 12(1):44-73 describen métodos para preparar bibliotecas de secuenciación en gotitas que comprenden células individuales que implican una indexación específica del ARN en gotitas.

Compendio

La invención está definida por las reivindicaciones adjuntas.

- 30 En la presente memoria se proporcionan métodos para identificar la dinámica de transición del estado celular marcando el ARN recién sintetizado. Se capturan los transcriptomas de ARN tanto completos como recién sintetizados, lo que permite la caracterización de la dinámica del transcriptoma entre puntos temporales a nivel de una sola célula. También se proporcionan en la presente memoria métodos que se centran en la secuenciación unicelular de ARNm de interés, abordando así el poder limitado de la corriente para detectar cambios en la abundancia de cualquier transcrito determinado. Además, se proporcionan métodos que superan la tasa de pérdida de células y las eficiencias de reacción limitadas para dar como resultado la elaboración de perfiles de un mayor número de células individuales de lo que antes era posible.

- 40 En una realización, un método incluye proporcionar una pluralidad de núcleos o células en una primera pluralidad de compartimentos, donde cada compartimento comprende un subconjunto de núcleos o células, y marcar ARN recién sintetizado en los subconjuntos de células o núcleos obtenidos de las células. Las moléculas de ARN en cada subconjunto de núcleos o células se procesan para generar núcleos o células indexadas, donde el procesamiento incluye añadir a los ácidos nucleicos de ARN presentes en cada subconjunto de núcleos o células una primera secuencia índice específica del compartimento para dar como resultado ácidos nucleicos de ADN indexados presentes en núcleos o células indexadas, y luego combinar los núcleos o células indexadas para generar núcleos o células indexadas agrupadas.

- 45 En otra realización, un método incluye proporcionar una pluralidad de núcleos o células en una primera pluralidad de compartimentos, donde cada compartimento comprende un subconjunto de núcleos o células. Cada subconjunto se pone en contacto con la transcriptasa inversa y un cebador que se aparea con un ácido nucleico de ARN predeterminado, dando como resultado, ácidos nucleicos de ADN bicatenarios con el cebador y la secuencia de nucleótidos de ADN correspondiente de los ácidos nucleicos de ARN molde. Las moléculas de ADN en cada subconjunto de núcleos o células se procesan para generar núcleos o células indexadas, donde el procesamiento incluye añadir a los ácidos nucleicos del ADN presentes en cada subconjunto de núcleos o células una primera secuencia índice específica del compartimento para dar como resultado, ácidos nucleicos indexados presentes en núcleos o células indexadas, y luego combinar los núcleos o células indexadas para generar núcleos o células indexadas agrupadas.

En otra realización, un método incluye proporcionar una pluralidad de núcleos o células en una primera pluralidad de compartimentos, donde cada compartimento comprende un subconjunto de núcleos o células. Cada subconjunto se pone en contacto con la transcriptasa inversa y un cebador que se aparea con un ácido nucleico de ARN predeterminado, dando como resultado, ácidos nucleicos de ADN bicatenario con el cebador y la secuencia de nucleótidos de ADN correspondiente de los ácidos nucleicos de ARN molde. Las moléculas de ADN en cada subconjunto de núcleos o células se procesan para generar núcleos o células indexadas, donde el procesamiento incluye añadir a los ácidos nucleicos del ADN presentes en cada subconjunto de núcleos o células una primera secuencia índice específica del compartimento para dar como resultado, ácidos nucleicos indexados presentes en núcleos o células indexadas, y luego combinar los núcleos o células indexadas para generar núcleos o células indexadas agrupadas. Los núcleos o células indexados agrupados se dividen y luego se procesan adicionalmente para añadir un segundo índice específico del compartimento a las moléculas de ADN, se combinan, se dividen y se procesan adicionalmente para añadir un tercer índice específico del compartimento a las moléculas de ADN.

Definiciones

Se entenderá que los términos usados en la presente memoria toman su significado habitual en la técnica pertinente, a menos que se especifique lo contrario. A continuación se exponen varios términos usados en la presente memoria y sus significados.

Tal como se usan en la presente memoria, los términos "organismo", "sujeto" se usan indistintamente y se refieren a microbios (p. ej., procariotas o eucariotas), animales y plantas. Un ejemplo de animal es un mamífero, como un ser humano.

Tal como se usa en la presente memoria, el término "tipo de célula" pretende identificar células basándose en su morfología, fenotipo, origen de desarrollo u otra característica celular distintiva conocida o reconocible. Se puede obtener una variedad de tipos de células diferentes de un solo organismo (o de la misma especie de organismo). Los tipos de células ejemplares incluyen, pero no se limitan a, gametos (incluidos gametos femeninos, p. ej., óvulos u óvulos, y gametos masculinos, p. ej., esperma), epiteliales del ovario, fibroblasto de ovario, testiculares, vejiga urinaria, células inmunitarias, células B, células T, células asesinas naturales, células dendríticas, células cancerosas, células eucariotas, células madre, células sanguíneas, células musculares, células grasas, células de la piel, células nerviosas, células óseas, células pancreáticas, células endoteliales, epiteliales pancreáticas, alfa pancreáticas, beta pancreáticas, endotelial pancreático, linfoblasto de médula ósea, linfoblasto B de médula ósea, macrófago de médula ósea, eritroblasto de médula ósea, dendrítica de médula ósea, adipocito de médula ósea, osteocito de médula ósea, condrocito de médula ósea, promieloblasto, megacarioblasto de médula ósea, vejiga, linfocito B del cerebro, glial cerebral, neurona, astrocito cerebral, neuroectodermo, macrófago cerebral, microglía cerebral, epitelial cerebral, neurona cortical, fibroblasto cerebral, epitelial mamario, epitelial de colon, linfocito B de colon, epitelial mamario, mioepitelial mamario, fibroblasto mamario, enterocito de colon, epitelial de cuello uterino, epitelial del conducto mamario, epitelial de la lengua, dendrítica de la amígdala, linfocito B de la amígdala, linfoblasto de sangre periférica, linfoblasto T de sangre periférica, linfocito T cutáneo de sangre periférica, asesino natural de sangre periférica, linfoblasto B de sangre periférica, monocito de sangre periférica, mieloblasto de sangre periférica, monoblasto de sangre periférica, promieloblasto de sangre periférica, macrófago de sangre periférica, basófilo de sangre periférica, endotelial del hígado, mastocitos hepáticos, epiteliales del hígado, linfocitos B del hígado, endotelial del bazo, epitelial del bazo, linfocitos B del bazo, hepatocitos del hígado, hígado, fibroblastos, epitelial del pulmón, epitelial de los bronquios, fibroblasto del pulmón, linfocito B de pulmón, Schwann de pulmón, escamosa de pulmón, macrófago de pulmón, osteoblasto de pulmón, neuroendocrino, alveolar de pulmón, epitelial de estómago y fibroblasto de estómago.

Tal como se usa en la presente memoria, el término "tejido" pretende significar una colección o agregación de células que actúan juntas para realizar una o más funciones específicas en un organismo. Las células pueden ser opcionalmente morfológicamente similares. Los tejidos ejemplares incluyen, pero no se limitan a, embrionario, epidídimo, ojos, músculos, piel, tendones, venas, arterias, sangre, corazón, bazo, ganglios linfáticos, huesos, médula ósea, pulmones, bronquios, tráquea, intestino e intestino delgado, intestino grueso, colon, recto, glándula salival, lengua, vesícula biliar, apéndice, hígado, páncreas, cerebro, estómago, piel, riñón, uréter, vejiga, uretra, gónada, testículo, ovario, útero, trompa de Falopio, timo, hipófisis, tiroides, suprarrenal o paratiroides. El tejido puede derivarse de cualquiera de una variedad de órganos de un organismo humano u otro. Un tejido puede ser un tejido sano o un tejido no sano. Ejemplos de tejidos no saludables incluyen, pero no se limitan a, neoplasias malignas en el tejido reproductor, pulmón, mama, colorrectal, próstata, nasofaringe, estómago, testículos, piel, sistema nervioso, huesos, ovarios, hígado, tejidos hematológicos, páncreas, útero, riñón, tejidos linfoides, etc. Las neoplasias malignas pueden ser de una variedad de subtipos histológicos, por ejemplo, carcinoma, adenocarcinoma, sarcoma, fibroadenocarcinoma, neuroendocrino o no diferenciados.

Tal como se usa en la presente memoria, el término "compartimento" pretende indicar un área o volumen que separa o aísla algo de otras cosas. Los compartimentos ejemplares incluyen, pero no se limitan a, viales, tubos, pocillos, gotitas, bolos, perlas, vasos, características de la superficie o áreas o volúmenes separados por fuerzas físicas tales como flujo de fluido, magnetismo, corriente eléctrica o similares. En una realización, un compartimento es un pocillo de una placa de múltiples pocillos, tal como una placa de 96 o 384 pocillos. Como se usa en la presente memoria, una gotita puede incluir una perla de hidrogel, que es una perla para encapsular uno o más núcleos o células, e incluye

una composición de hidrogel. En algunas realizaciones, la gotita es una gotita homogénea de material de hidrogel o es una gotita hueca que tiene una cubierta de hidrogel polimérico. Ya sea homogénea o hueca, una gota puede ser capaz de encapsular uno o más núcleos o células. En algunas realizaciones, la gotita es una gotita estabilizada con tensioactivo.

Como se usa en la presente memoria, un "complejo de transposoma" se refiere a una enzima de integración y un ácido nucleico que incluye un sitio de reconocimiento de integración. Un "complejo de transposoma" es un complejo funcional formado por una transposasa y un sitio de reconocimiento de transposasa que es capaz de catalizar una reacción de transposición (ver, por ejemplo, Gunderson et al., patente internacional WO 2016/130704). Ejemplos de enzimas de integración incluyen, pero no se limitan a, una integrasa o una transposasa. Ejemplos de sitios de reconocimiento de integración incluyen, pero no se limitan a, un sitio de reconocimiento de transposasa.

Tal como se usa en la presente memoria, se pretende que el término "ácido nucleico" sea coherente con su uso en la técnica e incluye ácidos nucleicos que existen de forma natural o análogos funcionales de los mismos. Los análogos funcionales particularmente útiles son capaces de hibridarse con un ácido nucleico de una manera específica de secuencia o capaces de usarse como plantilla para la replicación de una secuencia de nucleótidos particular. Los ácidos nucleicos que existen de forma natural generalmente tienen una cadena principal que contiene enlaces fosfodiéster. Una estructura análoga puede tener un enlace a una estructura principal alternativo que incluye cualquiera de una variedad de los conocidos en la técnica. Los ácidos nucleicos que existen de manera natural generalmente tienen un azúcar desoxirribosa (p. ej., que se encuentra en el ácido desoxirribonucleico (ADN)) o un azúcar ribosa (p. ej., que se encuentra en el ácido ribonucleico (ARN)). Un ácido nucleico puede contener cualquiera de una variedad de análogos de estos restos de azúcar que se conocen en la técnica. Un ácido nucleico puede incluir bases nativas o no nativas. A este respecto, un ácido desoxirribonucleico nativo puede tener una o más bases seleccionadas del grupo que consiste en adenina, timina, citosina o guanina y un ácido ribonucleico puede tener una o más bases seleccionadas del grupo que consiste en adenina, uracilo, citosina o guanina. En la técnica se conocen bases no nativas útiles que pueden incluirse en un ácido nucleico. Ejemplos de bases no nativas incluyen un ácido nucleico bloqueado (LNA), un ácido nucleico en puente (BNA) y bases pseudocomplementarias (Trilink Biotechnologies, San Diego, CA). Las bases de LNA y BNA se pueden incorporar en un oligonucleótido de ADN y aumentar la fuerza y la especificidad de la hibridación del oligonucleótido. Las bases de LNA y BNA y los usos de dichas bases son conocidos por el experto en la técnica y son rutinarios. A menos que se indique lo contrario, el término "ácido nucleico" incluye ARNm natural y no natural, ARN no codificante, p. ej., ARN sin poli-A en el extremo 3', ácidos nucleicos derivados de un ARN, p. ej., ADNc y ADN.

Como se usa en la presente memoria, el término "diana", cuando se usa en referencia a un ácido nucleico, pretende ser un identificador semántico para el ácido nucleico en el contexto de un método o composición descrita en la presente memoria y no limita necesariamente la estructura o función del ácido nucleico más allá de lo que se indica explícitamente. Un ácido nucleico diana puede ser esencialmente cualquier ácido nucleico de secuencia conocida o desconocida. Puede ser, por ejemplo, un fragmento de ADN genómico (p. ej., ADN cromosómico), ADN extracromosómico tal como un plásmido, ADN libre de células, ARN (p. ej., ARN o ARN no codificante), proteínas (p. ej., proteínas celulares o de la superficie celular) o ADNc. La secuenciación puede dar como resultado la determinación de la secuencia de la totalidad o de una parte de la molécula diana. Las dianas pueden derivarse de una muestra de ácido nucleico primario, tal como un núcleo. En una realización, las dianas se pueden procesar en plantillas adecuadas para la amplificación mediante la colocación de secuencias universales en uno o ambos extremos de cada fragmento diana. Las dianas también se pueden obtener a partir de una muestra de ARN primaria mediante transcripción inversa en ADNc. En una realización, la diana se usa en referencia a un subconjunto de ADN, ARN o proteínas presentes en la célula. La secuenciación dirigida usa la selección y el aislamiento de genes, regiones o proteínas de interés, normalmente mediante amplificación por PCR (p. ej., cebadores específicos de región) o métodos de captura basados en hibridación o anticuerpos. El enriquecimiento dirigido puede ocurrir en varias etapas del método. Por ejemplo, se puede obtener una representación de ARN diana usando cebadores específicos de la diana en la etapa de transcripción inversa o enriquecimiento basado en hibridación de un subconjunto de una biblioteca más compleja. Un ejemplo es la secuenciación del exoma o el ensayo L1000 (Subramanian et al., 2017, Cell, 171;1437-1452). La secuenciación dirigida puede incluir cualquiera de los procesos de enriquecimiento conocidos por un experto en la técnica.

Como se usa en la presente memoria, el término "universal", cuando se usa para describir una secuencia de nucleótidos, se refiere a una región de secuencia que es común a dos o más moléculas de ácido nucleico donde las moléculas también tienen regiones de secuencia que difieren entre sí. Una secuencia universal que está presente en diferentes miembros de una colección de moléculas puede permitir la captura de múltiples ácidos nucleicos diferentes usando una población de ácidos nucleicos de captura universal, p. ej., oligonucleótidos de captura que son complementarios a una porción de la secuencia universal, p. ej., una secuencia de captura universal. Los ejemplos no limitantes de secuencias de captura universal incluyen secuencias que son idénticas o complementarias a los cebadores P5 y P7. De manera similar, una secuencia universal presente en diferentes miembros de una colección de moléculas puede permitir la replicación (p. ej., secuenciación) o amplificación de múltiples ácidos nucleicos diferentes usando una población de cebadores universales que son complementarios a una porción de la secuencia universal, p. ej., una secuencia de anclaje universal. En una realización, se usan secuencias de anclaje universales

como un sitio con el que se aparea un cebador universal (p. ej., un cebador de secuenciación para la lectura 1 o la lectura 2) para la secuenciación. Por lo tanto, un oligonucleótido de captura o un cebador universal incluye una secuencia que puede hibridarse específicamente con una secuencia universal.

Los términos "P5" y "P7" pueden usarse cuando se hace referencia a una secuencia de captura universal o un oligonucleótido de captura. Los términos "P5'" (P5 primo) y "P7'" (P7 primo) se refieren al complemento de P5 y P7, respectivamente. Se entenderá que en los métodos presentados en la presente memoria se puede usar cualquier secuencia de captura universal adecuada o un oligonucleótido de captura, y que el uso de P5 y P7 son realizaciones ejemplares únicamente. Los usos de oligonucleótidos de captura tales como P5 y P7 o sus complementos en células de flujo se conocen en la técnica, como lo ejemplifican las descripciones de las patentes internacionales WO 2007/010251, WO 2006/064199, WO 2005/065814, WO 2015/106941, WO 1998/044151 y WO 2000/018957. Por ejemplo, cualquier cebador de amplificación directa adecuado, ya sea inmovilizado o en disolución, puede ser útil en los métodos presentados en la presente memoria para la hibridación con una secuencia complementaria y la amplificación de una secuencia. De manera similar, cualquier cebador de amplificación inversa adecuado, ya sea inmovilizado o en disolución, puede ser útil en los métodos presentados en la presente memoria para la hibridación con una secuencia complementaria y la amplificación de una secuencia. Un experto en la técnica entenderá cómo diseñar y usar secuencias de cebadores que sean adecuadas para la captura y/o amplificación de ácidos nucleicos como se presenta en la presente memoria.

Tal como se usa en la presente memoria, el término "cebador" y sus derivados se refieren generalmente a cualquier ácido nucleico que pueda hibridarse con una secuencia diana de interés. Normalmente, el cebador funciona como un sustrato sobre el que se pueden polimerizar los nucleótidos mediante una polimerasa o al que se puede ligar una secuencia de nucleótidos tal como un índice; sin embargo, en algunas realizaciones, el cebador puede incorporarse a la cadena de ácido nucleico sintetizada y proporcionar un sitio con el que otro cebador puede hibridarse para cebar la síntesis de una nueva cadena que es complementaria a la molécula de ácido nucleico sintetizada. El cebador puede incluir cualquier combinación de nucleótidos o análogos de los mismos. En algunas realizaciones, el cebador es un oligonucleótido o polinucleótido monocatenario. Los términos "polinucleótido" y "oligonucleótido" se usan indistintamente en la presente memoria para referirse a una forma polimérica de nucleótidos de cualquier longitud, y pueden incluir ribonucleótidos, desoxirribonucleótidos, análogos de los mismos o mezclas de los mismos. Debe entenderse que los términos incluyen, como equivalentes, análogos de ADN, ARN, ADNc o conjugados anticuerpo-oligo elaborados a partir de análogos de nucleótidos y que son aplicables a polinucleótidos monocatenarios (tales como sentido o antisentido) y bicatenarios. El término como se usa en la presente memoria también abarca ADNc, es decir, ADN complementario o copia producido a partir de un molde de ARN, por ejemplo mediante la acción de la transcriptasa inversa. Este término se refiere únicamente a la estructura primaria de la molécula. Por lo tanto, el término incluye ácido desoxirribonucleico ("ADN") tri-, bi- y monocatenario, así como ácido ribonucleico ("ARN") tri-, bi- y monocatenario.

Como se usa en la presente memoria, el término "adaptador" y sus derivados, p. ej., adaptador universal, se refiere generalmente a cualquier oligonucleótido lineal que pueda unirse a una molécula de ácido nucleico de la descripción. En algunas realizaciones, el adaptador es sustancialmente no complementario al extremo 3' o al extremo 5' de cualquier secuencia diana presente en la muestra. En algunas realizaciones, las longitudes de adaptador adecuadas están en el intervalo de aproximadamente 10 a 100 nucleótidos, aproximadamente 12 a 60 nucleótidos o aproximadamente 15 a 50 nucleótidos de longitud. Generalmente, el adaptador puede incluir cualquier combinación de nucleótidos y/o ácidos nucleicos. En algunos aspectos, el adaptador puede incluir uno o más grupos escindibles en una o más posiciones. En otro aspecto, el adaptador puede incluir una secuencia que sea sustancialmente idéntica, o sustancialmente complementaria, a al menos una porción de un cebador, por ejemplo un cebador universal. En algunas realizaciones, el adaptador puede incluir un código de barras (también denominado en la presente memoria como etiqueta o índice) para ayudar con la corrección, identificación o secuenciación de errores situados más abajo. Los términos "adaptador" y "adaptador" se utilizan indistintamente.

Tal como se usa en la presente memoria, el término "cada uno", cuando se usa en referencia a una colección de elementos, tiene como objetivo identificar un elemento individual de la colección, pero no necesariamente se refiere a todos los elementos de la colección a menos que el contexto indique claramente lo contrario.

Tal como se usa en la presente memoria, el término "transporte" se refiere al movimiento de una molécula a través de un fluido. El término puede incluir transporte pasivo, como el movimiento de moléculas a lo largo de su gradiente de concentración (p. ej., difusión pasiva). El término también puede incluir transporte activo mediante el cual las moléculas pueden moverse a lo largo de su gradiente de concentración o en contra de su gradiente de concentración. Por tanto, el transporte puede incluir la aplicación de energía para mover una o más moléculas en una dirección deseada o hasta una posición deseada tal como un sitio de amplificación.

Como se usa en la presente memoria, "amplificar", "que amplifica" o "reacción de amplificación" y sus derivados, se refieren generalmente a cualquier acción o proceso mediante el cual al menos una porción de una molécula de ácido nucleico se replica o copia en al menos una molécula de ácido nucleico adicional. La molécula de ácido nucleico adicional incluye opcionalmente una secuencia que es sustancialmente idéntica o sustancialmente complementaria a al menos alguna porción de la molécula de ácido nucleico molde. La molécula de ácido nucleico molde puede ser

monocatenaria o bicatenaria y la molécula de ácido nucleico adicional puede ser independientemente monocatenaria o bicatenaria. La amplificación incluye opcionalmente la replicación lineal o exponencial de una molécula de ácido nucleico. En algunas realizaciones, dicha amplificación se puede realizar usando condiciones isotérmicas; en otras realizaciones, dicha amplificación puede incluir termociclado. En algunas realizaciones, la amplificación es una

amplificación múltiple que incluye la amplificación simultánea de una pluralidad de secuencias diana en una única reacción de amplificación. En algunas realizaciones, "amplificación" incluye la amplificación de al menos una porción de los ácidos nucleicos basados en ADN y ARN solos o en combinación. La reacción de amplificación puede incluir cualquiera de los procesos de amplificación conocidos por un experto en la técnica. En algunas realizaciones, la reacción de amplificación incluye la reacción en cadena de la polimerasa (PCR).

Como se usa en la presente memoria, "condiciones de amplificación" y sus derivados generalmente se refieren a condiciones adecuadas para amplificar una o más secuencias de ácidos nucleicos. Dicha amplificación puede ser lineal o exponencial. En algunas realizaciones, las condiciones de amplificación pueden incluir condiciones isotérmicas o, alternativamente, pueden incluir condiciones de termociclado, o una combinación de condiciones isotérmicas y de termociclado. En algunas realizaciones, las condiciones adecuadas para amplificar una o más secuencias de ácidos nucleicos incluyen condiciones de reacción en cadena de la polimerasa (PCR). Normalmente, las condiciones de amplificación se refieren a una mezcla de reacción que es suficiente para amplificar ácidos nucleicos tales como una o más secuencias diana flanqueadas por una secuencia universal, o para amplificar una secuencia diana amplificada ligada a uno o más adaptadores. Generalmente, las condiciones de amplificación incluyen un catalizador para la amplificación o para la síntesis de ácidos nucleicos, por ejemplo una polimerasa; un cebador que posee cierto grado de complementariedad con el ácido nucleico que se va a amplificar; y nucleótidos, tales como trifosfatos de desoxirribonucleótidos (dNTP) para promover la extensión del cebador una vez hibridado con el ácido nucleico. Las condiciones de amplificación pueden requerir hibridación o apareamiento de un cebador con un ácido nucleico, extensión del cebador y una etapa de desnaturalización en la que el cebador extendido se separa de la secuencia de ácido nucleico que se somete a amplificación. Normalmente, pero no necesariamente, las condiciones de amplificación pueden incluir termociclado; en algunas realizaciones, las condiciones de amplificación incluyen una pluralidad de ciclos donde se repiten las etapas de apareamiento, extensión y separación. Normalmente, las condiciones de amplificación incluyen cationes como Mg^{2+} o Mn^{2+} y también puede incluir varios modificadores de la fuerza iónica.

Como se usa en la presente memoria, "reamplificación" y sus derivados se refieren generalmente a cualquier proceso mediante el cual al menos una porción de una molécula de ácido nucleico amplificada se amplifica adicionalmente mediante cualquier proceso de amplificación adecuado (denominado en algunas realizaciones como una amplificación "secundaria"), produciendo así una molécula de ácido nucleico reamplificada. No es necesario que la amplificación secundaria sea idéntica al proceso de amplificación original mediante el cual se produjo la molécula de ácido nucleico amplificada; tampoco es necesario que la molécula de ácido nucleico reamplificada sea completamente idéntica o completamente complementaria a la molécula de ácido nucleico amplificada; todo lo que se requiere es que la molécula de ácido nucleico reamplificada incluya al menos una porción de la molécula de ácido nucleico amplificada o su complemento. Por ejemplo, la reamplificación puede implicar el uso de diferentes condiciones de amplificación y/o diferentes cebadores, incluidos cebadores específicos de diana diferentes a los de la amplificación primaria.

Tal como se usa en la presente memoria, el término "reacción en cadena de la polimerasa" ("PCR") se refiere al método de Mullis Patentes de EE.UU. núms. 4.683.195 y 4.683.202, que describen un método para aumentar la concentración de un segmento de un polinucleótido de interés en una mezcla de ADN genómico sin clonación ni purificación. Este proceso para amplificar el polinucleótido de interés consiste en introducir un gran exceso de dos cebadores oligonucleotídicos en la mezcla de ADN que contiene el polinucleótido de interés deseado, seguido de una serie de ciclados térmicos en presencia de una ADN polimerasa. Los dos cebadores son complementarios a sus respectivas cadenas del polinucleótido bicatenario de interés. La mezcla se desnaturaliza primero a una temperatura más alta y luego los cebadores se aparean con secuencias complementarias dentro del polinucleótido de la molécula de interés. Después del apareamiento, los cebadores se extienden con una polimerasa para formar un nuevo par de hebras complementarias. Las etapas de desnaturalización, apareamiento de cebadores y extensión con polimerasa se pueden repetir muchas veces (lo que se denomina termociclado) para obtener una alta concentración de un segmento amplificado del polinucleótido de interés deseado. La longitud del segmento amplificado del polinucleótido de interés deseado (amplicón) está determinada por las posiciones relativas de los cebadores entre sí y, por lo tanto, esta longitud es un parámetro controlable. Al repetir el proceso, el método se denomina PCR. Debido a que los segmentos amplificados deseados del polinucleótido de interés se convierten en las secuencias de ácido nucleico predominantes (en términos de concentración) en la mezcla, se dice que están "amplificados por PCR". En una modificación del método tratado anteriormente, las moléculas de ácido nucleico diana se pueden amplificar por PCR usando una pluralidad de pares de cebadores diferentes, en algunos casos, uno o más pares de cebadores por molécula de ácido nucleico diana de interés, formando así una reacción de PCR múltiple.

Como se define en la presente memoria, "amplificación múltiple" se refiere a la amplificación selectiva y no aleatoria de dos o más secuencias diana dentro de una muestra usando al menos un cebador específico de la diana. En algunas realizaciones, la amplificación múltiple se realiza de manera que algunas o todas las secuencias diana se amplifiquen dentro de un único recipiente de reacción. El "plexo" o "plex" de una amplificación múltiple determinada se refiere generalmente al número de secuencias específicas de diana diferentes que se amplifican durante esa amplificación

múltiple única. En algunas realizaciones, el plexo puede ser de aproximadamente 12 plex, 24 plex, 48 plex, 96 plex, 192 plex, 384 plex, 768 plex, 1536 plex, 3072 plex, 6144 plex o superior. También es posible detectar las secuencias diana amplificadas mediante varias metodologías diferentes (p. ej., electroforesis en gel seguida de densitometría, cuantificación con un bioanalizador o PCR cuantitativa, hibridación con una sonda marcada; incorporación de cebadores biotinilados seguida de detección del conjugado avidina-enzima; incorporación de trifosfatos de desoxinucleótidos marcados con ^{32}P en la secuencia diana amplificada).

Como se usa en la presente memoria, "secuencias diana amplificadas" y sus derivados se refiere generalmente a una secuencia de ácido nucleico producida mediante la amplificación de las secuencias diana usando cebadores específicos de la diana y los métodos proporcionados en la presente memoria. Las secuencias diana amplificadas pueden ser del mismo sentido (es decir, la hebra positiva) o antisentido (es decir, la hebra negativa) con respecto a las secuencias diana.

Como se usan en la presente memoria, los términos "ligar", "ligadura" y sus derivados se refieren generalmente al proceso para unir covalentemente dos o más moléculas entre sí, por ejemplo, unir covalentemente dos o más moléculas de ácido nucleico entre sí. En algunas realizaciones, la ligadura incluye unir escisiones entre nucleótidos adyacentes de ácidos nucleicos. En algunas realizaciones, la ligadura incluye formar un enlace covalente entre un extremo de una primera y un extremo de una segunda molécula de ácido nucleico. En algunas realizaciones, la ligadura puede incluir la formación de un enlace covalente entre un grupo fosfato 5' de un ácido nucleico y un grupo hidroxilo 3' de un segundo ácido nucleico, formando así una molécula de ácido nucleico ligada. Generalmente, para los fines de esta descripción, se puede ligar una secuencia diana amplificada a un adaptador para generar una secuencia diana amplificada ligada a un adaptador.

Como se usa en la presente memoria, "ligasa" y sus derivados se refiere generalmente a cualquier agente capaz de catalizar la ligadura de dos moléculas de sustrato. En algunas realizaciones, la ligasa incluye una enzima capaz de catalizar la unión de escisiones entre nucleótidos adyacentes de un ácido nucleico. En algunas realizaciones, la ligasa incluye una enzima capaz de catalizar la formación de un enlace covalente entre un fosfato 5' de una molécula de ácido nucleico y un hidroxilo 3' de otra molécula de ácido nucleico, formando así una molécula de ácido nucleico ligada. Las ligasas adecuadas pueden incluir, pero no se limitan a, ligasa de ADN T4, ligasa de ARN T4 y ligasa de ADN de *E. coli*.

Como se usa en la presente memoria, "condiciones de ligadura" y sus derivados generalmente se refiere a condiciones adecuadas para ligar dos moléculas entre sí. En algunas realizaciones, las condiciones de ligadura son adecuadas para sellar escisiones o espacios entre ácidos nucleicos. Tal como se usa en la presente memoria, el término escisión o espacio es consistente con el uso del término en la técnica. Normalmente, se puede ligar una escisión o un espacio en presencia de una enzima, tal como ligasa, a una temperatura y un pH apropiados. En algunas realizaciones, la ligasa de ADN T4 puede unir una escisión entre ácidos nucleicos a una temperatura de aproximadamente 70-72 °C.

El término "célula de flujo", tal como se usa en la presente memoria, se refiere a una cámara que comprende una superficie sólida a través de la cual pueden hacerse fluir uno o más reactivos fluidos. Ejemplos de células de flujo y sistemas fluidicos relacionados y plataformas de detección que se pueden usar fácilmente en los métodos de la presente descripción se describen, por ejemplo, en Bentley et al., *Nature* 456:53-59 (2008), los documentos WO 04/018497; US 7.057.026; WO 91/06678; WO 07/123744; US 7.329.492; US 7.211.414; US 7.315.019; US 7.405.281 y US 2008/0108082.

Como se usa en la presente memoria, el término "amplicón", cuando se usa en referencia a un ácido nucleico, significa el producto de copiar el ácido nucleico, en donde el producto tiene una secuencia de nucleótidos que es igual o complementaria a al menos una porción de la secuencia de nucleótidos del ácido nucleico. Un amplicón puede producirse mediante cualquiera de una variedad de métodos de amplificación que usan el ácido nucleico, o un amplicón del mismo, como una plantilla incluyendo, por ejemplo, extensión de polimerasa, reacción en cadena de la polimerasa (PCR), amplificación por círculo rodante (RCA), extensión de ligadura o reacción en cadena de ligadura. Un amplicón puede ser una molécula de ácido nucleico que tiene una única copia de una secuencia de nucleótidos particular (p. ej., un producto de PCR) o múltiples copias de la secuencia de nucleótidos (p. ej., un producto concatémico de RCA). Un primer amplicón de un ácido nucleico diana suele ser una copia complementaria. Los amplicones posteriores son copias que se crean, después de la generación del primer amplicón, a partir del ácido nucleico diana o del primer amplicón. Un amplicón posterior puede tener una secuencia que sea sustancialmente complementaria al ácido nucleico diana o sustancialmente idéntica al ácido nucleico diana.

Tal como se usa en la presente memoria, el término "sitio de amplificación" se refiere a un sitio en o sobre una matriz donde se pueden generar uno o más amplicones. Un sitio de amplificación puede configurarse además para contener, retener o unir al menos un amplicón que se genera en el sitio.

Tal como se usa en la presente memoria, el término "matriz" se refiere a una población de sitios que pueden diferenciarse entre sí según su posición relativa. Diferentes moléculas que se encuentran en diferentes sitios de una matriz se pueden diferenciar entre sí según las posiciones de los sitios en la matriz. Un sitio individual de una matriz puede incluir una o más moléculas de un tipo particular. Por ejemplo, un sitio puede incluir una única molécula de

ácido nucleico diana que tiene una secuencia particular o un sitio puede incluir varias moléculas de ácido nucleico que tienen la misma secuencia (y/o secuencia complementaria de la misma). Los sitios de una matriz pueden ser diferentes elementos situados en el mismo sustrato. Los elementos ejemplares incluyen, sin limitación, pocillos en un sustrato, perlas (u otras partículas) en o sobre un sustrato, proyecciones desde un sustrato, crestas en un sustrato o canales en un sustrato. Los sitios de una matriz pueden ser sustratos separados, cada uno de los cuales lleva una molécula diferente. Se pueden identificar diferentes moléculas unidas a sustratos separados según las posiciones de los sustratos en una superficie a la que están asociados los sustratos o según las posiciones de los sustratos en un líquido o gel. Los conjuntos ejemplares en los que se colocan sustratos separados en una superficie incluyen, sin limitación, aquellos que tienen perlas en pocillos.

Como se usa en la presente memoria, el término "capacidad", cuando se usa en referencia a un sitio y material de ácido nucleico, significa la cantidad máxima de material de ácido nucleico que puede ocupar el sitio. Por ejemplo, el término puede referirse al número total de moléculas de ácido nucleico que pueden ocupar el sitio en una condición particular. También se pueden usar otras medidas que incluyen, por ejemplo, la masa total de material de ácido nucleico o el número total de copias de una secuencia de nucleótidos particular que puede ocupar el sitio en una condición particular. Normalmente, la capacidad de un sitio para un ácido nucleico diana será sustancialmente equivalente a la capacidad del sitio para amplicones del ácido nucleico diana.

Como se usa en la presente memoria, el término "agente de captura" se refiere a un material, producto químico, molécula o resto de los mismos que es capaz de unirse, retener o enlazarse a una molécula diana (p. ej., un ácido nucleico diana). Los agentes de captura ejemplares incluyen, sin limitación, un ácido nucleico de captura (también denominado en la presente memoria oligonucleótido de captura) que es complementario a al menos una porción de un ácido nucleico diana, un miembro de un par de unión receptor-ligando (p. ej., avidina, estreptavidina, biotina, lectina, carbohidrato, proteína de unión a ácido nucleico, epítopo, anticuerpo, etc.) capaz de unirse a un ácido nucleico diana (o resto de unión unido al mismo), o un reactivo químico capaz de formar un enlace covalente con un ácido nucleico diana (o resto de unión unido al mismo).

Tal como se usa en la presente memoria, el término "resto reportero" puede referirse a cualquier etiqueta, marca, índices, códigos de barras o grupo identificables que permita determinar la composición, identidad y/o la fuente de un analito que se investiga. En algunas realizaciones, un resto reportero puede incluir un anticuerpo que se une específicamente a una proteína. En algunas realizaciones, el anticuerpo puede incluir una etiqueta detectable. En algunas realizaciones, el reportero puede incluir un anticuerpo o reactivo de afinidad marcado con una etiqueta de ácido nucleico. La etiqueta de ácido nucleico puede ser detectable, por ejemplo, mediante un ensayo de ligadura por proximidad (PLA) o un ensayo de extensión por proximidad (PEA) o una lectura basada en secuenciación (Shahi et al. Scientific Reports volumen 7, número de artículo: 44447, 2017) o CITE-seq (Stoeckius et al. Nature Methods 14:865-868, 2017).

Tal como se usa en la presente memoria, el término "población clonal" se refiere a una población de ácidos nucleicos que es homogénea con respecto a una secuencia de nucleótidos particular. La secuencia homogénea normalmente tiene una longitud de al menos 10 nucleótidos, pero puede ser incluso más larga incluyendo, por ejemplo, al menos 50, 100, 250, 500 o 1000 nucleótidos de longitud. Una población clonal puede derivarse de un único ácido nucleico diana o de un ácido nucleico molde. Normalmente, todos los ácidos nucleicos de una población clonal tendrán la misma secuencia de nucleótidos. Se entenderá que puede ocurrir un pequeño número de mutaciones (p. ej., debidas a artefactos de amplificación) en una población clonal sin apartarse de la clonalidad.

Tal como se usa en la presente memoria, el término "identificador molecular único" o "UMI" se refiere a una etiqueta molecular, ya sea aleatoria, no aleatoria o semialeatoria, que puede unirse a un ácido nucleico. Cuando se incorpora a un ácido nucleico, se puede usar un UMI para corregir el sesgo de amplificación posterior contando directamente los identificadores moleculares únicos (UMI) que se secuencian después de la amplificación.

Como se usa en la presente memoria, un compuesto "exógeno", p. ej., una enzima exógena, se refiere a un compuesto que no se encuentra normal o naturalmente en una composición particular. Por ejemplo, cuando la composición particular incluye un lisado celular, una enzima exógena es una enzima que no se encuentra normal o naturalmente en el lisado celular.

Como se usa en la presente memoria, "proporcionar" en el contexto de una composición, un artículo, un ácido nucleico o un núcleo significa fabricar la composición, artículo, ácido nucleico o núcleo, comprar la composición, artículo, ácido nucleico o núcleo, u obtener de otro modo el compuesto, composición, artículo o núcleo.

El término "y/o" significa uno o todos los elementos enumerados o una combinación de dos o más de los elementos enumerados.

Las palabras "preferido" y "preferiblemente" se refieren a realizaciones de la descripción que pueden proporcionar ciertos beneficios, bajo ciertas circunstancias. Sin embargo, también pueden preferirse otras realizaciones, en las mismas u otras circunstancias. Además, la enumeración de una o más realizaciones preferidas no implica que otras realizaciones no sean útiles y no pretende excluir otras realizaciones del alcance de la descripción.

Los términos "comprende" y sus variaciones no tienen un significado limitante cuando estos términos aparecen en la descripción y las reivindicaciones.

Se entiende que dondequiera que se describan realizaciones en la presente memoria con el lenguaje "incluir", "incluye" o "que incluye" y similares, también se proporcionan realizaciones asimismo análogas descritas en términos de "que consiste en" y/o "que consiste esencialmente en".

A menos que se especifique lo contrario, "un", "una", "el/la" y "al menos uno" se usan indistintamente y significan uno o más de uno.

También en la presente memoria, las enumeraciones de intervalos numéricos por puntos finales incluyen todos los números incluidos dentro de ese intervalo (p. ej., 1 a 5 incluye 1, 1,5, 2, 2,75, 3, 3,80, 4, 5, etc.).

Para cualquier método descrito en la presente memoria que incluya etapas discretas, las etapas pueden realizarse en cualquier orden factible. Y, según corresponda, cualquier combinación de dos o más etapas podrá realizarse simultáneamente.

La referencia a lo largo de esta memoria a "una realización", "una realización", "ciertas realizaciones" o "algunas realizaciones", etc., significa que un elemento, configuración, composición o característica particular descrita en relación con la realización está incluida en al menos una realización de la descripción. Por tanto, las apariciones de tales frases en varios lugares a lo largo de esta memoria no se refieren necesariamente a la misma realización de la descripción. Además, los elementos, configuraciones, composiciones o características particulares se pueden combinar de cualquier manera adecuada en una o más realizaciones.

Breve descripción de las figuras

La siguiente descripción detallada de realizaciones ilustrativas de la presente descripción se puede entender mejor cuando se lee junto con los siguientes dibujos.

La FIG. 1 muestra un diagrama de bloques general de un método ilustrativo general para la indexación combinatoria unicelular según la presente descripción.

La FIG. 2 muestra un diagrama de bloques general de un método ilustrativo general para la indexación combinatoria unicelular según la presente descripción.

La FIG. 3 muestra un diagrama de bloques general de un método ilustrativo general para la indexación combinatoria unicelular según la presente descripción.

La FIG. 4 muestra que sci-RNA-seq3 permite la elaboración de perfiles de ~ 2 millones de células de 61 embriones de ratón en 5 etapas de desarrollo en un experimento. (A) esquema experimental y flujo de trabajo de sci-RNA-seq3. (B) Comparación del rendimiento por experimento aquí con informes recientes. (C) Diagrama de dispersión del recuento de UMI de ratón frente a humano de células HEK293T y NIH/3T3. (D) Diagrama de barras que muestra el número de células perfiladas de cada uno de los 61 embriones de ratón. (E) Diagrama de caja que muestra el número de genes y UMI detectados por célula. (F) Diagrama de dispersión de lecturas únicas alineadas con los genes Xist (cromosoma X) frente a cromosoma Y por embrión de ratón. (G) Trayectoria pseudotemporal de perfiles pseudo-masivos de secuencia de ARN de embriones de ratón. (H) Mapa de calor de los cambios en la expresión del gen marcador E9.5 a E13.5 a partir de perfiles pseudo-masivos de secuencia de ARN de embriones de ratón.

La FIG. 5 muestra los análisis relacionados con el rendimiento y el control de calidad para sci-RNA-seq3. (A) Diagrama de barras que muestra el número de pocillos RT usados para cada uno de los 61 embriones de ratón. (B) Histograma que muestra la distribución de lecturas de secuenciación sin procesar de cada pocillo de PCR en sci-RNA-seq3. (C) Diagrama de caja que muestra el número de UMI por célula de células HEK293T y NIH/3T3. (D) Diagrama de caja que muestra la proporción de lecturas asignadas a las especies esperadas para las células HEK293T (humana) y NIH/3T3 (ratón). (E) Diagrama de caja que compara el número de UMI por célula (reducido a 20.000 lecturas sin procesar por célula) para sci-RNA-seq3 frente a sci-RNA-seq. (F) Correlación entre las mediciones de expresión génica en perfiles añadidos de células HEK293T de sci-RNA-seq3 frente a sci-RNA-seq.

La FIG. 6 muestra análisis adicionales relacionados con el control de calidad y el rendimiento para sci-RNA-seq3. (A) Diagrama de dispersión que muestra la correlación entre el número de pocillos RT utilizados y el número de células recuperadas por embrión. (B-D) Diagrama de barras que compara el número de lecturas de secuenciación sin procesar (B), genes detectados (C) y UMI (D) por célula entre sci-RNA-seq3 y otros métodos. (E) Diagrama de caja que muestra el número de UMI detectados por célula de embriones en cinco etapas de desarrollo. (F) Diagrama de barras que muestra el número de embriones macho y hembra perfilados en cada etapa de desarrollo.

La FIG. 7 muestra perfiles pseudo-masivos de secuencia de ARN de embriones de ratón que se separan fácilmente por etapa de desarrollo. (A) t-SNE de los transcriptomas añadidos de células individuales derivadas de cada uno de 61 embriones de ratón da como resultado cinco grupos estrechamente agrupados que coinciden perfectamente con

sus etapas de desarrollo. (B) Trayectoria de pseudotiempo de perfiles de secuencia de ARN pseudo-masivos de embriones de ratón, idéntica a la Fig. 4G, pero coloreada por pseudotiempo. (C) Los 61 embriones perfilados se ordenaron por pseudotiempo. Los tres embriones E10.5 más antiguos y los tres más recientes (en pseudotiempo) se muestran en fotografías y parecen ser morfológicamente distintos.

5 La FIG. 8 muestra la identificación de los principales tipos de células de la organogénesis del ratón. (A) Visualización con t-SNE de 2.026.641 células de embriones de ratón, coloreadas por identificación de grupo a partir de agrupaciones de Louvain y anotadas en función de genes marcadores. El mismo t-SNE se representa a continuación, mostrando solo células de cada etapa de desarrollo. Los grupos eritroides primitivos (transitorios) y eritroides definitivos (en expansión) están recuadrados para ilustrar su dinámica proliferativa (B). Diagrama de puntos que muestra la expresión de un gen marcador seleccionado por tipo de célula. El tamaño del punto codifica el porcentaje de células dentro de un tipo de célula y su color codifica el nivel de expresión medio.

10 La FIG. 9 muestra la identificación de los principales tipos de células de la organogénesis del ratón y los correspondientes conjuntos de genes marcadores específicos del tipo de célula. (A) Visualización con t-SNE de la transición del estado celular de E9.5 a E13.5, coloreada por etapas de desarrollo. Este es el mismo t-SNE que se muestra en la Fig. 8A, pero en cada gráfico solo se muestran las células derivadas de un único punto temporal. (B) Mapa de calor que muestra la expresión relativa de genes en los principales tipos de células identificados. (C) Diagrama de barras que muestra el número de genes marcadores en cada tipo de célula con una expresión más del doble en comparación con el segundo tipo de célula con mayor expresión (FDR del 5 %).

15 La FIG. 10 muestra que las células derivadas de embriones replicados del mismo momento no exhiben efectos de lote obvios. (A-E) Visualización con t-SNE de células embrionarias de ratón de diferentes etapas de desarrollo: E9.5 (A), E10.5 (B), E11.5 (C), E12.5 (D), E13.5 (E), coloreado por ID de embrión en cada etapa.

20 La FIG. 11 muestra la dinámica del número de tipos de células durante la organogénesis en ratones. (A) Diagrama de barras que muestra el número de células perfiladas para cada tipo de célula, divididas por etapa de desarrollo. (B) Mapa de calor que muestra el número de células relativo de cada tipo de célula (filas) en 61 embriones de ratón (columnas). Se calculó una estimación del número absoluto de células por tipo de célula por embrión multiplicando la proporción con que ese tipo de célula contribuyó a un embrión determinado por el número total estimado de células en esa etapa de desarrollo. Para la presentación, estas estimaciones se normalizan en cada fila por el número máximo de células estimado para un tipo de célula determinado en los 61 embriones. Los embriones se clasifican de izquierda a derecha según el pseudotiempo de desarrollo. (C) Diagrama de líneas que muestra el cambio relativo en el número de células para los linajes eritroide primitivo y eritroide definitivo, calculado como en el panel B. Las líneas discontinuas muestran la expresión relativa de genes marcadores para eritroide primitivo (Hbb-bh1) y eritroide definitivo (Hbb-bs). Los puntos de datos de embriones individuales se ordenaron según el pseudotiempo de desarrollo y se suavizaron mediante el método loess.

25 La FIG. 12 muestra la agrupación de Louvain y la visualización con t-SNE de subgrupos de cada uno de los 38 tipos de células principales. Como la heterogeneidad del tipo de célula era fácilmente evidente en muchos de los 38 grupos que se muestran en la Fig. 8A, adoptamos una estrategia iterativa, repitiendo la agrupación de Louvain en cada tipo de célula principal para identificar subgrupos. Después de que se eliminaron los subgrupos dominados por uno o dos embriones y se fusionaron subgrupos muy similares, se obtuvo un total de 655 subgrupos (también denominados "subtipos" para distinguirlos de los 38 tipos de células principales identificados por el agrupamiento inicial).

30 La FIG. 13 muestra la agrupación de Louvain y la visualización con t-SNE de subgrupos de cada uno de los 38 tipos de células principales en las etapas de desarrollo. Idéntico a la Fig. 12, pero con células coloreadas según la etapa de desarrollo en lugar de la ID del subgrupo.

35 La FIG. 14 muestra que la sensibilidad para detectar tipos de células es una función de la cobertura celular. (A) Visualización con t-SNE de todas las células (gráfico de la izquierda, $n = 2.026.641$) y subconjunto reducido de muestra (gráfico de la derecha, $n = 50.000$), coloreado por los ID de los grupos de Louvain de la Fig. 8A. (B) Visualización con t-SNE de todas las células endoteliales (gráfico de la izquierda, $n = 35.878$) y aquellas del subconjunto reducido de muestra (gráfico de la derecha, $n = 1.173$), coloreadas por el ID del grupo de Louvain calculado en base a las 35.878 células endoteliales. (C) Visualización con t-SNE de las 1.173 células endoteliales, coloreada por el ID del grupo de Louvain calculado en base a las 1.173 células endoteliales.

40 La FIG. 15 muestra los 655 subtipos de células que se derivan de múltiples embriones y están definidos por conjuntos de marcadores. (A) Histograma que muestra la distribución de subgrupos con respecto al número de células (mediana 1.869; intervalo 51-65.894). (B) Histograma que muestra la distribución de subgrupos con respecto al número de embriones contribuyentes (>5 células para calificar como contribuyente). (C) Histograma que muestra la distribución de subgrupos con respecto a la relación de células derivadas del embrión que más contribuye. (D) Histograma que muestra la distribución de subgrupos con respecto al número de genes marcadores (expresión al menos 1,5 veces mayor en comparación con el segundo subtipo de célula con mayor expresión dentro del mismo grupo principal; FDR del 5 %).

La FIG. 16 muestra la dinámica de los números de subtipos celulares durante la organogénesis en ratones. (A) Mapa de calor que muestra la expresión relativa de genes en los 655 subgrupos identificados. (B) Mapa de calor que muestra el número de células relativo de cada subtipo de células (filas) en 61 embriones de ratón (columnas). Se calculó una estimación del número absoluto de células por subtipo de célula por embrión como en la Fig. 11B. (C) Visualización con t-SNE de los 61 embriones de ratón basada únicamente en las proporciones de los 655 subtipos de células en cada embrión.

La FIG. 17 muestra la identificación y caracterización de los subtipos de células epiteliales y la cresta ectodérmica apical (AER) de la extremidad. (A) Visualización con t-SNE y anotación basada en marcadores de subtipos de células epiteliales. (B) Visualización con t-SNE de todas las células epiteliales coloreadas por el nivel de expresión de *Fgf8*. (C) Imágenes de hibridación in situ de *Fgf8* en embriones E10.5 (izquierda) y E11.5 (derecha). (D) Visualización con t-SNE de todas las células epiteliales coloreadas por el nivel de expresión de *Fndc3a*. (E) Imágenes de hibridación in situ de *Fndc3a* en el embrión E10.5. Flecha: sitio de expresión génica. (F) Diagrama de caja que muestra la proporción de células de AER por embrión en diferentes etapas de desarrollo. (G) Trayectoria pseudotemporal de los transcriptomas unicelulares de AER, coloreada según la etapa de desarrollo. (H) Diagrama de líneas que muestra la expresión relativa de genes marcadores de AER a lo largo del pseudotiempo de desarrollo.

La FIG. 18 muestra la identificación de los subtipos de epitelio de ratón. Diagrama de puntos que muestra la expresión de un gen marcador seleccionado por subtipo epitelial. El tamaño del punto codifica el porcentaje de células dentro de un tipo de célula y su color codifica el nivel de expresión medio.

La FIG. 19 muestra la dinámica de la expresión génica de las células de la cresta ectodérmica apical (AER) de las extremidades a lo largo del pseudotiempo. (A) Mapa de calor que muestra la expresión génica diferencial dependiente del pseudotiempo suavizada (FDR del 1 %) en células de AER, generado por una regresión binomial negativa y escalada como porcentaje de la expresión génica máxima. Cada fila indica un gen diferente, y estos se dividen en subconjuntos que están activados (arriba), reprimidos (medio) o exhiben una dinámica transitoria (abajo) entre E9.5 y E13.5. (B-C) Diagramas que muestran el valor *q* transformado en $-\log_{10}$ y la puntuación combinada basada en *enrichR* de términos de Reactoma enriquecido (B) y factores de transcripción (C) para genes cuya expresión disminuye significativamente en el desarrollo de AER. Los principales términos de la vía enriquecida (Reactome2016) para genes que disminuyen significativamente incluyen la progresión del ciclo celular (ciclo celular mitótico, *qval* = 0,0002) y el metabolismo de la glucosa (metabolismo de los carbohidratos, *qval* = 0,0002). Los TF más enriquecidos con dianas de genes decrecientes incluyen factores pluripotentes como *Isl1* (*qval* <10-5), *Pou5f1* (*qval* = 0,002) y *Nanog* (*qval* = 0,003).

La FIG. 20 muestra la caracterización de las trayectorias celulares durante la diferenciación del mesénquima de las extremidades. (A) Visualización con UMAP 3D de células mesenquimales de las extremidades coloreadas por etapa de desarrollo (la izquierda y la derecha representan vistas desde dos direcciones). (B) Diagrama de dispersión que muestra la expresión normalizada de *Pitx1* y *Tbx5* en células mesenquimales de las extremidades. Solo se mostraron las células en las que se detectaron *Pitx1* y/o *Tbx5*. (C) Diagrama de volcán que muestra los genes expresados diferencialmente (FDR del 5 %, coloreado en rojo) entre las extremidades anteriores y posteriores. Los principales genes expresados diferencialmente están marcados. Eje X: cambio de veces transformado a $\log_2$ entre las extremidades anteriores y posteriores para cada gen. Eje Y: *qval* transformado a $-\log_{10}$ de la prueba de diferenciación. (D) Misma visualización que el panel A, coloreada por la expresión génica normalizada de marcadores proximales/condrocitos (*Sox6*, *Sox9*), distales (*Hoxd13*, *Tfap2b*), anteriores (*Pax9*, *Alx4*) o posteriores (*Hand2*, *Shh*). (E) Imágenes de hibridación in situ de *Hoxd13* en embriones E10.5 a E13.5. (F) Misma visualización que los paneles A y D, coloreada por la expresión génica normalizada de *Cpa2*. Su patrón de expresión dentro de esta trayectoria nos llevó a predecir que *Cpa2* es un marcador distal del mesénquima de la extremidad en desarrollo, como *Hoxd13*. (G) Imágenes de hibridación in situ de *Cpa2* en embriones E10.5 y E11.5. (H) Resumen combinado de resultados para las trayectorias del AER y del mesénquima de las extremidades.

La FIG. 21 muestra la caracterización de las trayectorias del destino celular durante el desarrollo del mesénquima de las extremidades. (A) Mapa de calor que muestra los principales genes expresados diferencialmente entre diferentes etapas de desarrollo para las células mesenquimales de las extremidades. (B) Diagrama de barras que muestra el valor *p* ajustado transformado a $-\log_{10}$ de factores de transcripción enriquecidos para genes significativamente sobreexpresados durante el desarrollo del mesénquima de las extremidades. (C) Visualización con t-SNE de células mesenquimales de las extremidades coloreadas por las extremidades anteriores (*Tbx5*+) y posteriores (*Pitx1*±). No se muestran las células sin expresión o con ambas expresiones en *Tbx5* y *Pitx1*.

La FIG. 22 muestra la expresión de marcadores espacialmente restringidos en las extremidades. Cada panel ilustra un gen marcador diferente. Los colores indican recuentos de UMI que se han escalado para el tamaño de la biblioteca, se han transformado logarítmicamente y luego se han aplicado a puntuaciones Z para permitir la comparación entre genes. Las células sin expresión de un marcador determinado se excluyen para evitar el trazado excesivo. (A) Marcador de extremidades posteriores *Pitx1* y marcador de extremidades anteriores *Tbx5*. (B) Primera fila: marcadores de extremidad proximal *Sox6* (que también marca los condrocitos) y *Sox9*. Segunda fila: marcadores de extremidad distal *Hoxd13* y *Tfap2b*. Tercera fila: Marcadores de miembro anterior *Pax9* y *Alx4*. Cuarta fila: marcadores de extremidad posterior *Shh* y *Hand2*.

La FIG. 23 muestra módulos de genes espacialmente restringidos en las extremidades. Se agruparon un total de 1.191 genes mediante agrupación jerárquica. El dendrograma se cortó en 8 módulos utilizando la función *cutree* en R y se calculó la expresión agregada de genes en cada módulo. Los colores indican recuentos UMI añadidos para cada módulo que se han escalado para el tamaño de la biblioteca, se han transformado logarítmicamente y luego se han aplicado a puntuaciones Z para permitir la comparación entre módulos. Las células sin expresión de un módulo determinado se excluyen para evitar el trazado excesivo.

La FIG. 24 muestra la caracterización de ocho trayectorias de desarrollo principales presentes durante la organogénesis del ratón. (A) Visualización con UMAP 3D de nuestro conjunto de datos general; arriba: vistas desde dos direcciones; abajo: vista ampliada de las trayectorias mesenquimales (izquierda) y del tubo neural/notocorda (derecha), coloreadas por etapa de desarrollo. (B) Mapa de calor que muestra la proporción de células de cada uno de los 38 tipos de células principales asignadas a cada una de las 8 trayectorias principales. Las columnas representan los ocho linajes principales, marcados por los colores en la barra superior (ver clave en el panel A). (C) Visualización con UMAP 3D de subtrayectorias epiteliales coloreadas por etapa de desarrollo.

La FIG. 25 muestra la caracterización de ocho trayectorias de desarrollo principales presentes durante la organogénesis del ratón. (A) Igual que la Fig. 24A, pero con colores correspondientes a los 38 grupos de células principales. (B-C) Gráfico de área que muestra la proporción estimada (B) y el número de células absoluto estimado (C) de células por embrión derivadas de cada una de las ocho trayectorias celulares principales de E9.5 a E13.5.

La FIG. 26 muestra la visualización con UMAP de las ocho trayectorias celulares principales coloreadas por la ID del grupo celular principal.

La FIG. 27 muestra la visualización con UMAP de las ocho trayectorias celulares principales coloreadas por etapa de desarrollo.

La FIG. 28 muestra la visualización con UMAP de subtipos de células epiteliales. Coloreado según los 29 subtipos epiteliales que se muestran en la Fig. 17A.

La FIG. 29 muestra la resolución de trayectorias celulares en la miogénesis. Los bordes de los gráficos principales que definen las trayectorias presentadas por Monocle 3 se muestran como segmentos de línea de color azul claro. (A) Las células supuestamente involucradas en la miogénesis se aislaron de la trayectoria de las células mesenquimales in silico y luego se usaron para construir una subtrayectoria de miocitos (Métodos). (B) Células en la subtrayectoria de miocitos coloreadas según la etapa de desarrollo. (C) Células en la trayectoria de los miocitos, coloreadas por su expresión de reguladores transcripcionales seleccionados de la miogénesis. Las células sin expresión detectable para un gen determinado se omiten de su gráfico. (D) Células clasificadas por etapa de desarrollo según los marcadores que se muestran en el panel C (Métodos).

La FIG. 30 muestra el perfil conjunto del transcriptoma total y recién sintetizado por sci-fate. (A) flujo de trabajo de sci-fate con etapas clave resumidas en el texto. (B) Esquema del experimento. Las células A549 se trataron con dexametasona en función del tiempo. Las células de todas las condiciones de tratamiento se marcaron con S4U dos horas antes de la recolección por sci-fate. (C) Diagrama de violín que muestra la relación de lecturas marcadas con S4U por célula en seis tiempos de tratamiento. (D) Diagrama de violín que muestra la relación de lecturas marcadas con S4U en lecturas exónicas e intrónicas. Para todos los diagramas de caja: líneas horizontales gruesas, medianas; bordes superior e inferior de la caja, primer y tercer cuartiles, respectivamente; bigotes, 1,5 veces el intervalo del intercuartil; círculos, valores atípicos. (E) Visualización con UMAP de células A549 mediante transcriptoma completo (izquierda), transcriptoma recién sintetizado (centro) y ambos (derecha). (F) Similar a (E), coloreado por la identificación del grupo identificado por el transcriptoma completo. (G) Visualización con UMAP de células A549 mediante información conjunta, coloreada por la expresión normalizada de genes marcadores G2/M por nivel de ARN (izquierda) y nivel de ARN recién sintetizado (derecha). Los recuentos de UMI para estos genes se escalan por el tamaño de la biblioteca, se transforman logarítmicamente, se añaden y luego se aplican a puntuaciones Z.

La FIG. 31 muestra análisis relacionados con el rendimiento y el control de calidad para Sci-Fate. (A) Diagrama de dispersión de los recuentos de UMI de ratón (NIH/3T3) frente a humano (HEK293T) por célula en la condición de sci-fate. (B-D) Diagrama de caja que muestra la relación de lecturas marcadas con S4U, el número de UMI y la pureza (proporción de lecturas aplicadas a las especies esperadas) por célula de células HEK293T (número de células $n = 932$) y NIH/3T3 (número de células $n = 438$). Para todos los diagramas de caja: líneas horizontales gruesas, medianas; bordes superior e inferior de la caja, primer y tercer cuartiles, respectivamente; bigotes, 1,5 veces el intervalo del intercuartil; círculos, valores atípicos. (E-F) Correlación (correlación de Spearman) entre mediciones de expresión génica en perfiles agregados de células HEK293T (E) y NIH/3T3 (F) de sci-fate (eje y) frente a células sci-RNA-seq (eje x).

La FIG. 32 muestra el rendimiento de sci-fate en células A549 tratadas con dexametasona. (A, B) Diagrama de violín que muestra el número de UMI (A) y genes (B) por célula en seis condiciones de tratamiento. Para todos los diagramas de caja: líneas horizontales gruesas, medianas; bordes superior e inferior de la caja, primer y tercer cuartiles, respectivamente; bigotes, 1,5 veces el intervalo del intercuartil; círculos, valores atípicos. (C) Diagrama de correlación

que muestra el coeficiente de correlación de Pearson entre diferentes condiciones de tratamiento para el transcriptoma completo añadido (arriba a la derecha) y el transcriptoma recién sintetizado (abajo a la izquierda). (D) Visualización con UMAP de células A549 mediante el transcriptoma recién sintetizado, coloreada por la identificación del grupo identificada por el transcriptoma recién sintetizado. (E) mapa de calor que muestra la proporción de células de cada grupo definido por el transcriptoma completo, que caen en cada grupo de células por el transcriptoma recién sintetizado. (F-G) Visualización con UMAP de células A549 mediante transcriptoma total y recién sintetizado, coloreada por la expresión normalizada de genes marcadores de fase S mediante expresión de ARN total (F) y ARN recién sintetizado (G). Los recuentos de UMI para estos genes se escalan para el tamaño de la biblioteca, se transforman logarítmicamente, se añaden y luego se aplican a puntuaciones Z.

La FIG. 33 muestra la caracterización de los módulos TF que impulsan la transición del estado de la célula. (A) Vínculos identificados (azul) entre factores de transcripción (naranja) y genes regulados (gris). Los módulos TF relacionados con la progresión del ciclo celular o la respuesta GR están marcados. (B) Visualización con UMAP de células A549 ordenadas por módulos TF del ciclo celular, coloreadas por ARNm recién sintetizado de marcadores de fase S y fase G2/M (arriba), tres fases del ciclo celular (abajo a la izquierda) y nueve estados del ciclo celular mediante análisis de la agrupación no supervisada (abajo a la derecha). (C) Visualización con UMAP de células A549 ordenadas por módulos TF de respuesta GR, coloreadas por el tiempo de tratamiento DEX (izquierda), la actividad CEBPB y FOXO1 (centro) y la identificación del grupo del análisis de agrupamiento no supervisado (derecha). Para calcular la actividad de TF, los recuentos de UMI recién sintetizados para estos genes se escalan para el tamaño de la biblioteca, se transforman logarítmicamente, se añaden y luego se aplican a puntuaciones Z. (D) Una tabla que muestra la relación observada (negro) del estado celular por el estado combinatorio de los módulos del ciclo celular (eje x) y los módulos de respuesta GR (eje y). El número rojo es la relación esperada suponiendo una selección independiente. (E) Mapa de calor que muestra la proporción de estados celulares definidos por los estados combinatorios de los módulos TF en cada grupo principal identificado mediante análisis de agrupamiento basado en el transcriptoma conjunto completo y recién sintetizado.

La FIG. 34 muestra módulos TF que impulsan la transición del estado de celular en células A549 tratadas con DEX. (A) Dianas génicas identificadas (gris) de CEBPB (naranja). Sólo se muestran las uniones con coeficiente de correlación regularizado de LASSO $> 0,6$. (B) Visualización con UMAP de células A549 mediante transcriptoma completo y recién sintetizado coloreado por expresión de CEBPB (izquierda) y actividad (derecha). (C) similar a (B), coloreado por la expresión YOD1 (izquierda) y la actividad YOD1 (derecha). (D) similar a (B), coloreado por la expresión GTF2IRD1 (izquierda) y la actividad GTF2IRD1 (derecha). (E) similar a (B), coloreado por la expresión de E2F1 (izquierda), la actividad de E2F1 (centro) y la expresión añadida del transcriptoma completo para genes unidos a E2F1 (derecha). (F) Mapa de calor que muestra el valor absoluto del coeficiente de correlación de Pearson entre módulos TF. 29 módulos TF se agruparon en cinco grupos mediante análisis de agrupamiento jerárquico.

La FIG. 35 muestra estados celulares caracterizados por estados combinatorios de módulos TF funcionales. (A) Esquema que muestra la estrategia para caracterizar los estados celulares mediante estados combinatorios de módulos TF funcionales. (B) Visualización con Umap de todas las células mediante el transcriptoma tanto completo como recién sintetizado, coloreado con la identificación del grupo principal identificada mediante el algoritmo de agrupamiento de picos de densidad en el espacio UMAP.

La FIG. 36 muestra la caracterización de >6.000 trayectorias de transición de estado de una sola célula. (A) Esquema que muestra la corrección de la memoria y el análisis de unión celular para construir una trayectoria de transición unicelular con detalles descritos en el texto y el método. (B) Gráfico 3D de células coloreadas por el tiempo de tratamiento DEX (también como coordenadas z). Las coordenadas x e y corresponden al espacio UMAP del transcriptoma completo y recién sintetizado en la Fig. 30E (izquierda). Las células principales y secundarias unidas están conectadas con líneas grises. (C) Similar a (B), excepto que las coordenadas x e y corresponden al espacio UMAP mediante dinámica del transcriptoma unicelular en seis puntos temporales. (D) Diagramas de líneas que muestran la dinámica del estado celular de diferentes estados de respuesta de GR (arriba) y estados del ciclo celular (abajo) en cada grupo de trayectoria celular (izquierda) o en todas las células (derecha) independientemente del análisis de unión celular. (E) Red de transición de estado celular. Los nodos son 27 estados de célula caracterizados en la Fig. 33D y las uniones son caminos de transición identificados entre estados celulares. Las uniones con probabilidades de transición bajas ($< 0,1$) se filtran. Cuadrados con líneas discontinuas que muestran los estados de ejemplo con dinámica de transición reversible. (F) Diagrama de correlación que muestra la correlación de las proporciones del estado celular entre las condiciones de tratamiento. Las correlaciones positivas se muestran en azul y las negativas en color rojo. La forma de la elipse está correlacionada con los coeficientes de correlación (en la elipse). (G) Diagrama de dispersión que muestra la correlación de las proporciones del estado celular entre los grupos de tratamiento con DEX de 10 horas observados y las proporciones del estado celular previstas. La predicción se basa en las probabilidades de transición del estado celular y la proporción del estado celular en el grupo de tratamiento sin DEX. La línea azul representa la línea de regresión lineal. (H) Diagrama de dispersión que muestra la correlación de las probabilidades de transición del estado celular calculadas por datos completos (0-10 horas) o datos parciales (0-6 horas), junto con la línea de regresión lineal.

La FIG. 37 muestra una nueva velocidad de detección de ARN y una estimación de la velocidad de degradación del ARN. (A) diagrama de dispersión que muestra la correlación entre el eje x: diferencias del transcriptoma completo

normalizado entre células sin DEX y tratadas con DEX de 2 horas, y eje y: diferencias del transcriptoma normalizado recién sintetizado entre células tratadas sin DEX y con DEX de 2 horas. La línea azul es la línea de regresión lineal. Tanto el transcriptoma completo como el transcriptoma recién sintetizado de cada punto temporal están normalizados por el tamaño de la biblioteca del transcriptoma completo de ese punto temporal. (B) Diagrama de correlación que muestra la correlación de la velocidad estimada de degradación génica entre las condiciones de tratamiento. Las correlaciones positivas se muestran en azul y las negativas en color rojo. La forma de la elipse está correlacionada con los coeficientes de correlación (en la elipse).

La FIG. 38 muestra la red de transición del estado celular para la predicción del estado de la célula. (A) Diagrama de dispersión que muestra la correlación entre los estados celulares observados en cada momento de tratamiento y el estado celular previsto por las probabilidades de transición del estado celular y la proporción del estado celular en el grupo de tratamiento sin DEX. La línea azul representa la línea de regresión lineal. (B) Diagrama de dispersión que muestra la correlación de las proporciones del estado celular entre los grupos de tratamiento con DEX de 10 horas observados y los valores previstos. Los valores previstos se basan en las probabilidades de transición del estado celular estimadas mediante datos parciales (0 a 6 horas) y la proporción del estado celular en el grupo de tratamiento sin DEX. La línea azul representa la línea de regresión lineal.

La FIG. 39 muestra que las probabilidades de transición del estado celular están reguladas por el panorama de estabilidad del estado cercano. (A) Diagrama de dispersión que muestra la relación entre la distancia de transición (distancia de Pearson) y la probabilidad de transición entre estados de célula, junto con la línea suave roja LOESS de ggplot2. (B) Gráfico 3D que muestra el panorama de inestabilidad de los estados celulares. El eje X representa los estados de respuesta de GR (estado de respuesta de nada a bajo a alto). El eje Y representa los estados del ciclo celular ordenados desde los estados G0/G1 a G2/M. El eje Z representa la inestabilidad del estado de la célula, definida por la probabilidad de que las células dentro de cada estado de la célula salten a otros estados después de 2 horas. (C) Diagrama de dispersión que muestra la relación entre la inestabilidad del estado celular y el cambio de proporción celular antes y después de tratamiento con DEX de 10 horas, junto con la línea suave roja LOESS de ggplot2. (D) Diagrama de dispersión que muestra la correlación entre la inestabilidad del estado y la entropía de transición del estado con la línea de regresión lineal (azul). (E) Diagrama de caja que muestra el r cuadrado con validación cruzada para predecir la probabilidad de transición interestatal solo por la distancia de transición o combinando la distancia de transición y el panorama de inestabilidad estatal mediante una red neuronal densamente conectada.

La FIG. 40 muestra una toma de navegador de un exón diana en el gen LMO2. El exón diana se indica en la pista "Exones diana". Tenga en cuenta que cebadores 12 RT que pasaron nuestros filtros abarcan el exón de la manera más uniforme posible. Las dos pistas inferiores, 'Cebadores_más/menos' y 'Capturado_más/menos' muestran el mapeo de lecturas de los cebadores RT y los correspondientes transcritos capturados. Una diferencia en el recuento de lecturas indica el número de eventos de cebado fuera del objetivo.

La FIG. 41 muestra una tabla que compara el rango de genes en un conjunto de datos de secuencia de ARN nuclear total ENCODE y una biblioteca de captura de RT múltiple in situ. Los genes diana se indican en negrita. De los 12 genes más abundantes en la biblioteca diana, 8 eran genes a los que nos dirigimos. Las dos últimas columnas son 'Rango', en nuestro conjunto de datos, 'ENCODE', clasificación en el conjunto de datos ENCODE. Los ARN no dirigidos abundantes incluyen los ARN ribosómicos mitocondriales RNR2 y RNR1, y el muy abundante ARNnc nuclear MALAT1. Tenga en cuenta que el gen LMO2 pasa del 4.627^o gen más detectado en el conjunto de datos ENCODE (de 26.281 genes) al 3^o más detectado en nuestra biblioteca.

La FIG. 42 muestra el enriquecimiento del transcrito capturado sobre el ARN total. Diagrama de dispersión de lecturas en una biblioteca tomada como objetivo frente a lecturas en ARN nuclear ENCODE. Los genes a los que se dirige la RT se indican en rojo; tenga en cuenta que generalmente se salen de la diagonal, lo que indica que se enriquecieron por encima del nivel esperado dada la expresión en el conjunto de datos de referencia. Los abundantes ARNnc nucleares MALAT1 y XIST se indican en azul: caen por encima de la diagonal, lo que indica un enriquecimiento exitoso contra estos ARN. Un eRNA tomado como objetivo no está enriquecido (naranja). El enriquecimiento medio por encima del nivel esperado para los 9 genes a los que se dirige este experimento es de 45,3 veces. En general, las lecturas de genes tomados como objetivo representan el 31 % del total de lecturas aplicadas a genes en este conjunto de datos.

Los dibujos esquemáticos no están necesariamente a escala. Los números iguales usados en las figuras se refieren a componentes, etapas y similares. Sin embargo, se entenderá que el uso de un número para referirse a un componente en una figura determinada no pretende limitar el componente en otra figura marcada con el mismo número. Además, el uso de números diferentes para referirse a componentes no pretende indicar que los componentes numerados diferentes no puedan ser iguales o similares a otros componentes numerados.

Descripción detallada de realizaciones ilustrativas

En una realización, el método proporcionado en la presente memoria se puede usar para producir bibliotecas de secuenciación de indexación combinatoria (sci) de células individuales que incluyen transcriptomas de una pluralidad de células individuales. Por ejemplo, el método se puede usar para obtener información de secuencia para transcriptomas de células completas, transcriptomas de ARN recién sintetizado o la combinación. En otra realización,

el método proporcionado en la presente memoria se puede usar para producir bibliotecas de secuenciación sci que incluyen información de secuencia de una subpoblación de ácidos nucleicos de ARN. Por ejemplo, cuando una región reguladora no codificante se toma como objetivo para la perturbación, una región codificante *cis* a la región reguladora se puede analizar la expresión alterada. En otro ejemplo, se pueden realizar experimentos con atlas celulares con la lectura restringida a un número limitado de ARNm que sean altamente informativos.

El método puede incluir uno o más de proporcionar núcleos o células aislados, distribuir subconjuntos de núcleos o células aislados en compartimentos, procesar los núcleos o células aislados para que incluyan fragmentos de ácido nucleico y añadir un índice específico del compartimento a los fragmentos de ácido nucleico. Opcionalmente, el método puede incluir exponer células a una condición predeterminada y/o marcar ARN recién sintetizado en las células. El método puede estar dirigido a obtener información que incluya el transcriptoma de una célula o una subpoblación de ácidos nucleicos de ARN. Estas etapas pueden ocurrir esencialmente en cualquier orden y se pueden combinar de diferentes maneras. Opcionalmente, se pueden aislar núcleos de las células después de exponer las células a una condición predeterminada y marcar el ARN recién sintetizado.

Proporcionar núcleos o células aisladas

El método proporcionado pero que no forma parte de la invención de la presente memoria puede incluir proporcionar las células o núcleos aislados de una pluralidad de células (FIG. 1, bloque 10; FIG. 2, bloque 22). Las células pueden ser de cualquier organismo y de cualquier tipo de célula o tejido del organismo. En una realización, las células pueden ser células embrionarias, p. ej., células obtenidas de un embrión. En una realización, las células o núcleos pueden ser de cáncer o de un tejido enfermo. El método puede incluir además disociar células y/o aislar los núcleos. El número de núcleos o células puede ser al menos dos. El límite superior depende de las limitaciones prácticas del equipo (p. ej., placas de múltiples pocillos, número de índices) usado en otras etapas del método como se describe en la presente memoria. El número de núcleos o células que se pueden usar no pretende ser limitante y puede ascender a miles de millones. Por ejemplo, en una realización el número de núcleos o células no puede ser mayor que 100.000.000, no mayor que 10.000.000, no mayor que 1.000.000.000, no mayor que 100.000.000, no mayor que 10.000.000, no mayor que 1.000.000, no mayor que 100.000, no mayor que 10.000, no mayor que 1.000, no mayor que 500 o no mayor que 50. El experto reconocerá que en algunas realizaciones las moléculas de ácido nucleico en cada núcleo representan el transcriptoma completo de ese núcleo, p. ej., el transcriptoma completo, el transcriptoma recién sintetizado, o ambos.

En aquellas realizaciones que usan núcleos aislados, los núcleos se pueden obtener mediante extracción y fijación. Opcional y preferiblemente, el método de obtención de núcleos aislados no incluye tratamiento enzimático. En esas realizaciones en las que se produce el transcriptoma recién sintetizado, los núcleos no se aíslan hasta después de que la célula haya sido expuesta a condiciones adecuadas para marcar los transcritos recién sintetizados.

En una realización, los núcleos se aíslan de células individuales que son adherentes o están en suspensión. Los expertos en la técnica conocen métodos para aislar núcleos de células individuales. Los núcleos suelen aislarse de las células presentes en un tejido. El método para obtener núcleos aislados incluye típicamente preparar el tejido, aislar los núcleos del tejido preparado y luego fijar los núcleos. En una realización, todas las etapas se realizan en hielo.

La preparación del tejido incluye congelar instantáneamente el tejido en nitrógeno líquido y luego reducir el tamaño del tejido a trozos de 1 mm o menos de diámetro. El tejido se puede reducir de tamaño sometiendo al tejido a un picado o una fuerza contundente. El picado se puede lograr con una cuchilla para cortar el tejido en trozos pequeños. Se puede aplicar una fuerza contundente rompiendo el tejido con un martillo o un objeto similar, y la composición resultante del tejido triturado se denomina polvo.

El aislamiento de los núcleos se puede lograr incubando los trozos o el polvo en tampón de lisis celular durante al menos 1 a 20 minutos, tal como 5, 10 o 15 minutos. Los tampones útiles son aquellos que promueven la lisis celular pero conservan la integridad de los núcleos. Un ejemplo de tampón de lisis celular incluye Tris-HCl 10 mM, pH 7,4, NaCl 10 mM, MgCl₂ 3 mM, IGEPAL CA-630 al 0,1 %, SUPERasa en inhibidor de RNasa al 1 % (20 U/μL, Ambion) y BSA al 1 % (20 mg/ml, NEB). Los métodos estándar de aislamiento de núcleos suelen usar uno o más compuestos exógenos, como enzimas exógenas, para ayudar en el aislamiento. Ejemplos de enzimas útiles que pueden estar presentes en un tampón de lisis celular incluyen, pero no se limitan a, inhibidores de proteasa, DNasa, lisozima, proteinasa K, tensioactivos, lisostafina, zimolasa, celulosa, proteasa o glucanasa, y similares (Islam et al. *Micromachines* (Basilea), 2017, 8(3):83; www.sigmaaldrich.com/life-science/biochemicals/biochemical-products.html?TablePage=14573107). En una realización, una o más enzimas exógenas no están presentes en un tampón de lisis celular útil en el método descrito en la presente memoria. Por ejemplo, una enzima exógena, (i) no se añade a las células antes de mezclar las células y el tampón de lisis, (ii) no está presente en un tampón de lisis celular antes de mezclarlo con las células, (iii) no se añade a la mezcla de células y tampón de lisis celular, o una combinación de los mismos. El experto reconocerá que estos niveles de los componentes pueden alterarse algo sin reducir la utilidad del tampón de lisis celular para aislar núcleos. Los núcleos extraídos luego se purifican mediante una o más rondas de lavado con un tampón de núcleos. Un ejemplo de tampón de núcleos incluye Tris-HCl 10 mM, pH 7,4, NaCl 10 mM, MgCl₂ 3 mM, inhibidor de SUPERasa en RNasa al 1 % (20 U/μL, Ambion) y BSA al 1 % (20 mg/ml, NEB). Al igual que un tampón de lisis celular, las enzimas exógenas también pueden estar ausentes de un tampón de núcleos

usado en un método de la presente descripción. El experto reconocerá que estos niveles de los componentes pueden alterarse algo sin reducir la utilidad del tampón de núcleos para aislar núcleos. El experto reconocerá que BSA y/o tensioactivos pueden ser útiles en los tampones usados para el aislamiento de núcleos.

Los núcleos aislados se fijan mediante exposición a un agente reticulante. Un ejemplo útil de agente reticulante incluye, pero no se limita a, paraformaldehído. El paraformaldehído puede estar en una concentración del 1 % al 8 %, tal como el 4 %. El tratamiento de núcleos con paraformaldehído puede incluir añadir paraformaldehído a una suspensión de núcleos e incubar a 0 °C. Opcional y preferentemente, a la fijación le sigue un lavado en un tampón de núcleos.

Los núcleos fijos aislados se pueden usar inmediatamente o en alícuotas y congelarse instantáneamente en nitrógeno líquido para su uso posterior. Cuando se preparan para su uso después de la congelación, los núcleos descongelados se pueden permeabilizar, por ejemplo, con tritonX-100 al 0,2 % durante 3 minutos en hielo, y sonicarse brevemente para reducir la aglomeración de los núcleos.

Las técnicas convencionales de extracción de núcleos de tejido normalmente incuban tejidos con enzima específica de tejido (p. ej., tripsina) a alta temperatura (p. ej., 37 °C) durante 30 minutos a varias horas, y luego lisan las células con tampón de lisis celular para la extracción de núcleos. El método de aislamiento de los núcleos descrito en la presente memoria tiene varias ventajas: (1) No se introducen enzimas artificiales y todas las etapas se realizan en hielo. Esto reduce la posible perturbación de los estados celulares (p. ej., el estado del transcriptoma). (2) El nuevo método ha sido validado en la mayoría de los tipos de tejidos, incluidos el cerebro, el pulmón, el riñón, el bazo, el corazón, el cerebelo y muestras de enfermedades como los tejidos tumorales. En comparación con las técnicas convencionales de extracción de núcleos de tejido que usan diferentes enzimas para diferentes tipos de tejido, la nueva técnica puede reducir potencialmente el sesgo al comparar estados celulares de diferentes tejidos. (3) El nuevo método también reduce el costo y aumenta la eficiencia al eliminar la etapa de tratamiento enzimático. (4) En comparación con otras técnicas de extracción de núcleos (p. ej., el triturador de tejido Dounce), la nueva técnica es más sólida para diferentes tipos de tejido (p. ej., el método Dounce necesita optimizar los ciclos de Dounce para diferentes tejidos) y permite procesar grandes piezas de muestras en alto rendimiento (p. ej., el método Dounce está limitado al tamaño del triturador).

Opcionalmente, los núcleos aislados pueden estar libres de nucleosomas o pueden someterse a condiciones que agoten los núcleos de nucleosomas, generando núcleos empobrecidos en nucleosomas.

Distribución de subconjuntos

El método proporcionado en la presente memoria incluye distribuir subconjuntos de núcleos o células aislados en una pluralidad de compartimentos (FIG. 1, bloque 11; FIG. 2, bloque 23; FIG. 3, bloque 32). El método puede incluir múltiples etapas de distribución, donde una población de núcleos o células aisladas (también denominadas en la presente memoria grupo) se divide en subconjuntos. Normalmente, los subconjuntos de núcleos o células aislados, p. ej., subconjuntos presentes en una pluralidad de compartimentos, se indexan con índices específicos de compartimentos y luego se agrupan. Por consiguiente, el método normalmente incluye al menos una etapa de "división y combinación" que consiste en tomar núcleos o células aisladas agrupadas y distribuirlos y añadir un índice específico del compartimento, donde el número de etapas de "división y combinación" puede depender del número de índices diferentes que se añaden a los fragmentos de ácido nucleico. Cada subconjunto inicial de núcleos o células antes de la indexación puede ser único respecto de otros subconjuntos. Por ejemplo, cada primer subconjunto puede ser de una muestra única o estar expuesto a una condición única. Después de la indexación, los subconjuntos se pueden agrupar después de la indexación, dividir en subconjuntos, indexar y agrupar nuevamente según sea necesario hasta que se añada un número suficiente de índices a los fragmentos de ácido nucleico. Este proceso asigna índices únicos o combinaciones de índices a cada célula o núcleo individual. Una vez completada la indexación, p. ej., después de añadirse uno, dos, tres o más índices, se pueden lisar los núcleos o células aislados. En algunas realizaciones, la adición de un índice y la lisis pueden ocurrir simultáneamente.

El número de núcleos o células presentes en un subconjunto, y por lo tanto en cada compartimento, puede ser al menos 1. En una realización, el número de núcleos o células presentes en un subconjunto no es superior a 100.000.000, no superior a 10.000.000, no superior a 1.000.000, no superior a 100.000, no superior a 10.000, no superior a 4.000, no superior a 3.000, no superior a 2.000, o no superior a 1.000, no superior a 500, o no superior a 50. En una realización, el número de núcleos o células presentes en un subconjunto puede ser de 1 a 1.000, de 1.000 a 10.000, de 10.000 a 100.000, de 100.000 a 1.000.000, de 1.000.000 a 10.000.000 o de 10.000.000 a 100.000.000. En una realización, el número de núcleos o células presentes en cada subconjunto es aproximadamente igual. El número de núcleos presentes en un subconjunto, y por lo tanto en cada compartimento, se basa en parte en el deseo de reducir las colisiones de índices, que es la presencia de dos núcleos o células que tienen la misma combinación de índices y terminan en el mismo compartimento en esta etapa del método. Los métodos para distribuir núcleos o células en subconjuntos son conocidos por el experto en la técnica y son rutinarios. Si bien se puede usar la citometría de clasificación de células activadas por fluorescencia (FACS), en algunas realizaciones se prefiere el uso de una dilución simple. En una realización, no se usa citometría FACS. Opcionalmente, los núcleos de diferentes ploidías pueden bloquearse y enriquecerse mediante tinción, p. ej., tinción con DAPI (4',6-diamidino-2-fenilindol). La tinción también se puede usar para discriminar células individuales de dobletes durante la clasificación.

El número de compartimentos en las etapas de distribución (y la posterior adición de un índice) puede depender del formato usado. Por ejemplo, el número de compartimentos puede ser de 2 a 96 compartimentos (cuando se usa una placa de 96 pocillos), de 2 a 384 compartimentos (cuando se usa una placa de 384 pocillos) o de 2 a 1536 compartimentos (cuando se usa una placa de 1536 pocillos). En una realización, se pueden usar múltiples placas. En una realización, cada compartimento puede ser una gotita. Cuando el tipo de compartimento usado es una gotita que contiene dos o más núcleos o células, se puede usar cualquier número de gotitas, tal como al menos 10.000, al menos 100.000, al menos 1.000.000 o al menos 10.000.000 de gotitas. Los subconjuntos de núcleos o células aislados normalmente se indexan en compartimentos antes de agruparlos.

En algunas realizaciones, el compartimento es una gotita o un pocillo. El transcriptoma, el transcriptoma recién sintetizado o su subpoblación de una célula o núcleo se pueden marcar con un índice único o una combinación de índices en una gotita o pocillo. Las bibliotecas indexadas derivadas de las particiones de gotitas o pozos se pueden agrupar para su posterior procesamiento y secuenciación. Ejemplos de tales métodos incluyen, pero no se limitan a, sistemas de análisis de células individuales de 10X genomics (Pleasanton, CA), Biorad (Hercules, CA) y CellSee (Ann Arbor, MI).

Exposición a una condición predeterminada

En una realización opcional, cada subconjunto de células está expuesto a un agente o perturbación (FIG. 1, bloque 12). Un agente puede ser esencialmente cualquier cosa que provoque un cambio en una célula. Por ejemplo, un agente puede alterar el transcriptoma de una célula, alterar la estructura de la cromatina de una célula, alterar la actividad de una proteína en la célula, alterar el ADN de una célula, alterar el estado de metilación, alterar la edición del ADN de una célula, o provocar otros cambios. Los ejemplos de agentes incluyen, pero no se limitan a, un compuesto tal como una proteína (que incluye un anticuerpo), una proteína no ribosómica, un policétido, una molécula orgánica (que incluye una molécula orgánica de 900 Daltons o menos), una molécula inorgánica, una molécula de ARN o ARNi, un carbohidrato, una glicoproteína, un ácido nucleico o una combinación de los mismos. En una realización, un agente provoca una perturbación genética, por ejemplo una proteína de edición de ADN como CRISPR o Talen. En una realización, un agente es un fármaco terapéutico. En una realización, la célula puede ser una célula de tipo salvaje, y en otra realización, la célula puede modificarse genéticamente para incluir una perturbación genética, por ejemplo, activación genética o bloqueo genético (Szlachta et al., Nat Commun., 2018, 9:4275). Se pueden exponer subconjuntos de células al mismo agente, pero se pueden alterar diferentes variables en los compartimentos, lo que permite probar múltiples variables en un solo experimento. Por ejemplo, se pueden probar diferentes dosis, diferente duración de la exposición y diferentes tipos de células en una sola placa de múltiples pocillos. En una realización, las células pueden expresar una proteína que tiene una actividad conocida y el efecto de un agente sobre la actividad se evalúa en diferentes condiciones. El uso de secuencias índice para marcar fragmentos de ácidos nucleicos permite la identificación posterior de los ácidos nucleicos que se originan en un subconjunto específico de núcleos o células, p. ej., de un pocillo de una placa de múltiples pocillos.

Etiquetado de ácidos nucleicos

En una realización opcional, se marcan ácidos nucleicos tales como ARN, ADNc o ADN producidos por una célula (FIG. 1, bloque 13). Los métodos actuales para técnicas genómicas unicelulares capturan una instantánea del estado celular, por lo que no proporcionan información sobre la dinámica de transición celular. Los inventores han descubierto que marcar el ARN recién sintetizado permite la captura tanto del transcriptoma completo como del transcriptoma recién sintetizado a nivel unicelular usando indexación por división y combinación, indexación combinatoria o cualquier método de indexación unicelular. El transcriptoma completo y el ARN recién sintetizado reciben el mismo índice único o una combinación de índices, lo que permite asignar el estado actual (p. ej., preexistente) y los ácidos nucleicos recién sintetizados a la misma célula. Esto permite la caracterización de la dinámica de transición del estado celular regulada por factores intrínsecos (p. ej., el programa del ciclo celular intrínseco de una célula) y extrínsecos (p. ej., la respuesta de una célula a un estímulo externo como un fármaco terapéutico). Además, en algunas realizaciones se permite la captura tanto del transcriptoma completo como del transcriptoma recién sintetizado a nivel unicelular, junto con la información del transcriptoma degradado de su estado pasado (memoria de estado pasado). La memoria del estado pasado de cada célula se puede corregir mediante la velocidad de degradación del ARNm (corrección de la memoria), de modo que cada célula se puede caracterizar por la dinámica del transcriptoma entre dos o más puntos temporales.

Existen varios métodos para marcar el ácido nucleico recién sintetizado de modo que pueda distinguirse del ácido nucleico existente previamente, y esencialmente se puede usar cualquier método. Normalmente, se incorpora una etiqueta a los ácidos nucleicos a medida que se sintetizan. Un tipo de método incluye la incorporación de un análogo de nucleósido que añade una mutación identificable. Por ejemplo, la adición del análogo de nucleósido 4-tiouridina (S4U) a una molécula de ARN da como resultado una mutación puntual durante una etapa de transcripción inversa para dar como resultado un ADNc de primera hebra mutado que tiene conversiones de timina a citosina (Sun y Chen, 2018, Metabolic Labeling of Newly Synthesized RNA with 4sU to in Parallel Assess RNA Transcription and Decay. En: Lamandé S. (eds) mRNA Decay. Methods in Molecular Biology, vol. 1720. Humana Press, Nueva York, NY). Esta mutación puntual se puede identificar durante las etapas de secuenciación y análisis comparando la secuencia con una referencia. Otro tipo de método incluye la incorporación de un nucleótido marcado con un hapteno que puede

usarse para purificar aquellos ARN que contienen el hapteno. Los ejemplos incluyen nucleótidos biotinilados (Luo et al., 2011, Nucl. Acids Res., 39(19):8559-8571) y nucleótidos modificados con digoxigenina (disponibles en Jena Bioscience GmbH). Un tercer tipo de método incluye la incorporación de un nucleótido que puede modificarse con una reacción química, p. ej., un nucleótido funcionalizado mediante clic, y la adición de un hapteno (Bharmal et al., 2010, J Biomol Tech., 21(3 Supl):S43, y disponible en Jena Bioscience GmbH y disponible en Thermo Fisher Scientific). Otro tipo de método incluye la incorporación de un nucleótido mutagénico tal como, pero no se limitan a, 8-oxo-dGTP y dPTP (disponibles en Jena Bioscience GmbH).

Normalmente se usan condiciones predeterminadas en una célula y no en núcleos aislados; sin embargo, el etiquetado del ácido nucleico a medida que se sintetiza se puede realizar utilizando células o núcleos aislados de las células.

En algunas realizaciones, el etiquetado puede incluir ADNc o ADN recién sintetizado. El marcado se puede usar como un identificador de una condición específica o un subconjunto de células o núcleos. Por ejemplo, se pueden usar diferentes cantidades de etiqueta, p. ej., análogo de nucleósido, nucleótido marcado con hapteno, nucleótido funcionalizado con clic y/o nucleótido mutagénico y/o diferentes relaciones entre etiquetas para marcar específicamente el ARN, ADNc o ADN de un compartimento. En otra realización, se puede agregar una etiqueta en diferentes puntos de tiempo para capturar la dimensión de tiempo. Se pueden añadir diferentes etiquetas o diferentes relaciones de etiquetas para marcar de manera diferencial el ARN en diferentes momentos. En algunas realizaciones, el etiquetado puede ser parte del esquema de indexación para resolver células individuales. Por ejemplo, una etapa de extensión puede contener un conjunto único de nucleótidos para cada compartimento. El etiquetado puede ocurrir en una etapa de transcripción inversa, etapa de extensión, hibridación o etapa de amplificación como la PCR. En algunas realizaciones, esto permite la detección de dobles o múltiplos de células o colisiones entre células.

Procesamiento para proporcionar fragmentos de ácido nucleico

En una realización, se puede usar el procesamiento de núcleos o células aislados para fragmentar ácidos nucleicos de ADN en núcleos o células aislados en fragmentos de ácidos nucleicos (FIG. 1, bloque 14). La fragmentación de ácidos nucleicos puede ser útil para obtener moléculas que tengan longitudes que sean adecuadas para la secuenciación con los métodos descritos en la presente memoria. El procesamiento puede ser necesario cuando los ácidos nucleicos diana que se van a secuenciar derivan del ADN presente en los núcleos o las células; sin embargo, en algunas realizaciones el procesamiento es opcional cuando los ácidos nucleicos diana que se van a secuenciar derivan de ARN (p. ej., ARNm y/o ARN no codificante) presente en los núcleos o células, porque en algunas realizaciones no es necesario que las moléculas de ARN estén fragmentadas. En otras realizaciones, los ácidos nucleicos derivados de moléculas de ARN están fragmentados. La fragmentación puede ocurrir en cualquier etapa del método. Por ejemplo, el método ilustrativo que se muestra en la Figura 2 incluye la fragmentación después de la adición de dos índices a las moléculas de ácido nucleico.

El procesamiento de ácidos nucleicos en núcleos o células normalmente añade una secuencia de nucleótidos a uno o ambos extremos de los fragmentos de ácido nucleico generados por el procesamiento, y la secuencia de nucleótidos puede incluir, y normalmente lo hace, una o más secuencias universales. Se puede usar una secuencia universal como, por ejemplo, una "plataforma de aterrizaje" en una etapa posterior para aparear una secuencia de nucleótidos que se puede usar como cebador para la adición de otra secuencia de nucleótidos, tal como un índice, a un fragmento de ácido nucleico. La secuencia de nucleótidos de dicho cebador puede incluir opcionalmente una secuencia índice. El procesamiento de ácidos nucleicos en núcleos o células puede añadir uno o más identificadores moleculares únicos a uno o ambos extremos de los fragmentos de ácido nucleico generados por el procesamiento.

Se conocen diversos métodos para procesar ácidos nucleicos en núcleos o células en fragmentos de ácido nucleico. Los ejemplos incluyen CRISPR y enzimas similares a Talen, y enzimas que desenrollan el ADN (p. ej., helicasas) que pueden formar regiones monocatenarias con las que los fragmentos de ADN pueden hibridarse e iniciar la extensión o amplificación. Por ejemplo, se puede usar la amplificación basada en helicasa (Vincent et al., 2004, EMBO Rep., 5(8):795-800). En una realización, la extensión o amplificación se inicia con un cebador aleatorio. En una realización, se usa un complejo de transposoma.

El complejo de transposoma es una transposasa unida a un sitio de reconocimiento de transposasa y puede insertar el sitio de reconocimiento de la transposasa en un ácido nucleico diana dentro de un núcleo en un proceso a veces denominado "tagmentación". En algunos de estos eventos de inserción, una hebra del sitio de reconocimiento de la transposasa puede transferirse al ácido nucleico diana. Dicha hebra se denomina "hebra transferida". En una realización, un complejo de transposoma incluye una transposasa dimérica que tiene dos subunidades y dos secuencias de transposones no contiguas. En otra realización, una transposasa incluye una transposasa dimérica que tiene dos subunidades y una secuencia de transposón contigua. En una realización, el extremo 5' de una o ambas hebras del sitio de reconocimiento de la transposasa puede estar fosforilado.

Algunas realizaciones pueden incluir el uso de una transposasa Tn5 hiperactiva y un sitio de reconocimiento de transposasa de tipo Tn5 (Goryshin y Reznikoff, J. Biol. Chem., 273:7367 (1998)), o transposasa MuA y un sitio de reconocimiento de transposasa Mu que comprende secuencias terminales R1 y R2 (Mizuuchi, K., Cell, 35: 785, 1983; Savilahti, H, et al., EMBO J., 14: 4893, 1995). Las secuencias del extremo de mosaico de Tn5 (ME) también se pueden

usar optimizadas por un experto en la técnica.

Más ejemplos de sistemas de transposición que se pueden usar con ciertas realizaciones de las composiciones y métodos proporcionados en la presente memoria incluyen *Staphylococcus aureus* tn552 (Colegio et al., J. Bacteriol., 183: 2384-8, 2001; Kirby C et al., Mol. Microbiol., 43: 173-86, 2002), Ty1 (Devine y Boeke, Nucleic Acids Res., 22: 3765-72, 1994 y publicación internacional WO 95/23875), transposón Tn7 (Craig, NL, Science. 271: 1512, 1996; Craig, N L, revisión en: Curr Top Microbiol Immunol., 204:27-48, 1996), Tn/O e IS10 (Kleckner N, et al., Curr Top Microbiol Immunol., 204:49-82, 1996), transposasa Mariner (Lampe D J, et al., EMBO J., 15: 5470-9, 1996), Tc1 (Plasterk R H, Curr. Topics Microbiol. Immunol., 204: 125-43, 1996), Elemento P (Gloor, G B, Methods Mol. Biol., 260: 97-114, 2004), Tn3 (Ichikawa y Ohtsubo, J Biol. Chem. 265:18829-32, 1990), secuencias de inserción bacteriana (Ohtsubo y Sekine, Curr. Top. Microbiol. Immunol. 204: 1-26, 1996), retrovirus (Brown et al., Proc Natl Acad Sci USA, 86:2525-9, 1989), y retrotransposón de levadura (Boeke y Corces, Annu Rev Microbiol. 43:403-34, 1989). Más ejemplos incluyen IS5, Tn10, Tn903, IS911 y versiones diseñadas de enzimas de la familia de las transposasas (Zhang et al., (2009) PLoS Genet. 5:e1000689. Publicación electrónica del 16 de octubre de 2009; Wilson C. et al (2007) J. Microbiol. Methods 71:332-5).

Otros ejemplos de integrasas que pueden usarse con los métodos y composiciones proporcionados en la presente memoria incluyen integrasas retrovirales y secuencias de reconocimiento de integrasas para dichas integrasas retrovirales, tales como integrasas de VIH-1, VIH-2, SIV, PFV-1, RSV.

Las secuencias de transposones útiles con los métodos y composiciones descritos en la presente memoria se proporcionan en la publicación de solicitud de patente de EE.UU. núm. 2012/0208705, publicación de solicitud de patente de EE.UU. núm. 2012/0208724 y publicación de solicitud de patente internacional núm. WO 2012/061832. En algunas realizaciones, una secuencia de transposón incluye un primer sitio de reconocimiento de transposasa y un segundo sitio de reconocimiento de transposasa. En aquellas realizaciones en las que se usa un complejo de transposoma para introducir una secuencia índice, la secuencia índice puede estar presente entre los sitios de reconocimiento de la transposasa o en el transposón.

Algunos complejos de transposomas útiles en la presente memoria incluyen una transposasa que tiene dos secuencias de transposón. En algunas de dichas realizaciones, las dos secuencias de transposón no están unidas entre sí, en otras palabras, las secuencias de transposón no son contiguas entre sí. Se conocen ejemplos de tales transposomas en la técnica (ver, por ejemplo, la publicación de solicitud de patente de EE.UU. núm. 2010/0120098).

Normalmente, la tagmentación se usa para producir fragmentos de ácido nucleico que incluyen diferentes secuencias de nucleótidos en cada extremo (p. ej., una secuencia de cebador N5 en un extremo y un cebador N7 en el otro extremo). Esto se puede lograr usando dos tipos de complejos de transposoma, donde cada complejo de transposoma incluye una secuencia de nucleótidos diferente que forma parte de la hebra transferida. En algunas realizaciones, la tagmentación usada en la presente memoria inserta una secuencia de nucleótidos en los fragmentos de ácido nucleico. La inserción de la secuencia de nucleótidos da como resultado fragmentos de ácido nucleico que tienen un dúplex de ligadura en horquilla en un extremo y la secuencia de nucleótidos insertada en el complejo de transposoma en el otro extremo. La secuencia de nucleótidos insertada en el complejo de transposoma incluye una secuencia universal. La secuencia universal sirve como secuencia complementaria para la hibridación en la etapa de amplificación descrita en la presente memoria para introducir otro índice.

En algunas realizaciones, un complejo de transposoma incluye un ácido nucleico de secuencia de transposón que se une a dos subunidades de transposasa para formar un "complejo en bucle" o un "transposoma en bucle". En un ejemplo, un transposoma incluye una transposasa dimérica y una secuencia de transposón. Los complejos en bucle pueden garantizar que los transposones se inserten en el ADN diana manteniendo al mismo tiempo la información de ordenamiento del ADN diana original y sin fragmentar el ADN diana. Como se apreciará, las estructuras en bucle pueden insertar secuencias de ácido nucleico deseadas, tales como índices, en un ácido nucleico diana, manteniendo al mismo tiempo la conectividad física del ácido nucleico diana. En algunas realizaciones, la secuencia de transposones de un complejo de transposoma en bucle puede incluir un sitio de fragmentación de modo que la secuencia de transposones pueda fragmentarse para crear un complejo de transposoma que comprenda dos secuencias de transposón. Dichos complejos de transposoma son útiles para garantizar que los fragmentos de ADN diana vecinos, en los que se insertan los transposones, reciban combinaciones de códigos de barras que puedan ensamblarse sin ambigüedades en una etapa posterior del ensayo.

En una realización, la fragmentación de ácidos nucleicos se logra usando un sitio de fragmentación presente en los ácidos nucleicos. Normalmente, los sitios de fragmentación se introducen en los ácidos nucleicos diana mediante el uso de un complejo de transposoma. En una realización, después de fragmentar los ácidos nucleicos, la transposasa permanece unida a los fragmentos de ácido nucleico, de modo que los fragmentos de ácido nucleico derivados de la misma molécula de ADN genómico permanecen unidos físicamente (Adey et al., 2014, Genome Res., 24:2041-2049). Por ejemplo, un complejo de transposoma en bucle puede incluir un sitio de fragmentación. Se puede usar un sitio de fragmentación para escindir la asociación física, pero no la informativa, entre secuencias índice que se han insertado en un ácido nucleico diana. La escisión puede realizarse por medios bioquímicos, químicos u otros medios. En algunas realizaciones, un sitio de fragmentación puede incluir un nucleótido o una secuencia de nucleótidos que puede

fragmentarse por diversos medios. Ejemplos de sitios de fragmentación incluyen, pero no se limitan a, un sitio de endonucleasa de restricción, al menos un ribonucleótido escindible con una ARNasa, análogos de nucleótidos escindibles en presencia de un determinado agente químico, una unión diol escindible mediante tratamiento con peryodato, un grupo disulfuro escindible con un agente reductor químico, un resto escindible que puede estar sujeto a escisión fotoquímica, y un péptido escindible por una enzima peptidasa u otros medios adecuados (ver, por ejemplo, la publicación de solicitud de patente de EE.UU. núm. 2012/0208705, Publicación de solicitud de patente de EE.UU. núm. 2012/0208724 y patente internacional WO 2012/061832).

Un complejo de transposoma puede incluir opcionalmente una secuencia índice, también denominada índice de transposasa. La secuencia índice está presente como parte de la secuencia del transposón. En una realización, la secuencia índice puede estar presente en una hebra transferida, la hebra del sitio de reconocimiento de transposasa que se transfiere al ácido nucleico diana.

La tagmentación de los núcleos y el procesamiento de los fragmentos de ácidos nucleicos pueden ir seguidos de un proceso de limpieza para mejorar la pureza de las moléculas. Puede usarse cualquier proceso de limpieza adecuado, tal como electroforesis, cromatografía de exclusión de tamaño o similares. En algunas realizaciones, se pueden emplear perlas paramagnéticas de inmovilización reversible en fase sólida para separar las moléculas de ADN deseadas de, por ejemplo, cebadores no incorporados, y para seleccionar ácidos nucleicos en función del tamaño. Las perlas paramagnéticas de inmovilización reversible en fase sólida están disponibles comercialmente en Beckman Coulter (Agencourt AMPure XP), Thermofisher (MagJet), Omega Biotek (Mag-Bind), Promega Beads (Promega) y Kapa Biosystems (Kapa Pure Beads).

Adición de un índice específico del compartimento

Una secuencia índice, también denominada etiqueta o código de barras, es útil como marcador característico del compartimento en el que estaba presente un ácido nucleico particular. En consecuencia, un índice es una etiqueta de secuencia de ácido nucleico que está unida a cada uno de los ácidos nucleicos diana presentes en un compartimento particular, cuya presencia es indicativa de, o se usa para identificar, el compartimento en el que estaba presente una población de núcleos o células aislados en una etapa particular del método. La adición de un índice a fragmentos de ácido nucleico se logra con subconjuntos de núcleos o células aislados distribuidos en diferentes compartimentos (FIG. 1, bloque 15; FIG. 2, bloques 24, 26 y 30; FIG. 3, bloque 33 y 37).

Una secuencia índice puede tener cualquier número adecuado de nucleótidos de longitud, p. ej., 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20 o más. Una etiqueta de cuatro nucleótidos ofrece la posibilidad de multiplexar 256 muestras en la misma matriz, y una etiqueta de seis bases permite procesar 4096 muestras en la misma matriz.

En una realización, la adición de un índice se logra durante el procesamiento de ácidos nucleicos en fragmentos de ácidos nucleicos. Por ejemplo, se puede usar un complejo de transposoma que incluya un índice. En algunas realizaciones, se añade un índice después de que el procesamiento genera fragmentos de ácido nucleico que contienen una secuencia de nucleótidos en uno o ambos extremos. En otras realizaciones, no es necesario el procesamiento para añadir un índice. Por ejemplo, se puede añadir un índice directamente a los ácidos nucleicos de ARN sin fragmentar los ácidos nucleicos de ARN. Por consiguiente, la referencia a "fragmento de ácido nucleico" incluye ácidos nucleicos que resultan del procesamiento y ácidos nucleicos de ARN, y los ácidos nucleicos derivados de estos ácidos nucleicos.

Los métodos para añadir un índice incluyen, pero no se limitan a, ligadura, extensión (incluida la extensión usando transcriptasa inversa), hibridación, adsorción, interacciones específicas o no específicas de un cebador, amplificación o transposición. La secuencia de nucleótidos que se añade a uno o ambos extremos de los fragmentos de ácido nucleico también puede incluir una o más secuencias universales y/o identificadores moleculares únicos. Se puede usar una secuencia universal como, por ejemplo, una "plataforma de aterrizaje" en una etapa posterior para aparear una secuencia de nucleótidos que se puede usar como cebador para la adición de otra secuencia de nucleótidos, tal como otro índice y/u otra secuencia universal, a un fragmento de ácido nucleico. Por consiguiente, la incorporación de una secuencia índice puede usar un proceso que incluye uno, dos o más etapas, usando esencialmente cualquier combinación de ligadura, extensión, hibridación, adsorción, interacciones específicas o no específicas de un cebador, amplificación o transposición.

Por ejemplo, en realizaciones que incluyen el uso de fragmentos de ácido nucleico que se derivan de ARNm, se pueden usar diversos métodos para añadir un índice al ARNm en una o dos etapas. Por ejemplo, se puede añadir un índice usando los tipos de métodos usados para producir ADNc. Un cebador con una secuencia poli-T en el extremo 3' puede aparearse con moléculas de ARNm y extenderse usando una transcriptasa inversa. La exposición de los núcleos o células aislados a estos componentes en condiciones adecuadas para la transcripción inversa da como resultado una adición en una sola etapa del índice para dar como resultado una población de núcleos o células indexados, donde cada núcleo o célula contiene fragmentos de ácido nucleico indexados. Alternativamente, el cebador con una secuencia poli-T incluye una secuencia universal en lugar de un índice, y el índice se añade mediante una etapa posterior de ligadura, extensión del cebador, amplificación, hibridación o una combinación de las mismas. En algunas realizaciones, el código de barras se añade sin el uso de una secuencia universal. Los fragmentos de ácido

nucleico indexados pueden incluir, y normalmente lo hacen, en la hebra sintetizada la secuencia índice indicativa del compartimento particular.

En realizaciones que incluyen el uso de fragmentos de ácido nucleico derivados de ARN no codificante, se pueden usar varios métodos para añadir un índice al ARN no codificante en una o dos etapas. Por ejemplo, se puede añadir un índice usando un primer cebador que incluye una secuencia aleatoria y un cebador de cambio de plantilla, donde cualquiera de los cebadores puede incluir un índice. Se puede usar una transcriptasa inversa que tenga una actividad transferasa terminal para dar como resultado la adición de nucleótidos que no son de la plantilla al extremo 3' de la hebra sintetizada, y el cebador de cambio de plantilla incluye nucleótidos que se aparean con los nucleótidos que no son de la plantilla añadidos mediante la transcriptasa inversa. Un ejemplo de una enzima transcriptasa inversa útil es la transcriptasa inversa del virus de la leucemia murina de Moloney. En una realización particular, se usa el reactivo SMARTer™ disponible de Takara Bio USA, Inc. (Núm. de cat. 634926) para el uso del cambio de plantilla para añadir un índice al ARN no codificante y al ARNm si se desea. Alternativamente, el primer cebador y/o el cebador de cambio de plantilla pueden incluir una secuencia universal en lugar de un índice, y el índice se añade mediante una etapa posterior de ligadura, extensión del cebador, amplificación, hibridación o una combinación de los mismos. Los fragmentos de ácido nucleico indexados pueden incluir, y normalmente lo hacen, en la hebra sintetizada la secuencia índice indicativa del compartimento particular. Otras realizaciones incluyen perfiles 5' o 3' de ARN o perfiles de ARN de longitud completa.

En otra realización, se puede seleccionar como objetivo un ARNm específico y/o un ARN no codificante para la amplificación. La focalización permite la producción de bibliotecas de secuenciación enriquecidas para secuencias que tienen más probabilidades de proporcionar información útil, de dar como resultado una gran reducción en la profundidad de secuenciación y los costos asociados, y de aumentar el poder para detectar diferencias sutiles entre células. Se pueden seleccionar moléculas de ARN que incluyen uno o más ARNm y/o uno o más ARN no codificantes como probables para proporcionar información útil, y se pueden usar cebadores para aparear selectivamente los ácidos nucleicos de ARN predeterminados y amplificar una subpoblación de las moléculas de ARN totales presentes en una célula o núcleo. El experto reconocerá que las moléculas de ARN apropiadas para seleccionar dependen del experimento. Por ejemplo, en la evaluación de perturbaciones no codificantes, solo las regiones codificantes *cis* al elemento regulador que se está alterando se pueden probar para cambios en la expresión. Este enfoque puede reducir el fondo de las lecturas ribosómicas más que el uso de cebadores hexámeros aleatorios o poli-T. Este enfoque también permite apuntar a las uniones de empalme y a los exones resultantes de eventos alternativos en el sitio de inicio de la transcripción, proporcionando así información de isoformas que no se detecta fácilmente con los métodos *sci* convencionales.

La amplificación dirigida de moléculas de ARN puede ocurrir en varias etapas durante la producción de la biblioteca. En una realización, la amplificación dirigida de múltiples dianas se produce durante la transcripción inversa de moléculas de ARN. Un experimento puede incluir múltiples cebadores diferentes dirigidos a diferentes moléculas de ARN. En una realización, se pueden usar múltiples cebadores dirigidos a diferentes regiones de la misma molécula de ARN. El uso de múltiples cebadores dirigidos a diferentes regiones de la misma molécula de ARN permite múltiples oportunidades para que la molécula de ARN se transcriba de manera inversa en ADNc, lo que aumenta la probabilidad de detección de la molécula de ARN.

En una realización, los cebadores usados para la amplificación dirigida no incluyen un índice. Cuando no se añade un índice durante la reacción de amplificación, no es necesaria la distribución de células o núcleos en diferentes compartimentos, y la amplificación puede ocurrir como una reacción única con todas las moléculas de ARN y todos los cebadores presentes. En realizaciones en las que se añade un índice durante la reacción de amplificación, la distribución de las células o núcleos es útil, y la amplificación puede ocurrir como una única reacción en cada compartimento con todas las moléculas de ARN y todos los cebadores presentes, pero cada cebador en un compartimento tiene el mismo índice específico del compartimento.

En una realización, el diseño de cebadores para la captura múltiple de dianas puede incluir una o más de las siguientes consideraciones. Después de seleccionar un ARN para la amplificación dirigida, se puede recoger la secuencia del ARN y determinar todos los cebadores de transcriptasa inversa posibles - los cebadores candidatos. La longitud de cualquier cebador debe ser lo suficientemente larga para funcionar en una reacción de transcripción inversa y puede tener, por ejemplo, entre 20 y 30 nucleótidos de longitud.

Los cebadores candidatos se pueden filtrar según varios criterios, que incluyen, pero no se limitan a, el contenido de GC, la posición de las bases de GC en el cebador, la probabilidad de orientación fuera del sitio y la capacidad de mapeo. Un contenido de GC útil es del 40-60 %, correspondiente a temperaturas de fusión que están aproximadamente entre 55 y 70 °C. Se prefiere tener dos bases de guanina o citosina en los últimos 5 nucleótidos del extremo 3' del cebador para aumentar la probabilidad de que el cebador apareado sea un buen sustrato para la extensión mediante la enzima transcriptasa inversa.

Con respecto a la probabilidad de un cebado fuera del objetivo, los inventores descubrieron que, si bien los ARN diana estaban muy enriquecidos, una gran fracción de las lecturas todavía derivaban de otros ARN que abundaban dentro de las células. La mayoría de estos eventos de cebado fuera del objetivo fueron el resultado de aproximadamente 5 a

8 pares de bases de complementariedad entre el extremo 3' del cebador y el ARN fuera del objetivo. Los inventores encontraron útil considerar la abundancia del hexámero final del cebador candidato dentro del ARN celular total. Se determinó que los cebadores útiles incluían un último hexámero que (i) o no estaba presente en el ARN ribosomal o (ii) estaba representado en un nivel bajo dentro del ARN celular total.

5 Se describen ejemplos de hexámeros no presentes dentro del ARN ribosómico (los hexámeros "no tan aleatorios" o NSR de Armour et al., 2009, Nature Methods, 6(9):647-49). Se descubrió que era mucho menos probable que los cebadores que tenían esta característica tuvieran un cebado fuera del objetivo dentro del ARN ribosómico. Un método para determinar si un hexámero está representado en un nivel bajo dentro del ARN celular total puede incluir identificar la abundancia de cada hexámero en las moléculas de ARN dentro de una célula, por ejemplo, toda la transcripción
10 naciente, incluida la transcripción ribosómica, dentro del tipo de célula que se va a analizar según los métodos descritos en la presente memoria. El uso de cebadores candidatos que se encuentran en un nivel bajo de abundancia, p. ej., dentro del cuartil más bajo de abundancia, puede reducir el direccionamiento fuera del sitio.

Los cebadores candidatos también pueden evaluarse mediante mapeabilidad. Por ejemplo, cada candidato se puede alinear con las dianas usando un algoritmo tipo pajarita y permitiendo 3 desajustes. Esta etapa ayuda a garantizar que
15 cada cebador tendrá solo un sitio diana en el genoma.

En algunas realizaciones, la amplificación de múltiples dianas en la misma reacción, también denominada como captura múltiple de dianas, el control de las temperaturas de apareamiento de cebadores de transcriptasa inversa es útil para mantener la transcripción inversa específica y la amplificación de los ARN diana deseados. Por ejemplo, los protocolos típicos de transcripción inversa desnaturalizan una mezcla de ARN y cebador de transcripción inversa y se
20 enfrían a 4 °C para permitir el apareamiento. Una temperatura de apareamiento baja es demasiado permisiva y da como resultado eventos de apareamiento indeseables y fuera del objetivo. Para aumentar la probabilidad de que los únicos eventos de apareamiento que se extienden sean aquellos en los que todos los cebadores de transcripción inversa específicos se aparean con las dianas correctas, se mantiene una temperatura alta durante todo el proceso de transcripción inversa. En una realización, los componentes - p. ej., mezcla de células fijas, conjunto de cebadores
25 de transcripción inversa y dNTP - a 65 °C, se aparean a 53 °C, se añade una mezcla de enzima de transcripción inversa/tampón que está preequilibrada a 53 °C a la reacción de apareamiento, y se extienden a 53 °C durante 20 minutos. Por lo tanto, se reduce la posibilidad de que los cebadores de transcripción inversa se apareen a baja temperatura entre las etapas de desnaturalización y extensión. El experto reconocerá que se pueden realizar algunas modificaciones, por ejemplo alterando la temperatura o el tiempo, sin reducir la especificidad de la transcripción
30 inversa.

Se pueden usar otros métodos para la adición de un índice a un fragmento de ácido nucleico, y la forma en que se añade un índice no pretende ser limitante. Por ejemplo, en una realización la incorporación de una secuencia índice incluye ligar un cebador a uno o ambos extremos de los fragmentos de ácido nucleico. La ligadura del cebador de
35 ligadura puede verse favorecida por la presencia de una secuencia universal en los extremos de los fragmentos de ácido nucleico. Un ejemplo de cebador es un dúplex de ligadura en horquilla. El dúplex de ligadura puede ligarse a un extremo o preferiblemente a ambos extremos de fragmentos de ácido nucleico.

En otra realización, la incorporación de una secuencia índice incluye el uso de fragmentos de ácido nucleico monocatenario y la síntesis de la segunda hebra de ADN. En una realización, la segunda hebra de ADN se produce usando un cebador que incluye secuencias complementarias a los nucleótidos presentes en los extremos de los
40 fragmentos de ácido nucleico monocatenario.

En otra realización, la incorporación de un índice se produce en una, dos, tres o más rondas de códigos de barras divididos y agrupados, lo que da como resultado bibliotecas de células individuales indexadas únicas, dobles, triples o múltiples (p. ej., cuatro o más).

En otra realización, la incorporación de índices y mediador de amplificación (p. ej., una secuencia universal) es
45 beneficiosa, ya que permite bibliotecas de secuenciación de células individuales dirigidas y/o preparar bibliotecas de secuenciación de células individuales dirigidas.

Adición de secuencias universales para la inmovilización

En una realización, la adición de nucleótidos durante las etapas de procesamiento y/o indexación añade secuencias universales útiles en la inmovilización y secuenciación de los fragmentos. En otra realización, los fragmentos de ácido
50 nucleico indexados se pueden procesar adicionalmente para añadir secuencias universales útiles para inmovilizar y secuenciar los fragmentos de ácido nucleico. El experto reconocerá que en las realizaciones donde el compartimento es una gotita, las secuencias para inmovilizar fragmentos de ácido nucleico son opcionales. En una realización, la incorporación de secuencias universales útiles para inmovilizar y secuenciar los fragmentos incluye ligar adaptadores universales idénticos (también denominados 'adaptadores no coincidentes', cuyas características generales se describen en Gormley et al., documento US 7.741.463, y Bignell et al., documento US 8.053.192) a los extremos 5' y
55 3' de los fragmentos de ácido nucleico indexados. En una realización, el adaptador universal incluye todas las secuencias necesarias para la secuenciación, incluidas secuencias para inmovilizar los fragmentos de ácido nucleico

indexados en una matriz.

En una realización, se puede usar la ligadura de extremos romos. En otra realización, los fragmentos de ácido nucleico se preparan con nucleótidos sobresalientes únicos mediante, por ejemplo, la actividad de ciertos tipos de ADN polimerasa tal como Taq polimerasa o Klenow exo minus polimerasa que tiene una actividad transferasa terminal no dependiente de la plantilla que añade uno o más desoxinucleótidos, por ejemplo, desoxiadenosina (A) a los extremos 3' de los fragmentos de ácido nucleico indexados. En algunos casos, el nucleótido sobresaliente es más de una base. Dichas enzimas se pueden usar para añadir un único nucleótido 'A' al extremo romo 3' de cada hebra de los fragmentos de ácido nucleico. Por tanto, se podría añadir una 'A' al extremo 3' de cada hebra de los fragmentos diana bicatenarios mediante reacción con Taq o Klenow exo minus polimerasa, mientras que las secuencias adicionales que se añadirán a cada extremo del fragmento de ácido nucleico pueden incluir un saliente 'T' compatible presente en el extremo 3' de cada región de ácido nucleico bicatenario que se va a añadir. Esta modificación del extremo también evita la autoligadura de los ácidos nucleicos de modo que existe una tendencia hacia la formación de los fragmentos de ácido nucleico indexados flanqueados por las secuencias que se añaden en esta realización.

En otra realización, cuando el adaptador universal ligado a los fragmentos de ácido nucleico indexados no incluye todas las secuencias necesarias para la secuenciación, entonces se puede usar una etapa de amplificación, tal como PCR, para modificar aún más los adaptadores universales presentes en cada fragmento de ácido nucleico indexado antes para inmovilizar y secuenciar. Por ejemplo, se puede llevar a cabo una reacción de extensión del cebador inicial usando una secuencia de anclaje universal complementaria a una secuencia universal presente en el fragmento de ácido nucleico indexado, en la que se forman productos de extensión complementarios a ambas hebras de cada fragmento de ácido nucleico indexado individual. Normalmente, la PCR añade secuencias universales adicionales, como una secuencia de captura universal.

Después de añadir los adaptadores universales, ya sea mediante un método de una sola etapa de ligadura o hibridación de un adaptador universal que incluye todas las secuencias necesarias para la secuenciación, o mediante un método de dos etapas de ligadura de un adaptador universal y luego una amplificación para modificar aún más el adaptador universal, los fragmentos de índice finales incluirán una secuencia de captura universal y una secuencia de anclaje. El resultado de añadir adaptadores universales a cada extremo es una pluralidad o biblioteca de fragmentos de ácido nucleico indexados.

Los fragmentos indexados resultantes proporcionan colectivamente una biblioteca de ácidos nucleicos que pueden inmovilizarse y luego secuenciarse. El término biblioteca, también denominado en la presente memoria como biblioteca de secuenciación, se refiere a la colección de fragmentos de ácido nucleico de núcleos o células individuales que contienen secuencias universales conocidas en sus extremos 3' y 5'. La biblioteca incluye ácidos nucleicos del transcriptoma completo, ácidos nucleicos de moléculas de ARN recién sintetizadas o una combinación de ambos, y puede usarse para realizar la secuenciación del transcriptoma completo, el transcriptoma del ARN recién sintetizado o una combinación de ambos.

Los fragmentos de ácido nucleico indexados pueden someterse a condiciones que seleccionan un intervalo de tamaño predeterminado, tal como de 150 a 400 nucleótidos de longitud, tal como de 150 a 300 nucleótidos. Los fragmentos de ácido nucleico indexados resultantes se combinan y, opcionalmente, se pueden someter a un proceso de limpieza para mejorar la pureza de las moléculas de ADN mediante la eliminación de al menos una porción de adaptadores o cebadores universales no incorporados. Puede usarse cualquier proceso de limpieza adecuado, tal como electroforesis, cromatografía de exclusión por tamaño o similares. En algunas realizaciones, se pueden emplear perlas paramagnéticas de inmovilización reversible en fase sólida para separar las moléculas de ADN deseadas de adaptadores o cebadores universales no unidos y para seleccionar ácidos nucleicos en función del tamaño. Las perlas paramagnéticas de inmovilización reversible en fase sólida están disponibles comercialmente en Beckman Coulter (Agencourt AMPure XP), Thermofisher (MagJet), Omega Biotek (Mag-Bind), Promega Beads (Promega) y Kapa Biosystems (Kapa Pure Beads).

Una realización ilustrativa no limitante de la presente descripción se muestra en la FIG. 1. En esta realización, el método incluye proporcionar una pluralidad de células (FIG. 1, bloque 10). El método incluye además distribuir subconjuntos de células en una pluralidad de compartimentos (FIG. 1, bloque 11) y exponer las células a una condición predeterminada (FIG. 1, bloque 12). Las condiciones predeterminadas pueden variar, y normalmente lo hacen, entre diferentes compartimentos. Por ejemplo, diferentes compartimentos pueden incluir diferentes dosis de un agente, diferentes perturbaciones, diferentes duraciones de exposición, diferentes tipos de células y similares. A continuación, se marca el ARN recién sintetizado (Figura 1, bloque 13). El etiquetado puede ocurrir con células, o se pueden aislar núcleos de las células y el etiquetado ocurre con núcleos. En algunas realizaciones, el ARN celular se procesa para dar como resultado fragmentos (FIG. 1, bloque 14). A continuación, se indexa el ARN presente en las células o en los núcleos (figura 1, bloque 15). Son posibles varias realizaciones para indexar ARN. Por ejemplo, en una realización, todo el ARNm presente en una célula se indexa usando un cebador que incluye una región poli-T. En otra realización, se pueden indexar ácidos nucleicos de ARN específicos.

Otra realización ilustrativa no limitante de la presente descripción se muestra en la FIG. 2 y se describe en el Ejemplo 1. En esta realización, el método incluye proporcionar núcleos aislados de una pluralidad de células (FIG. 2, bloque

22). El método incluye además distribuir subconjuntos de los núcleos aislados en una primera pluralidad de compartimentos (FIG. 2, bloque 23). El número de compartimentos en la primera etapa de distribución (FIG. 2, bloque 23) puede depender del formato utilizado. Por ejemplo, el número de compartimentos puede ser de 2 a 96 compartimentos (cuando se usa una placa de 96 pocillos), de 2 a 384 compartimentos (cuando se usa una placa de 384 pocillos) o de 2 a 1536 compartimentos (cuando se usa una placa de 1536 pocillos). Alternativamente, se pueden usar otros compartimentos, como por ejemplo gotitas.

El método también incluye generar núcleos indexados (FIG. 2, bloque 24). En una realización, generar núcleos indexados incluye el uso de transcriptasa inversa con un cebador oligo-dT para añadir un índice, una secuencia de nucleótidos aleatoria y una secuencia universal. El índice en cada compartimento es único, p. ej., cada índice es específico del compartimento. La secuencia aleatoria se usa como un identificador molecular único (UMI) para marcar fragmentos de ácidos nucleicos únicos. La secuencia aleatoria también se puede usar para ayudar en la eliminación de duplicados en el procesamiento posterior. La secuencia universal sirve como secuencia complementaria para la hibridación en la etapa de ligadura descrita en la presente memoria. En otra realización, generar núcleos indexados incluye el uso de transcriptasa inversa con cebadores específicos para apuntar a moléculas de ARN predeterminadas. La transcripción inversa puede dar como resultado la adición de un índice, una secuencia de nucleótidos aleatoria y una secuencia universal a las moléculas de ARN diana. La exposición de los núcleos a estos componentes en condiciones adecuadas para la transcripción inversa da como resultado una población de núcleos indexados, donde cada núcleo contiene fragmentos de ácido nucleico indexados. Los fragmentos de ácido nucleico indexados pueden incluir, y normalmente lo hacen, en la hebra sintetizada la secuencia índice indicativa del compartimento particular. Un ejemplo de un fragmento de ácido nucleico indexado se muestra en la FIG. 1A del Ejemplo 1 (ver "Transcripción inversa indexada").

Los núcleos indexados de múltiples compartimentos se pueden combinar (FIG. 2, bloque 25). Los subconjuntos de estos núcleos indexados combinados, denominados en la presente memoria como núcleos indexados agrupados, se distribuyen entonces en una segunda pluralidad de compartimentos (FIG. 2, bloque 25). La distribución de núcleos en subconjuntos va seguida por la incorporación en los fragmentos de ácido nucleico indexados en cada compartimento de una segunda secuencia índice para generar fragmentos de doble índice. Esto da como resultado una indexación adicional de los fragmentos de ácido nucleico indexados (Figura 2, bloque 26).

En esta realización ilustrativa, la incorporación de la segunda secuencia índice incluye ligar un dúplex de ligadura en horquilla a los fragmentos de ácido nucleico indexados en cada compartimento. El uso de un dúplex de ligadura en horquilla para introducir una secuencia universal, un índice o una combinación de los mismos en el extremo de un fragmento de ácido nucleico diana normalmente usa un extremo del dúplex como cebador para una amplificación posterior. Por el contrario, un dúplex de ligadura en horquilla usado en esta realización no actúa como cebador. Una ventaja de usar un dúplex de ligadura en horquilla descrito en la presente memoria es una reducción de la autoligadura observada con muchos dúplex de ligadura en horquilla descritos en la técnica. En una realización, el dúplex de ligadura incluye cinco elementos: 1) una secuencia universal que es un complemento de la secuencia universal presente en el cebador oligo-dT, 2) un segundo índice, 3) un ideoxiU, 4) una secuencia de nucleótidos que puede formar una horquilla, y 5) el complemento inverso del segundo índice. Las segundas secuencias índice son únicas para cada compartimento en el que se colocaron los núcleos indexados distribuidos (FIG. 2, bloque 25) después de añadirse el primer índice mediante transcripción inversa. Un ejemplo de un fragmento de ácido nucleico de doble índice se muestra en la FIG. 1A del Ejemplo 1 (ver "Ligadura en horquilla indexada").

La eliminación del ideoxiU presente en la región en horquilla del dúplex de ligadura en horquilla incorporado en los fragmentos de ácido nucleico puede ocurrir antes, durante o después de la limpieza. La eliminación del residuo de uracilo se puede lograr mediante cualquier método disponible y, en una realización en cono, se usa el reactivo de escisión específico de uracilo (USER) disponible en NEB.

Los subconjuntos de estos núcleos de doble índice combinados, denominados en la presente memoria núcleos de doble índice agrupados, se distribuyen luego en una tercera pluralidad de compartimentos (FIG. 2, bloque 27). En una realización, se distribuyen de 100 a 30.000 núcleos en cada pocillo. En una realización, el número de núcleos en un pocillo es al menos 100, al menos 500, al menos 1.000 o al menos 5.000. En una realización, el número de núcleos en un pocillo no es mayor que 30.000, no mayor que 25.000, no mayor que 20.000 o no mayor que 15.000. En una realización, el número de núcleos presentes en un subconjunto puede ser de 100 a 1.000, de 1.000 a 10.000, de 10.000 a 20.000 o de 20.000 a 30.000. En una realización, se distribuyen 2.500 núcleos en cada pocillo. En una realización, el número de núcleos presentes en cada subconjunto es aproximadamente igual.

La distribución de núcleos de doble índice en subconjuntos va seguida por la síntesis de la segunda hebra de ADN (FIG. 2, bloque 28). Los ácidos nucleicos en los núcleos se procesan someténdolos a tagmentación (FIG. 2, bloque 29). Cada compartimento que contiene los núcleos de doble índice incluye un complejo de transposoma. En esta realización, la tagmentación se usa para producir fragmentos de ácido nucleico que incluyen diferentes secuencias de nucleótidos en cada extremo (p. ej., una secuencia de cebador N5 en un extremo y un cebador N7 en el otro extremo).

La tagmentación de núcleos va seguida por la incorporación en los fragmentos de ácido nucleico de doble índice en cada compartimento una tercera secuencia índice para generar fragmentos de índice triple, donde la tercera secuencia

índice en cada compartimento es diferente de la primera y segunda secuencias índice en los compartimentos. Esto da como resultado la indexación adicional de los fragmentos de ácido nucleico indexados (FIG. 2, bloque 30; véase también la FIG. 1A del Ejemplo ("tratamiento del USER, PCR indexada")) antes de la inmovilización y secuenciación. En una realización, las secuencias universales presentes en los extremos de los fragmentos de ácido nucleico de doble índice (p. ej., la secuencia de nucleótidos insertados dúplex de ligadura en horquilla en un extremo y la secuencia de nucleótidos insertada-complejo de transposoma en el otro extremo) se pueden usar para la unión de cebadores y extenderse en una reacción de amplificación. Normalmente se usan dos cebadores diferentes. Un cebador se hibrida con secuencias universales en el extremo 3' de una hebra de los fragmentos de ácido nucleico de doble índice, y un segundo cebador se hibrida con secuencias universales en el extremo 3' de la otra hebra de los fragmentos de ácido nucleico de doble índice. Por tanto, la secuencia de anclaje (p. ej., el sitio al que se aparea un cebador universal tal como un cebador de secuenciación para la lectura 1 o la lectura 2 para la secuenciación) presente en cada cebador puede ser diferente. Cada uno de los cebadores adecuados puede incluir secuencias universales adicionales, tales como una secuencia de captura universal (p. ej., el sitio al que se hibrida un oligonucleótido de captura, donde el oligonucleótido de captura puede inmovilizarse sobre una superficie de un sustrato sólido). Debido a que cada cebador incluye un índice, esta etapa da como resultado la adición de otra secuencia índice, una en cada extremo de los fragmentos de ácido nucleico para dar como resultado fragmentos con índice triple. En una realización, se pueden usar cebadores indexados, tales como un cebador P5 indexado y un cebador P7 indexado, para añadir el tercer índice. Los fragmentos de triple índice se agrupan y pueden someterse a una etapa de limpieza como se describe en la presente memoria.

La pluralidad de fragmentos de triple índice se puede preparar para la secuenciación. Después de agrupar los fragmentos de triple índice y someterlos a limpieza, se enriquecen, normalmente mediante inmovilización y/o amplificación, antes de la secuenciación (FIG. 2, bloque 31).

Otra realización ilustrativa no limitante de la presente descripción se muestra en la FIG. 3 y se describe en el Ejemplo 4. En esta realización, el método incluye proporcionar núcleos o células aislados (FIG. 3, bloque 30). Los núcleos o células pueden exponerse a transcriptasa inversa y cebadores específicos para dirigirse y enriquecer moléculas de ARN predeterminadas (FIG. 3, bloque 31). La exposición de los núcleos o células a estos componentes en condiciones adecuadas para la transcripción inversa da como resultado una población de núcleos o células, donde cada núcleo contiene fragmentos de ácido nucleico que están enriquecidos en secuencias presentes en las moléculas de ARN predeterminadas. El método incluye además distribuir subconjuntos de núcleos o células en una primera pluralidad de compartimentos (FIG. 3, bloque 32). El número de compartimentos en la primera etapa de distribución (FIG. 3, bloque 32) puede depender del formato usado. Por ejemplo, el número de compartimentos puede ser de 2 a 96 compartimentos (cuando se usa una placa de 96 pocillos), de 2 a 384 compartimentos (cuando se usa una placa de 384 pocillos) o de 2 a 1536 compartimentos (cuando se usa una placa de 1536 pocillos). Alternativamente, se pueden usar otros compartimentos, tal como gotitas.

El método también incluye generar núcleos o células indexadas (FIG. 3, bloque 33) incorporando en los fragmentos de ácido nucleico en cada compartimento una secuencia índice para generar fragmentos indexados.

En una realización, la incorporación de la secuencia índice incluye ligar un dúplex de ligadura en horquilla a los fragmentos de ácido nucleico indexados en cada compartimento. Los núcleos o células que contienen los fragmentos indexados se agrupan y luego se distribuyen subconjuntos de estos núcleos o células indexados combinados en una segunda pluralidad de compartimentos (FIG. 3, bloque 34).

La distribución de núcleos o células indexados en subconjuntos puede ir seguida por la síntesis de la segunda hebra de ADN (FIG. 3, bloque 35). Los ácidos nucleicos en los núcleos o células se procesan sometiéndolos a tagmentación (FIG. 3, bloque 36). Cada compartimento que contiene los núcleos indexados incluye un complejo de transposoma. En esta realización, la tagmentación se usa para producir fragmentos de ácido nucleico que incluyen diferentes secuencias de nucleótidos en cada extremo (p. ej., una secuencia de cebador N5 en un extremo y un cebador N7 en el otro extremo).

La tagmentación de núcleos puede ir seguida por la incorporación en los fragmentos de ácido nucleico indexados en cada compartimento de una segunda secuencia índice para generar fragmentos de doble índice, donde la segunda secuencia índice en cada compartimento es diferente de las primeras secuencias índice en los compartimentos. Esto da como resultado la indexación adicional de los fragmentos de ácido nucleico indexados (FIG. 3, bloque 37) antes de la inmovilización y secuenciación.

La pluralidad de fragmentos de doble índice se puede preparar para la secuenciación, donde los datos de secuenciación se enriquecen con secuencias presentes en las moléculas de ARN predeterminadas. Después de agrupar los fragmentos de doble índice y someterlos a limpieza, se enriquecen, normalmente mediante inmovilización y/o amplificación, antes de la secuenciación (FIG. 3, bloque 38).

Preparación de muestras inmovilizadas para la secuenciación

En la técnica se conocen métodos para unir fragmentos indexados de una o más fuentes a un sustrato. En una realización, los fragmentos indexados se enriquecen usando una pluralidad de oligonucleótidos de captura que tienen especificidad por los fragmentos indexados, y los oligonucleótidos de captura se pueden inmovilizar sobre una superficie de un sustrato sólido. Por ejemplo, los oligonucleótidos de captura pueden incluir un primer miembro de un par de unión universal, y en donde un segundo miembro del par de unión está inmovilizado sobre una superficie de un sustrato sólido. Asimismo, los métodos para amplificar fragmentos de doble índice inmovilizados incluyen, pero no se limitan a, amplificación en puente y exclusión cinética. Los métodos para la inmovilización y amplificación antes de la secuenciación se describen, por ejemplo, en Bignell et al. (documento US 8.053.192), Gunderson et al. (patente internacional WO2016/130704), Shen et al. (documento US 8.895.249), y Pipenburg et al. (documento US 9.309.502).

Se puede inmovilizar una muestra agrupada en la preparación para la secuenciación. La secuenciación se puede realizar como una serie de moléculas individuales o se puede amplificar antes de la secuenciación. La amplificación se puede llevar a cabo usando uno o más cebadores inmovilizados. El o los cebadores inmovilizados pueden ser, por ejemplo, un césped sobre una superficie plana o sobre un grupo de perlas. El grupo de perlas se puede aislar en una emulsión con una sola perla en cada "compartimento" de la emulsión. A una concentración de sólo una plantilla por "compartimento", sólo se amplifica una única plantilla en cada perla.

El término "amplificación en fase sólida" como se usa en la presente memoria se refiere a cualquier reacción de amplificación de ácido nucleico llevada a cabo en o en asociación con un soporte sólido de manera que todos o una parte de los productos amplificados se inmovilicen sobre el soporte sólido a medida que se forman. En particular, el término abarca la reacción en cadena de la polimerasa en fase sólida (PCR en fase sólida) y la amplificación isotérmica en fase sólida que son reacciones análogas a la amplificación en fase de solución estándar, excepto que uno o ambos de los cebadores de amplificación directa e inversa está(n) inmovilizado(s) en el soporte sólido. La PCR en fase sólida cubre sistemas tales como emulsiones, en los que un cebador está anclado a una perla y el otro está en disolución libre, y la formación de colonias en matrices de gel en fase sólida en las que un cebador está anclado a la superficie y el otro está en disolución libre.

En algunas realizaciones, el soporte sólido comprende una superficie con un patrón. Una "superficie con un patrón" se refiere a una disposición de diferentes regiones en o sobre una capa expuesta de un soporte sólido. Por ejemplo, una o más de las regiones pueden ser características en las que están presentes uno o más cebadores de amplificación. Las características pueden estar separadas por regiones intersticiales donde los cebadores de amplificación no están presentes. En algunas realizaciones, el patrón puede tener un formato x-y de características que están en filas y columnas. En algunas realizaciones, el patrón puede ser una disposición repetitiva de características y/o regiones intersticiales. En algunas realizaciones, el patrón puede ser una disposición aleatoria de características y/o regiones intersticiales. Superficies con un patrón ejemplares que se pueden usar en los métodos y composiciones descritos en la presente memoria se describen en las patentes de EE.UU. núms. 8.778.848, 8.778.849 y 9.079.148, y la publicación de EE.UU. núm. 2014/0243224.

En algunas realizaciones, el soporte sólido incluye una matriz de pocillos o depresiones en una superficie. Esto se puede fabricar como se conoce generalmente en la técnica usando una variedad de técnicas, que incluyen, pero no se limitan a, fotolitografía, técnicas de formación de patrones, técnicas de moldeo y técnicas de micrograbado. Como apreciarán los expertos en la técnica, la técnica utilizada dependerá de la composición y forma del sustrato de la matriz.

Las características en una superficie con un patrón pueden ser pocillos en una matriz de pocillos (p. ej., micropocillos o nanopocillos) sobre vidrio, silicio, plástico u otros soportes sólidos adecuados con un gel con un patrón covalentemente unido tal como poli(N-(5-azidoacetamidopentil)acrilamida-co-acrilamida) (PAZAM, véase, por ejemplo, la publicación de EE.UU. núm. 2013/184796, patente internacional WO 2016/066586, y patente internacional WO 2015/002813). El proceso crea almohadillas de gel utilizadas para la secuenciación que pueden ser estables durante ejecuciones de secuenciación con una gran cantidad de ciclos. El enlace covalente del polímero a los pocillos es útil para mantener el gel con sus características estructuradas durante toda la vida útil del sustrato estructurado durante una variedad de usos. Sin embargo, en muchas realizaciones no es necesario que el gel esté unido covalentemente a los pocillos. Por ejemplo, en algunas condiciones la acrilamida libre de silano (SFA, véase, por ejemplo, la patente de EE.UU. núm. 8.563.477) que no está unida covalentemente a ninguna parte del sustrato estructurado, puede usarse como material de gel.

En realizaciones particulares, se puede preparar un sustrato estructurado presentando un material de soporte sólido con pocillos (p. ej., micropocillos o nanopocillos), recubriendo el soporte presentado con un material de gel (p. ej., PAZAM, SFA o variantes químicamente modificadas de los mismos, tales como la versión azidolizada de SFA (azido-SFA)) y pulir el soporte recubierto de gel, por ejemplo mediante pulido químico o mecánico, reteniendo así el gel en los pocillos pero eliminando o inactivando sustancialmente todo el gel de las regiones intersticiales en la superficie del sustrato estructurado entre los pocillos. Los ácidos nucleicos cebadores se pueden unir al material de gel. Luego se puede poner en contacto una disolución de fragmentos indexados con el sustrato pulido de modo que los fragmentos indexados individuales sembrarán pocillos individuales mediante interacciones con cebadores adheridos al material

de gel; sin embargo, los ácidos nucleicos diana no ocuparán las regiones intersticiales debido a la ausencia o inactividad del material del gel. La amplificación de los fragmentos indexados se limitará a los pocillos, ya que la ausencia o inactividad del gel en las regiones intersticiales impide la migración hacia afuera de la colonia de ácido nucleico en crecimiento. El proceso puede fabricarse convenientemente, ser escalable y utilizar métodos convencionales de micro o nanofabricación.

Aunque la descripción abarca métodos de amplificación en "fase sólida" en los que solo se inmoviliza un cebador de amplificación (el otro cebador normalmente está presente en disolución libre), en una realización se prefiere que el soporte sólido esté provisto tanto con cebadores directos como inversos inmovilizados. En la práctica, habrá una "pluralidad" de cebadores directos idénticos y/o una "pluralidad" de cebadores inversos idénticos inmovilizados sobre el soporte sólido, ya que el proceso de amplificación requiere un exceso de cebadores para sostener la amplificación. Las referencias en la presente memoria a cebadores directos e inversos deben interpretarse en consecuencia como que abarcan una "pluralidad" de dichos cebadores a menos que el contexto indique lo contrario.

Como apreciará el lector experto, cualquier reacción de amplificación dada requiere al menos un tipo de cebador directo y al menos un tipo de cebador inverso específico para la plantilla que se va a amplificar. Sin embargo, en determinadas realizaciones, los cebadores directos e inversos pueden incluir porciones específicas de plantilla de secuencia idéntica y pueden tener una secuencia y estructura de nucleótidos completamente idéntica (incluida cualquier modificación no nucleotídica). En otras palabras, es posible llevar a cabo la amplificación en fase sólida usando sólo un tipo de cebador, y dichos métodos de cebador único están incluidos dentro del alcance de la descripción. Otras realizaciones pueden usar cebadores directos e inversos que contienen secuencias específicas de plantilla idénticas pero que difieren en algunas otras características estructurales. Por ejemplo, un tipo de cebador puede contener una modificación no nucleotídica que no está presente en el otro.

En todas las realizaciones de la descripción, los cebadores para la amplificación en fase sólida se inmovilizan preferiblemente mediante unión covalente de un solo punto al soporte sólido en o cerca del extremo 5' del cebador, dejando la porción específica de la plantilla del cebador libre para aparearse con su plantilla afín y el grupo hidroxilo 3' libre para la extensión del cebador. Para este fin se puede usar cualquier medio de unión covalente adecuado conocido en la técnica. La química de unión elegida dependerá de la naturaleza del soporte sólido y de cualquier derivatización o funcionalización que se le aplique. El propio cebador puede incluir un resto, que puede ser una modificación química no nucleotídica, para facilitar la unión. En una realización particular, el cebador puede incluir un nucleófilo que contiene azufre, tal como fosforotioato o tiofosfato, en el extremo 5'. En el caso de hidrogeles de poli(acrilamida) soportados en sólidos, este nucleófilo se unirá a un grupo bromoacetamida presente en el hidrogel. Un medio más particular de unir cebadores y plantillas a un soporte sólido es mediante la unión de fosforotioato 5' a un hidrogel comprendido por acrilamida polimerizada y N-(5-bromoacetamidilpentil)acrilamida (BRAPA), como se describe en la patente internacional WO 05/065814.

Ciertas realizaciones de la descripción pueden hacer uso de soportes sólidos que incluyen un sustrato o matriz inerte (p. ej., portaobjetos de vidrio, perlas de polímero, etc.) que ha sido "funcionalizado", por ejemplo mediante la aplicación de una capa o recubrimiento de un material intermedio que incluye grupos reactivos que permiten la unión covalente a biomoléculas, como los polinucleótidos. Ejemplos de dichos soportes incluyen, pero no se limitan a, hidrogeles de poli(acrilamida) soportados sobre un sustrato inerte tal como vidrio. En dichas realizaciones, las biomoléculas (p. ej., polinucleótidos) pueden estar unidas directamente de forma covalente al material intermedio (p. ej., el hidrogel), pero el propio material intermedio puede estar unido de forma no covalente al sustrato o matriz (p. ej., el sustrato de vidrio). Por consiguiente, el término "unión covalente a un soporte sólido" debe interpretarse en el sentido de que abarca este tipo de disposición.

Las muestras agrupadas se pueden amplificar en perlas en donde cada perla contiene un cebador de amplificación directa e inversa. En una realización particular, la biblioteca de fragmentos indexados se usa para preparar matrices agrupadas de colonias de ácido nucleico, análogas a las descritas en la publicación de EE.UU. núm. 2005/0100900, patente de EE.UU. núm. 7.115.400, patente internacional WO 00/18957 y patente internacional WO 98/44151 mediante amplificación en fase sólida y más particularmente amplificación isotérmica en fase sólida. Los términos "grupo" y "colonia" se usan indistintamente en la presente memoria para referirse a un sitio discreto sobre un soporte sólido que incluye una pluralidad de hebras de ácido nucleico inmovilizadas idénticas y una pluralidad de hebras de ácido nucleico complementarias inmovilizadas idénticas. El término "matriz agrupada" se refiere a una matriz formada a partir de dichos grupos o colonias. En este contexto, no debe entenderse que el término "matriz" requiere una disposición ordenada de grupos.

El término "fase sólida" o "superficie" se usa para referirse a una matriz plana en donde los cebadores están unidos a una superficie plana, por ejemplo, portaobjetos de vidrio, sílice o plástico para microscopio o dispositivos de celda de flujo similares; perlas, en donde cualquiera o los dos cebadores están unidos a las perlas y las perlas se amplifican; o una matriz de perlas sobre una superficie después de que las perlas hayan sido amplificadas.

Las matrices agrupadas se pueden preparar usando un proceso de termociclado, como se describe en la patente internacional WO 98/44151, o un proceso mediante el cual la temperatura se mantiene constante y los ciclos de extensión y desnaturalización se realizan mediante cambios de reactivos. Dichos métodos de amplificación isotérmica

se describen en la solicitud de patente número. WO 02/46456 y publicación de EE.UU. núm. 2008/0009420. Debido a las temperaturas más bajas útiles en el proceso isotérmico, esto es particularmente preferido en algunas realizaciones.

Se apreciará que cualquiera de las metodologías de amplificación descritas en la presente memoria o generalmente conocidas en la técnica se puede usar con cebadores universales o específicos de diana para amplificar fragmentos de ADN inmovilizados. Los métodos adecuados para la amplificación incluyen, pero no se limitan a, la reacción en cadena de la polimerasa (PCR), la amplificación por desplazamiento de hebra (SDA), la amplificación mediada por la transcripción (TMA) y la amplificación basada en secuencias del ácido nucleico (NASBA), como se describe en la patente de EE.UU. núm. 8.003.354. Los métodos de amplificación anteriores se pueden emplear para amplificar uno o más ácidos nucleicos de interés. Por ejemplo, se puede utilizar PCR, incluyendo PCR múltiple, SDA, TMA, NASBA y similares para amplificar fragmentos de ADN inmovilizados. En algunas realizaciones, en la reacción de amplificación se incluyen cebadores dirigidos específicamente al polinucleótido de interés.

Otros métodos adecuados para la amplificación de polinucleótidos pueden incluir las tecnologías de extensión y ligadura de oligonucleótidos, amplificación por círculo rodante (RCA) (Lizardi et al., Nat. Genet. 19:225-232 (1998)) y ensayo de ligadura de oligonucleótidos (OLA) (Ver en general las patentes de EE.UU. núms. 7.582.420, 5.185.243, 5.679.524 y 5.573.907; las patentes europeas EP 0 320 308 B1; EP 0 336 731 B1; EP 0 439 182 B1; las patentes internacionales WO 90/01069; WO 89/12696; y WO 89/09835). Se apreciará que estas metodologías de amplificación pueden diseñarse para amplificar fragmentos de ADN inmovilizados. Por ejemplo, en algunas realizaciones, el método de amplificación puede incluir reacciones de amplificación con sonda de ligadura o ensayo de ligadura de oligonucleótidos (OLA) que contienen cebadores dirigidos específicamente al ácido nucleico de interés. En algunas realizaciones, el método de amplificación puede incluir una reacción de extensión-ligadura del cebador que contiene cebadores dirigidos específicamente al ácido nucleico de interés. Como ejemplo no limitante de extensión del cebador y cebadores de ligadura que pueden diseñarse específicamente para amplificar un ácido nucleico de interés, la amplificación puede incluir cebadores usados para el ensayo GoldenGate (Illumina, Inc., San Diego, CA) como se ejemplifica por las patentes de EE.UU. núms. 7.582.420 y 7.611.869.

Las nanobolas de ADN también se pueden usar en combinación con métodos y composiciones como se describe en la presente memoria. Los métodos para crear y utilizar nanobolas de ADN para la secuenciación genómica se pueden encontrar, por ejemplo, en patentes de EE.UU. y publicaciones de patente de EE.UU. núms. 7.910.354, 2009/0264299, 2009/0011943, 2009/0005252, 2009/0155781, 2009/0118488 y como se describe, por ejemplo, en Drmanac et al., 2010, Science 327(5961): 78-81. Brevemente, después de que los adaptadores de fragmentación de ADN de la biblioteca genómica se ligan a los fragmentos, los fragmentos ligados al adaptador se redondean mediante ligadura con una ligasa circular y se lleva a cabo la amplificación en círculo rodante (como se describe en Lizardi et al., 1998, Nat. Genet. 19:225-232 y el documento US 2007/0099208 A1). La estructura concatamérica extendida de los amplicones promueve el enrollamiento, creando así nanobolas de ADN compactas. Las nanobolas de ADN se pueden capturar en sustratos, preferiblemente para crear una matriz ordenada o con un patrón de modo que se mantenga la distancia entre cada nanobola, permitiendo así la secuenciación de las nanobolas de ADN separadas. En algunas realizaciones tales como las usadas por Complete Genomics (Mountain View, California), se llevan a cabo rondas consecutivas de ligadura, amplificación y digestión del adaptador antes de la circularización para producir construcciones de cabeza a cola que tienen varios fragmentos de ADN genómico separados por secuencias de adaptador.

Los métodos de amplificación isotérmica ejemplares que pueden usarse en un método de la presente descripción incluyen, pero no se limitan a, amplificación de desplazamiento múltiple (MDA) como se ejemplifica, por ejemplo, por Dean et al., Proc. Natl. Acad. Sci. USA 99:5261-66 (2002) o amplificación de ácido nucleico por desplazamiento isotérmico de hebra ejemplificada por, por ejemplo la patente de EE.UU. núm. 6.214.587. Otros métodos no basados en PCR que pueden usarse en la presente descripción incluyen, por ejemplo, amplificación por desplazamiento de la hebra (SDA) que se describe, por ejemplo, en Walker et al., Molecular Methods for Virus Detection, Academic Press, Inc., 1995; las patentes de EE.UU. núms. 5.455.166, y 5.130.238, y Walker et al., Nucl. Acids Res. 20:1691-96 (1992) o amplificación por desplazamiento de hebra hiperramificada que se describe, por ejemplo, en Lage et al., Genome Res. 13:294-307 (2003). Se pueden usar métodos de amplificación isotérmica, por ejemplo, con la polimerasa Phi 29 de desplazamiento de hebra o el fragmento grande de ADN polimerasa Bst, 5'->3' exo- para la amplificación con cebador aleatorio de ADN genómico. El uso de estas polimerasas aprovecha su alta procesividad y actividad de desplazamiento de hebras. La alta procesividad permite a las polimerasas producir fragmentos de 10 a 20 kb de longitud. Como se describe anteriormente, se pueden producir fragmentos más pequeños en condiciones isotérmicas usando polimerasas que tienen baja procesividad y actividad de desplazamiento de hebras, tales como la polimerasa Klenow. La descripción adicional de las reacciones, condiciones y componentes de amplificación se describen en detalle en la descripción de la patente de EE.UU. núm. 7.670.810.

Otro método de amplificación de polinucleótidos que es útil en la presente descripción es la PCR etiquetada que usa una población de cebadores de dos dominios que tienen una región 5' constante seguida de una región 3' aleatoria como se describe, por ejemplo, en Grothues et al. Nucleic Acids Res. 21(5):1321-2 (1993). Las primeras rondas de amplificación se llevan a cabo para permitir una multitud de iniciaciones en ADN desnaturizado por calor basadas en la hibridación individual desde la región 3' sintetizada aleatoriamente. Debido a la naturaleza de la región 3', se contempla que los sitios de iniciación sean aleatorios en todo el genoma. Después, los cebadores no unidos pueden

eliminarse y puede tener lugar una replicación adicional usando cebadores complementarios para la región 5' constante.

En algunas realizaciones, la amplificación isotérmica se puede realizar usando amplificación de exclusión cinética (KEA), también denominada amplificación de exclusión (ExAmp). Se puede preparar una biblioteca de ácidos nucleicos de la presente descripción usando un método que incluye una etapa de hacer reaccionar un reactivo de amplificación para producir una pluralidad de sitios de amplificación que incluye cada uno de ellos una población sustancialmente clonal de amplicones de un ácido nucleico diana individual que ha sembrado el sitio. En algunas realizaciones, la reacción de amplificación continúa hasta que se genera una cantidad suficiente de amplicones para llenar la capacidad del sitio de amplificación respectivo. Llenar al máximo un sitio ya sembrado de esta manera inhibe que los ácidos nucleicos diana aterricen y se amplifiquen en el sitio, produciendo así una población clonal de amplicones en el sitio. En algunas realizaciones, se puede lograr una clonalidad aparente incluso si un sitio de amplificación no se llena al máximo de su capacidad antes de que llegue un segundo ácido nucleico diana al sitio. En algunas condiciones, la amplificación de un primer ácido nucleico diana puede avanzar hasta un punto en el que se realice un número suficiente de copias para superar o abrumar eficazmente la producción de copias de un segundo ácido nucleico diana que se transporta al sitio. Por ejemplo, en una realización que usa un proceso de amplificación en puente sobre una característica circular que tiene menos de 500 nm de diámetro, se ha determinado que después de 14 ciclos de amplificación exponencial para un primer ácido nucleico diana, la contaminación de un segundo ácido nucleico diana puede disminuir en el mismo sitio producirá una cantidad insuficiente de amplicones contaminantes para impactar negativamente el análisis de secuenciación por síntesis en una plataforma de secuenciación Illumina.

En algunas realizaciones, los sitios de amplificación en una matriz pueden ser, pero no necesariamente, completamente clones. Más bien, para algunas aplicaciones, un sitio de amplificación individual puede estar poblado predominantemente con amplicones de un primer fragmento indexado y también puede tener un nivel bajo de amplicones contaminantes de un segundo ácido nucleico diana. Una matriz puede tener uno o más sitios de amplificación que tengan un bajo nivel de amplicones contaminantes siempre que el nivel de contaminación no tenga un impacto inaceptable en un uso posterior de la matriz. Por ejemplo, cuando la matriz se va a usar en una aplicación de detección, un nivel aceptable de contaminación sería un nivel que no afecte la relación señal/ruido o la resolución de la técnica de detección de una manera inaceptable. En consecuencia, la clonalidad aparente será generalmente relevante para un uso o aplicación particular de una matriz realizada mediante los métodos descritos en la presente memoria. Los niveles ejemplares de contaminación que pueden ser aceptables en un sitio de amplificación individual para aplicaciones particulares incluyen, pero no se limitan a, como máximo un 0,1 %, 0,5 %, 1 %, 5 %, 10 % o 25 % de amplicones contaminantes. Una matriz puede incluir uno o más sitios de amplificación que tengan estos niveles ejemplares de amplicones contaminantes. Por ejemplo, hasta el 5 %, 10 %, 25 %, 50 %, 75 % o incluso el 100 % de los sitios de amplificación en una matriz pueden tener algunos amplicones contaminantes. Se entenderá que en una matriz u otra colección de sitios, al menos el 50 %, 75 %, 80 %, 85 %, 90 %, 95 % o 99 % o más de los sitios pueden ser clones o aparentemente clones.

En algunas realizaciones, la exclusión cinética puede ocurrir cuando un proceso ocurre a una velocidad suficientemente rápida para excluir efectivamente la ocurrencia de otro evento o proceso. Tomemos, por ejemplo, la elaboración de una matriz de ácido nucleico en la que los sitios de la matriz se siembran aleatoriamente con fragmentos indexados de una disolución y se generan copias de los fragmentos indexados en un proceso de amplificación para llenar cada uno de los sitios sembrados hasta su capacidad. De acuerdo con los métodos de exclusión cinética de la presente descripción, los procesos de siembra y amplificación pueden realizarse simultáneamente en condiciones donde la tasa de amplificación excede la tasa de siembra. Como tal, la velocidad relativamente rápida a la que se realizan copias en un sitio que ha sido sembrado por un primer ácido nucleico diana excluirá eficazmente a un segundo ácido nucleico de sembrar el sitio para la amplificación. Los métodos de amplificación por exclusión cinética se pueden realizar como se describe en detalle en la descripción de la publicación de solicitud de EE.UU. núm. 2013/0338042.

La exclusión cinética puede aprovechar una velocidad relativamente lenta para iniciar la amplificación (p. ej., una velocidad lenta para hacer una primera copia de un fragmento indexado) frente a una velocidad relativamente rápida para hacer copias posteriores del fragmento indexado (o de la primera copia del fragmento indexado). En el ejemplo del párrafo anterior, la exclusión cinética se produce debido a la velocidad relativamente lenta de siembra del fragmento indexado (p. ej., difusión o transporte relativamente lento) frente a la velocidad relativamente rápida a la que se produce la amplificación para llenar el sitio con copias de la semilla del fragmento indexado. En otra realización ejemplar, la exclusión cinética puede ocurrir debido a un retraso en la formación de una primera copia de un fragmento indexado que ha sembrado un sitio (p. ej., activación retrasada o lenta) frente a la velocidad relativamente rápida a la que se realizan copias posteriores para llenar el sitio. En este ejemplo, puede haberse sembrado un sitio individual con varios fragmentos indexados diferentes (p. ej., pueden estar presentes varios fragmentos indexados en cada sitio antes de la amplificación). Sin embargo, la formación de la primera copia para cualquier fragmento indexado determinado se puede activar aleatoriamente de modo que la velocidad media de formación de la primera copia sea relativamente lenta en comparación con la velocidad a la que se generan las copias posteriores. En este caso, aunque un sitio individual puede haber sido sembrado con varios fragmentos indexados diferentes, la exclusión cinética permitirá que solo se amplifique uno de esos fragmentos indexados. Más específicamente, una vez que se ha activado un primer fragmento indexado para su amplificación, el sitio se llenará rápidamente hasta su capacidad con sus copias, evitando

así que se realicen copias de un segundo fragmento indexado en el sitio.

En una realización, el método se lleva a cabo para (i) transportar simultáneamente fragmentos indexados a sitios de amplificación a una velocidad media de transporte, y (ii) amplificar los fragmentos indexados que están en los sitios de amplificación a una velocidad media de amplificación, en donde la velocidad media de amplificación excede la velocidad media de transporte (Patente de EE.UU. núm. 9.169.513). Por consiguiente, la exclusión cinética se puede lograr en tales realizaciones usando una velocidad de transporte relativamente lenta. Por ejemplo, se puede seleccionar una concentración suficientemente baja de fragmentos indexados para lograr una velocidad media de transporte deseada, dando las concentraciones más bajas como resultado velocidades medias de transporte más lentas. Alternativa o adicionalmente, se puede usar una disolución de alta viscosidad y/o la presencia de reactivos de apiñamiento molecular en la disolución para reducir las velocidades de transporte. Ejemplos de reactivos de apiñamiento molecular útiles incluyen, pero no se limitan a, polietilenglicol (PEG), ficoll, dextrano o alcohol polivinílico. Formulaciones y reactivos de apiñamiento molecular ejemplares se describen en la patente de EE.UU. núm. 7.399.590. Otro factor que se puede ajustar para lograr una velocidad de transporte deseada es el tamaño medio de los ácidos nucleicos diana.

Un reactivo de amplificación puede incluir componentes adicionales que faciliten la formación de amplicones y, en algunos casos, que aumenten la velocidad de formación de amplicones. Un ejemplo es una recombinasa. La recombinasa puede facilitar la formación de amplicones al permitir la invasión/extensión repetida. Más específicamente, la recombinasa puede facilitar la invasión de un fragmento indexado por la polimerasa y la extensión de un cebador por la polimerasa usando el fragmento indexado como plantilla para la formación de amplicones. Este proceso se puede repetir como una reacción en cadena donde los amplicones producidos en cada ronda de invasión/extensión sirven como plantillas en una ronda posterior. El proceso puede ocurrir más rápidamente que la PCR estándar, ya que no se requiere un ciclo de desnaturalización (p. ej., mediante calentamiento o desnaturalización química). Como tal, la amplificación facilitada por recombinasa se puede llevar a cabo de forma isotérmica. Generalmente es deseable incluir ATP u otros nucleótidos (o en algunos casos análogos no hidrolizables de los mismos) en un reactivo de amplificación facilitado por recombinasa para facilitar la amplificación. Una mezcla de recombinasa y proteína de unión monocatenaria (SSB) es particularmente útil ya que la SSB puede facilitar aún más la amplificación. Las formulaciones ejemplares para la amplificación facilitada por recombinasa incluyen las vendidas comercialmente como kits TwistAmp por TwistDx (Cambridge, Reino Unido). Los componentes útiles del reactivo de amplificación facilitado por recombinasa y las condiciones de reacción se describen en los documentos US 5.223.414 y US 7.399.590.

Otro ejemplo de un componente que puede incluirse en un reactivo de amplificación para facilitar la formación de amplicones y en algunos casos para aumentar la velocidad de formación de amplicones es una helicasa. La helicasa puede facilitar la formación de amplicones al permitir una reacción en cadena de formación de amplicones. El proceso puede ocurrir más rápidamente que la PCR estándar, ya que no se requiere un ciclo de desnaturalización (p. ej., mediante calentamiento o desnaturalización química). Como tal, la amplificación facilitada por helicasa se puede llevar a cabo de forma isotérmica. Una mezcla de helicasa y proteína de unión monocatenaria (SSB) es particularmente útil ya que la SSB puede facilitar aún más la amplificación. Las formulaciones ejemplares para la amplificación facilitada por helicasa incluyen las que se venden comercialmente como kits IsoAmp de Biohelix (Beverly, MA). Además, se describen ejemplos de formulaciones útiles que incluyen una proteína helicasa en los documentos US 7.399.590 y US 7.829.284.

Otro ejemplo más de un componente que puede incluirse en un reactivo de amplificación para facilitar la formación de amplicones y, en algunos casos, aumentar la velocidad de formación de amplicones es una proteína de unión al origen.

Uso en secuenciación/métodos de secuenciación

Después de la unión de los fragmentos indexados a una superficie, se determina la secuencia de los fragmentos indexados inmovilizados y amplificados. La secuenciación se puede llevar a cabo usando cualquier técnica de secuenciación adecuada, y los métodos para determinar la secuencia de fragmentos indexados inmovilizados y amplificados, incluida la resíntesis de hebras, son conocidos en la técnica y se describen, por ejemplo, en Bignell et al. (documento US 8.053.192), Gunderson et al. (patente internacional WO2016/130704), Shen et al. (documento US 8.895.249), y Pipenburg et al. (documento US 9.309.502).

Los métodos descritos en la presente memoria se pueden usar junto con una variedad de técnicas de secuenciación de ácidos nucleicos. Las técnicas particularmente aplicables son aquellas en donde los ácidos nucleicos se unen en posiciones fijas en una matriz de manera que sus posiciones relativas no cambian y en donde se obtienen imágenes de la matriz repetidamente. Son particularmente aplicables las realizaciones en las que se obtienen imágenes en diferentes canales de color, por ejemplo, coincidiendo con diferentes etiquetas usadas para distinguir un tipo de base de nucleótido de otro. En algunas realizaciones, el proceso para determinar la secuencia de nucleótidos de un fragmento indexado puede ser un proceso automatizado. Las realizaciones preferidas incluyen técnicas de secuenciación por síntesis ("SBS").

Las técnicas de SBS generalmente implican la extensión enzimática de una hebra de ácido nucleico naciente mediante

la adición iterativa de nucleótidos contra una hebra plantilla. En los métodos tradicionales de SBS, se puede proporcionar un monómero de un solo nucleótido a un nucleótido diana en presencia de una polimerasa en cada administración. Sin embargo, en los métodos descritos en la presente memoria, se puede proporcionar más de un tipo de monómero de nucleótido a un ácido nucleico diana en presencia de una polimerasa en una administración.

- 5 En una realización, un monómero de nucleótido incluye ácidos nucleicos bloqueados (LNA) o ácidos nucleicos en puente (BNA). El uso de LNA o BNA en un monómero de nucleótido aumenta la fuerza de hibridación entre un monómero de nucleótido y una secuencia cebadora de secuenciación presente en un fragmento indexado inmovilizado.

10 La SBS puede usar monómeros de nucleótido que tienen un resto terminador o aquellos que carecen de cualquier resto terminador. Los métodos que usan monómeros de nucleótido que carecen de terminadores incluyen, por ejemplo, pirosecuenciación y secuenciación utilizando nucleótidos marcados con fosfato γ, como se describe en más detalle en la presente memoria. En los métodos que usan monómeros de nucleótidos que carecen de terminadores, el número de nucleótidos añadidos en cada ciclo es generalmente variable y dependiente de la secuencia plantilla y del modo de administración de nucleótidos. Para las técnicas de SBS que utilizan monómeros de nucleótidos que

15 tienen un resto terminador, el terminador puede ser efectivamente irreversible bajo las condiciones de secuenciación usadas como es el caso de la secuenciación tradicional de Sanger que utiliza didesoxinucleótidos, o el terminador puede ser reversible como es el caso de los métodos de secuenciación desarrollados por Solexa (ahora Illumina, Inc.).

20 Las técnicas de SBS pueden usar monómeros de nucleótidos que tienen un resto de etiqueta o aquellos que carecen de un resto etiqueta. Por consiguiente, los eventos de incorporación pueden detectarse basándose en una característica de la etiqueta, tal como la fluorescencia de la etiqueta; una característica del monómero de nucleótido tal como peso molecular o carga; un subproducto de la incorporación del nucleótido, tal como la liberación de pirofosfato; o similar. En realizaciones en donde dos o más nucleótidos diferentes están presentes en un reactivo de secuenciación, los diferentes nucleótidos pueden distinguirse entre sí o, alternativamente, las dos o más etiquetas diferentes pueden ser indistinguibles según las técnicas de detección que se usen. Por ejemplo, los diferentes

25 nucleótidos presentes en un reactivo de secuenciación pueden tener diferentes etiquetas y pueden distinguirse usando ópticas apropiadas como lo ejemplifican los métodos de secuenciación desarrollados por Solexa (ahora Illumina, Inc.).

Las realizaciones preferidas incluyen técnicas de pirosecuenciación. La pirosecuenciación detecta la liberación de pirofosfato inorgánico (PPi) a medida que se incorporan nucleótidos particulares en la hebra naciente (Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlen, M. y Nyren, P. (1996) "Real-time DNA sequencing using detection of pyrophosphate release". *Analytical Biochemistry* 242 (1), 84-9; Ronaghi, M. (2001) "Pyrosequencing sheds light on DNA sequencing". *Genoma Res.* 11(1), 3-11; Ronaghi, M., Uhlen, M. y Nyren, P. (1998) "A sequencing method base don real-time pyrophosphate". *Science* 281 (5375), 363; Patentes de EE.UU. núms. 6.210.891; 6.258.568 y 6.274.320). En la pirosecuenciación, el PPi liberado se puede detectar al convertirse inmediatamente en trifosfato de adenosina (ATP) mediante la ATP sulfurasa, y el nivel de ATP generado se detecta mediante fotones producidos por luciferasa.

30 Los ácidos nucleicos que se van a secuenciar se pueden unir a características de una matriz y se pueden obtener imágenes de la matriz para capturar las señales quimioluminiscentes que se producen debido a la incorporación de nucleótidos en las características de la matriz. Se puede obtener una imagen después de tratar la matriz con un tipo de nucleótido particular (p. ej., A, T, C o G). Las imágenes obtenidas después de la adición de cada tipo de nucleótido diferirán con respecto a qué características se detectan en la matriz. Estas diferencias en la imagen reflejan el contenido de secuencia diferente de las características en la matriz. Sin embargo, las posiciones relativas de cada característica permanecerán sin cambios en las imágenes. Las imágenes se pueden almacenar, procesar y analizar usando los métodos descritos en la presente memoria. Por ejemplo, las imágenes obtenidas después del tratamiento de la matriz con cada tipo de nucleótido diferente se pueden manejar de la misma manera que se ejemplifica en la presente memoria para imágenes obtenidas de diferentes canales de detección para métodos de secuenciación

40 basados en terminadores reversibles.

En otro tipo ejemplar de SBS, la secuenciación del ciclo se logra mediante la adición gradual de nucleótidos terminadores reversibles que contienen, por ejemplo, una etiqueta colorante escindible o fotoblanqueable como se describe, por ejemplo, en la patente internacional WO 04/018497 y la patente de EE.UU. núm. 7.057.026. Solexa (ahora Illumina Inc.) comercializa este enfoque y también se describe en las patentes internacionales WO 91/06678 y

50 WO 07/123.744. La disponibilidad de terminadores marcados con fluorescencia en los que tanto la terminación se puede revertir como la etiqueta fluorescente escindida facilita la secuenciación eficaz de la terminación reversible cíclica (CRT). Las polimerasas también pueden diseñarse conjuntamente para incorporarse y extenderse eficazmente a partir de estos nucleótidos modificados.

En algunas realizaciones de secuenciación basada en terminadores reversibles, las etiquetas no inhiben sustancialmente la extensión en condiciones de reacción de SBS. Sin embargo, las etiquetas de detección pueden ser eliminables, por ejemplo, mediante escisión o degradación. Las imágenes se pueden capturar tras la incorporación de etiquetas en características de ácidos nucleicos dispuestas. En realizaciones particulares, cada ciclo implica la administración simultánea de cuatro tipos de nucleótidos diferentes a la matriz y cada tipo de nucleótido tiene una etiqueta espectralmente distinta. Luego se pueden obtener cuatro imágenes, cada una usando un canal de detección

60 que es selectivo para una de las cuatro etiquetas diferentes. Alternativamente, se pueden añadir diferentes tipos de

nucleótidos secuencialmente y se puede obtener una imagen de la matriz entre cada etapa de adición. En dichas realizaciones, cada imagen mostrará características de ácido nucleico que han incorporado nucleótidos de un tipo particular. Diferentes características estarán presentes o ausentes en las diferentes imágenes debido al diferente contenido de secuencia de cada característica. Sin embargo, la posición relativa de las características permanecerá sin cambios en las imágenes. Las imágenes obtenidas a partir de dichos métodos de SBS de terminador reversible se pueden almacenar, procesar y analizar cómo se describe en la presente memoria. Después de la etapa de captura de imágenes, se pueden eliminar las etiquetas y se pueden eliminar los restos terminadores reversibles para ciclos posteriores de adición y detección de nucleótidos. La eliminación de las etiquetas después de haberlas detectado en un ciclo particular y antes de un ciclo posterior puede proporcionar la ventaja de reducir la señal de fondo y las interferencias entre ciclos. En la presente memoria se describen ejemplos de etiquetas útiles y métodos de eliminación.

En realizaciones particulares, algunos o todos los monómeros de nucleótidos pueden incluir terminadores reversibles. En dichas realizaciones, los terminadores reversibles/fluoróforos escindibles pueden incluir fluoróforos unidos al resto ribosa mediante una unión éster 3' (Metzker, *Genoma Res.* 15:1767-1776 (2005)). Otros enfoques han separado la química del terminador de la escisión de la etiqueta de fluorescencia (Ruparel et al., *Proc Natl Acad Sci USA* 102: 5932-7 (2005)). Ruparel et al. describieron el desarrollo de terminadores reversibles que usaban un pequeño grupo alilo 3' para bloquear la extensión, pero que podían desbloquearse fácilmente mediante un tratamiento corto con un catalizador de paladio. El fluoróforo se unió a la base mediante un conector fotoescindible que podría escindirse fácilmente mediante una exposición de 30 segundos a luz UV de longitud de onda larga. Por lo tanto, se puede usar la reducción por disulfuro o la fotoescisión como conector escindible. Otro enfoque para la terminación reversible es el uso de la terminación natural que se produce después de la colocación de un tinte voluminoso en un dNTP. La presencia de un tinte voluminoso cargado en el dNTP puede actuar como un terminador eficaz a través de un impedimento estérico y/o electrostático. La presencia de un evento de incorporación evita incorporaciones adicionales a menos que se elimine el tinte. La escisión del tinte elimina el fluoróforo y revierte eficazmente la terminación. También se describen ejemplos de nucleótidos modificados en las patentes de EE.UU. núms. 7.427.673 y 7.057.026.

Sistemas y métodos de SBS ejemplares adicionales que se pueden utilizar con los métodos y sistemas descritos en la presente memoria se describen en las publicaciones de EE.UU. núms. 2007/0166705, 2006/0188901, 2006/0240439, 2006/0281109, 2012/0270305 y 2013/0260372, patente de EE.UU. núm. 7.057.026, publicación PCT núm. WO 05/065814, publicación de solicitud de patente de EE.UU. núm. 2005/0100900, y publicaciones PCT núms. WO 06/064199 y WO 07/010.251.

Algunas realizaciones pueden usar la detección de cuatro nucleótidos diferentes usando menos de cuatro etiquetas diferentes. Por ejemplo, la SBS se puede realizar usando métodos y sistemas descritos en los materiales incorporados de la publicación de EE.UU. núm. 2013/0079232. Como primer ejemplo, se puede detectar un par de tipos de nucleótidos en la misma longitud de onda, pero distinguirlos en función de una diferencia de intensidad para un miembro del par en comparación con el otro, o en función de un cambio en un miembro del par (p. ej. mediante modificación química, modificación fotoquímica o modificación física) que hace que aparezca o desaparezca una señal evidente en comparación con la señal detectada para el otro miembro del par. Como segundo ejemplo, se pueden detectar tres de cuatro tipos diferentes de nucleótidos en condiciones particulares, mientras que un cuarto tipo de nucleótido carece de una etiqueta que sea detectable en esas condiciones, o se detecta mínimamente en esas condiciones (p. ej., detección mínima debido a la fluorescencia de fondo, etc.). La incorporación de los primeros tres tipos de nucleótidos en un ácido nucleico se puede determinar en base a la presencia de sus respectivas señales y la incorporación del cuarto tipo de nucleótido en el ácido nucleico se puede determinar en base a la ausencia o detección mínima de cualquier señal. Como tercer ejemplo, un tipo de nucleótido puede incluir etiqueta(s) que se detecta(n) en dos canales diferentes, mientras que otros tipos de nucleótidos se detectan en no más de uno de los canales. Las tres configuraciones ejemplares antes mencionadas no se consideran mutuamente excluyentes y pueden usarse en diversas combinaciones. Una realización ejemplar que combina los tres ejemplos es un método de SBS basado en fluorescencia que usa un primer tipo de nucleótido que se detecta en un primer canal (p. ej., dATP que tiene una etiqueta que se detecta en el primer canal cuando se excita con una primera longitud de onda de excitación), un segundo tipo de nucleótido que se detecta en un segundo canal (p. ej., dCTP que tiene una etiqueta que se detecta en el segundo canal cuando se excita con una segunda longitud de onda de excitación), un tercer tipo de nucleótido que se detecta tanto en el primer como en el segundo canal (p. ej., dTTP que tiene al menos una etiqueta que se detecta en ambos canales cuando se excita con la primera y/o segunda longitud de onda de excitación) y un cuarto tipo de nucleótido que carece de una etiqueta que no se detecta, o se detecta mínimamente, en cualquiera de los canales (p. ej., dGTP que no tiene etiqueta).

Además, como se describe en los materiales incorporados de la publicación de EE.UU. núm. 2013/0079232, los datos de secuenciación se pueden obtener usando un solo canal. En los llamados enfoques de secuenciación de un colorante, el primer tipo de nucleótido se marca pero la etiqueta se elimina después de que se genera la primera imagen, y el segundo tipo de nucleótido se marca solo después de que se genera una primera imagen. El tercer tipo de nucleótido conserva su etiqueta tanto en la primera como en la segunda imagen, y el cuarto tipo de nucleótido permanece sin etiquetar en ambas imágenes.

Algunas realizaciones pueden usar técnicas de secuenciación mediante ligadura. Dichas técnicas utilizan ADN ligasa para incorporar oligonucleótidos e identificar la incorporación de dichos oligonucleótidos. Los oligonucleótidos

normalmente tienen diferentes etiquetas que se correlacionan con la identidad de un nucleótido particular en una secuencia con la que se hibridan los oligonucleótidos. Al igual que con otros métodos de SBS, se pueden obtener imágenes después del tratamiento de una serie de características de ácidos nucleicos con los reactivos de secuenciación marcados. Cada imagen mostrará características de ácidos nucleicos que han incorporado etiquetas de un tipo particular. Estarán presentes o ausentes diferentes características en las diferentes imágenes debido al contenido de secuencia diferente de cada característica, pero la posición relativa de las características permanecerá sin cambios en las imágenes. Las imágenes obtenidas a partir de métodos de secuenciación basados en ligadura se pueden almacenar, procesar y analizar como se describe en la presente memoria. Los sistemas y métodos de SBS ejemplares que se pueden usar con los métodos y sistemas descritos en la presente memoria se describen en las patentes de EE.UU. núms. 6.969.488, 6.172.218 y 6.306.597.

Algunas realizaciones pueden utilizar secuenciación de nanoporos (Deamer, D. W. y Akeson, M. "Nanopores and nucleic acids: prospects for ultrarapid sequencing". Trends Biotechnol. 18, 147-151 (2000); Deamer, D. y D. Branton, "Characterization of nucleic acids by nanopore analysis", Acc. Chem. Res. 35:817-825 (2002); Li, J., M. Gershow, D. Stein, E. Brandin y J. A. Golovchenko, "DNA molecules and configurations in a solid-state nanopore microscope" Nat. Mater. 2:611-615 (2003)). En dichas realizaciones, el fragmento indexado pasa a través de un nanoporo. El nanoporo puede ser una proteína de membrana biológica o de poro sintético, como la α -hemolisina. A medida que el fragmento indexado pasa a través del nanoporo, cada par de bases puede identificarse midiendo las fluctuaciones en la conductancia eléctrica del poro. (patente de EE.UU. núm. 7.001.792; Soni, G. V. y Meller, "A. progress toward ultrafast DNA sequencing using solid-state nanopores". Clin. Chem. 53, 1996-2001 (2007); Healy, K. "Nanopore-based single-molecule DNA analysis". Nanomed. 2, 459-481 (2007); Cockroft, S. L., Chu, J., Amorin, M. y Ghadiri, M. R. "A single-molecule nanopore device detects DNA polymerase activity with single-nucleotide resolution". J. Am. Chem. Soc. 130, 818-820 (2008)). Los datos obtenidos de la secuenciación de nanoporos se pueden almacenar, procesar y analizar como se describe en la presente memoria. En particular, los datos pueden tratarse como una imagen de acuerdo con el tratamiento ejemplar de imágenes ópticas y otras imágenes que se describe en la presente memoria.

Algunas realizaciones pueden usar métodos que implican el seguimiento en tiempo real de la actividad de la ADN polimerasa. Las incorporaciones de nucleótidos se pueden detectar mediante interacciones de transferencia de energía por resonancia de fluorescencia (FRET) entre una polimerasa portadora de fluoróforo y nucleótidos marcados con fosfato y como se describe, por ejemplo, en las patentes de EE.UU. núms. 7.329.492 y 7.211.414, o las incorporaciones de nucleótidos se pueden detectar con guías de ondas de modo cero como se describe, por ejemplo, en la patente de EE.UU. núm. 7.315.019, y usando análogos de nucleótidos fluorescentes y polimerasas diseñadas como se describe, por ejemplo, en la patente de EE.UU. núm. 7.405.281 y la publicación de EE.UU. núm. 2008/0108082. La iluminación se puede restringir a un volumen de escala de zeptolitro alrededor de una polimerasa fijada a la superficie, de modo que la incorporación de nucleótidos marcados de manera fluorescente se pueda observar con un fondo bajo (Levene, M. J. et al. "Zero-mode waveguides for single-molecule analysis at high concentrations". Science 299, 682-686 (2003); Lundquist, P. M. et al. "Parallel confocal detection of single molecules in real time". Opt. Lett. 33, 1026-1028 (2008); Korklach, J. et al. "Selective aluminum passivation for targeted immobilization of single DNA polymerase molecules in zero-mode waveguide nano structures". Proc. Natl. Acad. Sci. USA 105, 1176-1181 (2008)). Las imágenes obtenidas mediante dichos métodos se pueden almacenar, procesar y analizar como se describe en la presente memoria.

Algunas realizaciones de SBS incluyen la detección de un protón liberado tras la incorporación de un nucleótido en un producto de extensión. Por ejemplo, la secuenciación basada en la detección de protones liberados puede usar un detector eléctrico y técnicas asociadas que están disponibles comercialmente en Ion Torrent (Guilford, CT, una subsidiaria de Life Technologies) o métodos y sistemas de secuenciación descritos en las publicaciones de EE.UU. núms. 2009/0026082; 2009/0127589; 2010/0137143; y 2010/0282617. Los métodos descritos en la presente memoria para amplificar ácidos nucleicos diana usando exclusión cinética se pueden aplicar fácilmente a sustratos usados para detectar protones. Más específicamente, los métodos descritos en la presente memoria se pueden usar para producir poblaciones clonales de amplicones que se usan para detectar protones.

Los métodos de SBS anteriores se pueden llevar a cabo ventajosamente en formatos múltiples de modo que se manipulen simultáneamente múltiples fragmentos indexados diferentes. En realizaciones particulares, se pueden tratar diferentes fragmentos indexados en un recipiente de reacción común o sobre una superficie de un sustrato particular. Esto permite la distribución conveniente de reactivos de secuenciación, la eliminación de reactivos que no reaccionaron y la detección de eventos de incorporación de manera múltiple. En realizaciones que usan ácidos nucleicos diana unidos a la superficie, los fragmentos indexados pueden estar en un formato de matriz. En un formato de matriz, los fragmentos indexados normalmente se pueden unir a una superficie de una manera espacialmente distinguible. Los fragmentos indexados pueden unirse mediante unión covalente directa, unión a una perla u otra partícula o unión a una polimerasa u otra molécula que esté unida a la superficie. La matriz puede incluir una única copia de un fragmento indexado en cada sitio (también denominado como característica) o pueden estar presentes múltiples copias que tengan la misma secuencia en cada sitio o característica. Se pueden producir múltiples copias mediante métodos de amplificación tales como amplificación en puente o PCR en emulsión como se describe con mayor detalle en la presente memoria.

Los métodos descritos en la presente memoria pueden usar matrices que tienen características en cualquiera de una

variedad de densidades que incluyen, por ejemplo, al menos aproximadamente 10 características/cm², 100 características/cm², 500 características/cm², 1.000 características/cm², 5.000 características/cm², 10.000 características/cm², 50.000 características/cm², 100.000 características/cm², 1.000.000 características/cm², 5.000.000 características/cm², o superiores.

- 5 Una ventaja de los métodos descritos en la presente memoria es que proporcionan una detección rápida y eficaz de una pluralidad de cm², en paralelo. En consecuencia, la presente descripción proporciona sistemas integrados capaces de preparar y detectar ácidos nucleicos usando técnicas conocidas en la técnica, como las ejemplificadas en la presente memoria. Por lo tanto, un sistema integrado de la presente descripción puede incluir componentes fluidicos capaces de administrar reactivos de amplificación y/o reactivos de secuenciación a uno o más fragmentos indexados
- 10 inmovilizados, incluyendo el sistema componentes tales como bombas, válvulas, depósitos, líneas fluidicas y similares. Se puede configurar y/o usar una celda de flujo en un sistema integrado para la detección de ácidos nucleicos diana. Celdas de flujo ejemplares se describen, por ejemplo, en la publicación de EE.UU. núm. 2010/0111768 y la serie de EE.UU. núm. 13/273.666. Como se ejemplifica para las celdas de flujo, uno o más de los componentes fluidicos de un sistema integrado se pueden usar para un método de amplificación y para un método de detección. Tomando como
- 15 ejemplo una realización de secuenciación de ácido nucleico, uno o más de los componentes fluidicos de un sistema integrado se pueden usar para un método de amplificación descrito en la presente memoria y para la administración de reactivos de secuenciación en un método de secuenciación tal como los ejemplificados anteriormente. Alternativamente, un sistema integrado puede incluir sistemas fluidicos separados para llevar a cabo métodos de amplificación y para llevar a cabo métodos de detección. Ejemplos de sistemas de secuenciación integrados que son
- 20 capaces de crear ácidos nucleicos amplificados y también determinar la secuencia de los ácidos nucleicos incluyen, sin limitación, la plataforma MiSeqTM (Illumina, Inc., San Diego, CA) y los dispositivos descritos en la serie de EE.UU. núm. 13/273.666.

- También se proporcionan en la presente memoria, pero no forman parte de la invención, composiciones. Durante la práctica de los métodos descritos en la presente memoria pueden resultar diversas composiciones. Por ejemplo, puede
- 25 resultar una composición que incluye fragmentos de ácido nucleico indexados, en donde los fragmentos de ácido nucleico indexados derivan de ARN recién sintetizado. En una realización, se marca el ARN recién sintetizado. También se proporciona una placa de múltiples pocillos, en donde un pocillo de la placa de múltiples pocillos incluye fragmentos de ácido nucleico indexados.

- También se proporcionan en la presente memoria, pero no forman parte de la invención, kits. En una realización, un kit sirve para preparar una biblioteca de secuenciación donde se marca el ARN recién sintetizado. En una realización, el kit incluye una etiqueta de nucleótido descrito en la presente memoria. En otra realización, el kit incluye uno o más cebadores para el apareamiento con ARN, donde al menos un cebador es para la amplificación dirigida de uno o más ácidos nucleicos predeterminados. En una realización adicional, el kit incluye los componentes para añadir al menos tres índices a los ácidos nucleicos. Un kit también puede incluir otros componentes útiles para producir una biblioteca de secuenciación. Por ejemplo, el kit puede incluir al menos una enzima que media en la ligadura, la extensión del cebador o la amplificación para procesar moléculas de ARN para incluir un índice. El kit puede incluir ácidos nucleicos con secuencias índice. El kit también puede incluir otros componentes útiles para añadir un índice a un ácido nucleico, tal como un complejo de transposoma. El kit también puede incluir uno o más cebadores para el apareamiento con ARN. Los cebadores pueden ser para la producción de un transcriptoma completo (p. ej., un cebador que incluye una secuencia poli-T) o para la amplificación dirigida de uno o más ácidos nucleicos predeterminados.
- 30
- 35
- 40

- Los componentes de un kit normalmente se encuentran en un material de embalaje adecuado en una cantidad suficiente para al menos un ensayo o uso. Opcionalmente se pueden incluir otros componentes, como tampones y disoluciones. Normalmente también se incluyen instrucciones para el uso de los componentes empaquetados. Tal como se usa en la presente memoria, la frase "material de embalaje" se refiere a una o más estructuras físicas
- 45 utilizadas para alojar el contenido del kit. El material de embalaje se construye mediante métodos rutinarios, generalmente para proporcionar un ambiente estéril y libre de contaminantes. El material de embalaje puede tener una etiqueta que indique que los componentes se pueden usar para producir una biblioteca de secuenciación. Además, el material de embalaje contiene instrucciones que indican cómo se emplean los materiales del kit. Tal como se usa en la presente memoria, el término "paquete" se refiere a un recipiente tal como vidrio, plástico, papel, papel de aluminio y similares, capaz de contener dentro de límites fijos los componentes del kit. Las "instrucciones de uso" normalmente incluyen una expresión tangible que describe la concentración del reactivo o al menos un parámetro del método de ensayo, como las cantidades relativas de reactivo y muestra que se van a mezclar, períodos de tiempo de mantenimiento para las mezclas de reactivo/muestra, temperatura, condiciones del tampón y similares.
- 50

Ejemplos

- 55 La presente descripción se ilustra mediante los siguientes ejemplos.

Ejemplo 1

El panorama transcripcional dinámico de la organogénesis de mamíferos con resolución unicelular

Durante la organogénesis de los mamíferos, las células de las tres capas germinales se transforman en un embrión que incluye la mayoría de los principales órganos internos y externos. Los reguladores clave de los defectos del desarrollo se pueden estudiar durante este período crucial, pero las tecnologías actuales carecen del rendimiento y la resolución para obtener una visión global de los estados moleculares y las trayectorias de un número de tipos de células que se diversifica y expande rápidamente. Aquí nos propusimos investigar la dinámica transcripcional del desarrollo del ratón durante la organogénesis con resolución unicelular. Con un protocolo mejorado basado en indexación combinatoria de células individuales ("sci-RNA-seq3"), perfilamos más de 2 millones de células derivadas de 61 embriones de ratón en etapas entre 9,5 y 13,5 días de gestación (E9.5 a E13.5; 10 a 15 réplicas por punto temporal). Identificamos cientos de tipos de células en expansión, contracción y transitorias, muchas de las cuales solo se detectan debido a la profundidad de la cobertura celular obtenida aquí, y definimos los conjuntos correspondientes de genes marcadores específicos del tipo de célula, varios de los cuales validamos mediante montaje completo en hibridación in situ. Exploramos la dinámica de la proliferación y la expresión génica dentro de los tipos de células a lo largo del tiempo, incluidos análisis enfocados de la cresta ectodérmica apical, el mesénquima de las extremidades y el músculo esquelético. Con un nuevo algoritmo, identificamos las principales trayectorias de desarrollo unicelular de la organogénesis del ratón y, dentro de ellas, descubrimos ejemplos de caminos distintos hacia el mismo punto final, es decir, ramificación y convergencia. Estos datos constituyen un recurso fundamental para la biología del desarrollo de los mamíferos y están disponibles de una manera que facilitará su anotación continua por parte de la comunidad investigadora.

Introducción

La organogénesis de los mamíferos es un proceso asombroso. En un corto período de tiempo, las células de las tres capas germinales se transforman en un embrión adecuado que incluye la mayoría de sus principales órganos internos y externos. Aunque es posible cultivar y estudiar embriones humanos muy tempranos in vitro¹, el acceso a material correspondiente a etapas posteriores del desarrollo embrionario humano es limitado. En consecuencia, la mayoría de los estudios sobre la organogénesis de los mamíferos se basan en organismos modelo y, en particular, el ratón.

En comparación con los humanos, los ratones se desarrollan rápidamente, con sólo 21 días entre la fertilización y el nacimiento de las crías. La implantación del blastocisto de ratón (32-64 células) se produce en el día embrionario 4 (E4.0). A esto le sigue la gastrulación y la formación de capas germinales primarias (E6.5-E7.5; 660-15.000 células)^{2,3}. Durante este tiempo se forma la línea primitiva y se produce la asignación de los distintos linajes del embrión en una secuencia anteroposterior⁴. En las etapas tempranas de los somitas (E8.0-E8.5), el embrión pasa de la gastrulación a la organogénesis temprana asociada con la formación de la placa neural y el tubo cardíaco (60.000-90.000 células). La organogénesis clásica comienza en E9.5. En los cuatro días siguientes (E9.5-E13.5), el embrión de ratón se expande de unos pocos cientos de miles de células a más de diez millones de células y, al mismo tiempo, desarrolla órganos sensoriales, órganos gastrointestinales y respiratorios, su médula espinal, su sistema esquelético y sistema hematopoyético. Como era de esperar, este período crítico del desarrollo del ratón se ha estudiado intensamente. De hecho, la mayoría de los reguladores clave de los defectos del desarrollo pueden estudiarse durante esta ventana^{5,6}.

Un paradigma convencional para los estudios de organogénesis en ratones implica centrarse en un sistema orgánico individual en una etapa restringida de desarrollo y combinar estudios de inactivación de genes con fenotipado por morfología anatómica, hibridación in situ, inmunohistoquímica^{7,8} o, más recientemente, perfiles de transcriptoma o epigenoma⁹. Aunque estos estudios enfocados han generado conocimientos fundamentales sobre el desarrollo de los mamíferos, las tecnologías subyacentes carecen del rendimiento y la resolución para obtener una visión global de los procesos moleculares dinámicos en marcha en las poblaciones y subpoblaciones de células diversas y en rápida expansión durante la organogénesis.

El "perfil de escopeta" del contenido molecular de células individuales representa una vía prometedora para abordar estas deficiencias y seguir avanzando en nuestra comprensión del desarrollo de los mamíferos. Por ejemplo, la aplicación de métodos de ARN-seq unicelular ha revelado recientemente una enorme heterogeneidad en neuronas y miocardiocitos durante el desarrollo del ratón^{10,11}. Aunque recientemente se publicaron dos atlas transcripcionales unicelulares del ratón y representan recursos importantes para el campo^{12,13}, en su mayoría están restringidos a órganos adultos y no intentan caracterizar la aparición y la dinámica temporal de los tipos de células de mamíferos durante el desarrollo.

La indexación combinatoria de células individuales ("sci-") es un marco metodológico que emplea códigos de barras de división y combinación para marcar de forma única el contenido de ácido nucleico de un gran número de células individuales o núcleos¹⁴⁻²¹. Recientemente desarrollamos un protocolo 'sci-' para transcriptomas ('sci-RNA-seq') y lo aplicamos para generar una 'cobertura celular de escopeta' 50 veces mayor del nematodo *Caenorhabditis elegans* en la etapa L2 19. Aunque el rendimiento de los métodos 'sci-' aumenta exponencialmente con el número de rondas de indexación, este potencial aún no se ha aprovechado plenamente debido a otros factores como la tasa de pérdida celular y la eficacia limitada de reacción de algunas etapas^{19,21}. Para abordar esto, desarrollamos y optimizamos ampliamente sci-RNA-seq de 3 niveles (sci-RNA-seq3), lo que dio por resultado un flujo de trabajo que puede perfilar más de un millón de células por experimento. Como anteriormente¹⁹, se pueden codificar múltiples muestras (p. ej., réplicas, puntos temporales, etc.) durante la primera ronda de indexación y procesarlas simultáneamente.

Aquí nos propusimos investigar la dinámica transcripcional del desarrollo del ratón durante la organogénesis con resolución unicelular utilizando sci-RNA-seq3. En un experimento, perfilamos más de 2 millones de células individuales derivadas de 61 embriones de ratón entre E9.5 y E13.5 (de 10 a 15 réplicas por punto temporal). A partir de estos datos, identificamos 38 tipos de células principales, así como más de 600 tipos de células granulares más (aquí denominados "subtipos" para distinguirlos de los 38 tipos de células principales). En total, descubrimos miles de nuevos genes marcadores candidatos para tipos y subtipos celulares, y validamos ejemplos representativos mediante hibridación in situ de montaje completo. Cuantificamos la dinámica de proliferación y la expresión génica en tipos de células transitorias y en expansión durante la mitad de la gestación, incluidos análisis enfocados de la cresta ectodérmica apical, el mesénquima de las extremidades y el músculo esquelético. Con un nuevo algoritmo, definimos las principales trayectorias de desarrollo unicelular de la organogénesis del ratón y, dentro de ellas, descubrimos ejemplos de caminos distintos hacia el mismo punto final, es decir, ramificación y convergencia. Todos los datos se ponen a disposición de forma gratuita de una manera que facilite su anotación continua por parte de la comunidad investigadora.

Resultados

Perfilado de 2 millones de células de 61 embriones de ratón en 5 etapas de desarrollo con sci-RNA-seq3

Para aumentar el rendimiento de sci-RNA-seq, exploramos más de 1000 condiciones experimentales. En relación con nuestra descripción original del método¹⁹, las principales mejoras introducidas por sci-RNA-seq3 (Fig. 4A, Métodos) son: (i) Desarrollamos una nueva estrategia de extracción y fijación nuclear, en donde los núcleos se extraen directamente de tejidos frescos sin ningún tratamiento enzimático. Después de su extracción, los núcleos se fijan en paraformaldehído al 4 % y se pueden almacenar en nitrógeno líquido antes de su posterior procesamiento. (ii) En comparación con nuestra descripción anterior de indexación de 3 niveles¹⁹, cambiamos de la tagmentación Tn5 indexada a la ligadura en horquilla indexada. (iii) Varias reacciones individuales, p. ej. transcripción inversa, se han optimizado aún más para mayor eficiencia. (iv) Se omite la etapa de clasificación FACS y se añaden las etapas de sonicación y filtración para minimizar la agregación de núcleos. Las etapas de preparación de la biblioteca de sci-RNA-seq3 pueden completarse por un solo individuo en una semana y superan con creces el rendimiento "por experimento" de los protocolos alternativos sc-RNA-seq (Fig. 4B).

Recogimos embriones de ratón C57BL/6 entre E9.5-E13.5 y los congelamos en nitrógeno líquido, incluyendo de 10 a 15 embriones de al menos tres camadas independientes por etapa. Posteriormente, aislamos núcleos de 61 embriones completos individuales y realizamos sci-RNA-seq3 (Fig. 4A). Los núcleos derivados de cada embrión se depositaron en diferentes pocillos durante la primera ronda de indexación, de modo que los perfiles de secuencia de ARN de los núcleos individuales pudieran vincularse a los embriones de los que se derivaron (Fig. 5A). Como control interno, también incorporamos una mezcla de células HEK293T y NIH/3T3 en dos pocillos durante la primera ronda de indexación. Después de completar el protocolo sci-RNA-seq3, la biblioteca resultante se secuenció en una ejecución de NovaSeq, lo que arrojó 11 mil millones de lecturas (Fig. 5B).

De este experimento, recuperamos 2.072.011 transcriptomas de células individuales (identificador molecular único o recuento UMI ≥ 200), incluidas 2.058.652 células de 61 embriones de ratón y 13.359 células de las células HEK293T o NIH/3T3. De manera tranquilizadora, los transcriptomas de las células HEK293T y NIH/3T3 se aplicaron abrumadoramente al genoma de una especie u otra, con 420 (3 %) colisiones (Fig. 4C). Con una profundidad de secuenciación de 23.207 lecturas por célula, observamos una mediana de 3.676 UMI por célula HEK293T y 5.163 UMI por célula NIH/3T3, con un 3,9 % y un 2,9 % de lecturas por célula aplicadas a especies incorrectas, respectivamente (Fig. 5C-D). Comparamos nuestro protocolo original sci-RNA-seq con sci-RNA-seq3 mediante el submuestreo de un conjunto de datos recopilado previamente¹⁹ a una profundidad de secuenciación equivalente por célula HEK293T o NIH/3T3. El protocolo sci-RNA-seq3, si bien obtuvo un aumento de 40 veces en el rendimiento, exhibió una eficacia comparable en términos del número de UMI detectados por célula (Fig. 5E). Además, los transcriptomas agregados de los perfiles de células individuales HEK293T derivados de sci-RNA-seq3 y sci-RNA-seq estaban altamente correlacionados (Pearson: 0,98, Fig. 5F).

Las 2.058.652 células derivadas de embriones se asignaron a los 61 embriones individuales en base a su código de barras de primera ronda (mediana de 35.272 células por embrión; Fig. 4D). El número de células recuperadas de cada embrión se correlacionó bien con el número de pocillos de primera ronda asignados (Spearman: 0,75, Fig. 6A). Con una profundidad de secuenciación relativamente baja (~ 5.000 lecturas por célula), identificamos una media de 519 genes (671 UMI) por célula (Fig. 4E). Esto es comparable o superior a otros estudios de scRNA-seq a partir de los cuales se han distinguido y anotado diversos tipos de células^{19,21,22}, a pesar de menos de un tercio como mucho de las lecturas de secuenciación sin procesar por célula (Fig. 6B-D). Los embriones en etapas posteriores (E12.5 y E13.5) exhibieron recuentos de UMI algo menores por célula, lo que sugiere una disminución del contenido de ARNm por núcleo durante el desarrollo (Fig. 6E).

En base a nuestras estimaciones aproximadas del número de células por embrión en cada punto temporal (Métodos), y sumando las 10 a 15 réplicas por punto temporal, estimamos que nuestra "cobertura celular de escopeta" del embrión de ratón es 0,8x en E9.5 (200.000 células por embrión; 152.000 perfiladas aquí), 0,3x en E10.5 (1,1 millones de células; 378.000 perfiladas), 0,2x en E11.5 (2 millones de células; 616.000 perfiladas), 0,08x en E12.5 (6 millones de células;

475.000 perfilado) y 0,03x en E13.5 (13 millones de células; 437.000 perfilado). Por lo tanto, aunque todavía no estamos "sobremuestreando", el número de células que estamos perfilando en cada etapa es equivalente a un porcentaje sustancial del contenido celular de un embrión de ratón individual (3-80 %).

Como control de la calidad de los datos, agregamos los transcriptomas unicelulares de cada individuo, lo que dio como resultado 61 "perfiles pseudomasivos" de embriones de ratón. Al contar el número de UMI que se asignan al transcrito Xist (solo expresada en mujeres) o a los transcritos del cromosoma Y, los embriones de ratón se separan fácilmente en grupos de machos ($x = 31$) y hembras ($n = 30$) (Fig. 4F) con representación equilibrada en términos del número de réplicas de machos frente a hembras en cada etapa (Fig. 6F).

Como control de calidad adicional, sometimos los transcriptomas 'pseudo-masivos' de los 61 embriones a una incrustación de vecinos estocásticos (t-SNE), lo que dio por resultado cinco grupos estrechamente agrupados que coincidían perfectamente con sus etapas de desarrollo (Fig. 7A). También ordenamos los embriones de ratón a lo largo de una trayectoria de 'pseudotiempo' usando Monocle²³, basado en los 1.000 genes principales expresados diferencialmente en distintos puntos temporales, y el orden resultante también coincide con las expectativas (Fig. 4G). En particular, hay dos huecos prominentes en la trayectoria del pseudotiempo a nivel embrionario, una entre E9.5 y E10.5 y la otra entre E11.5 y E12.5, lo que sugiere cambios dramáticos en el transcriptoma global durante estas ventanas. Asignamos a cada embrión un pseudotiempo, lo que potencialmente refleja una evaluación más detallada de la etapa de desarrollo (Fig. 7B). Por ejemplo, los embriones E10.5 colocados antes o después en el pseudotiempo de desarrollo eran morfológicamente distintos (Fig. 7C).

También examinamos los cambios en el transcriptoma global durante el desarrollo. 12.236 genes se expresaron diferencialmente en diferentes etapas de desarrollo (datos no mostrados); trazamos algunos de los genes más dinámicos en la Fig. 4H. Como era de esperar, observamos una mayor expresión de genes de hemoglobina adulta como *Hbb-bt* y *Hbb-bs*, y disminución de la expresión de genes de hemoglobina embrionaria como *Hbb-bhl* y *Hbb-x*. Genes con funciones conocidas en la diferenciación neuronal, incluidos *Cntn4*²⁴, *Neurod2*²⁵ y *neurod6*²⁶, exhiben una mayor expresión en etapas posteriores. Sin embargo, muchos genes altamente dinámicos no se han caracterizado previamente, p. ej. *Slc35f4*, *Prtg* y *Trim30a*. De todos modos, nuestra suposición - y de hecho, la motivación para recopilar datos de células individuales - es que la dinámica de la expresión genética del "embrión completo" está impulsada principalmente por cambios dramáticos en las proporciones relativas de distintos tipos de células, en lugar de cambios dentro de un solo tipo de células.

Identificación y anotación de los principales tipos y subtipos de células presentes durante la organogénesis del ratón

Para identificar los principales tipos de células, sometimos los 2.058.652 transcriptomas de células individuales (es decir, todos los embriones de todos los puntos temporales en total) hasta la agrupación de Louvain, que identificó 40 grupos distintos, y la visualización t-SNE (Fig. 8A). Es tranquilizador que, aunque observamos diferencias claras entre células derivadas de diferentes puntos temporales (Fig. 9A), las células derivadas de embriones replicados del mismo punto temporal se distribuyen de manera similar (Fig. 10). En base a los conjuntos de genes específicos de cada uno de estos 40 grupos, seleccionamos manualmente las asignaciones de tipos de células en comparación con los genes marcadores publicados (datos no mostrados). Para 37 grupos, pudimos asignarlos con confianza exactamente a un tipo de célula definido en la bibliografía, mientras que dos grupos correspondieron al linaje eritroide definitivo. Un grupo tenía recuentos de UMI anormalmente altos pero no tenía genes fuertemente específicos del grupo, lo que sugiere que puede ser un artefacto técnico de dobletes celulares. La fusión de los grupos definitivos del linaje eritroide y el descarte de este supuesto grupo doblete dieron como resultado 38 tipos de células principales (Fig. 8A). Para muchos grupos, genes marcadores altamente específicos facilitaron la identificación del tipo de célula (Fig. 8B, Fig. 9B-C, datos no mostrados). Por ejemplo, el grupo 6 (células epiteliales) expresó específicamente los genes marcadores bien caracterizados *Epcam* y *Trp63*^{27,28}, mientras que el grupo 29 (hepatocitos) se marcó específicamente por *alp* y *Alba* expresión¹². También se podrían anotar fácilmente grupos más pequeños, incluidos algunos correspondientes a tipos de células altamente especializadas. Por ejemplo, el grupo 36 se enriqueció con transcritos altamente expresados durante el desarrollo de la retina, como *tiro* y *tpm1*, lo que sugiere fuertemente que estos son melanocitos^{29,30}. El grupo 37 se enriqueció con transcritos expresados exclusivamente en la lente en desarrollo. Para los grupos correspondientes al mesénquima embrionario y al tejido conectivo, la identificación del tipo de célula fue más desafiante, en gran parte porque en la literatura actual hay menos genes marcadores altamente específicos disponibles.

De 26.183 genes, 17.789 genes (68 %) se expresaron diferencialmente (FDR del 5 %) en los 38 tipos de células principales (Fig. 9B, datos no mostrados). Entre estos, identificamos 2.863 marcadores específicos del tipo de célula, la gran mayoría de los cuales, hasta donde sabemos, no se habían asociado previamente con el tipo de célula respectivo (media 75 marcadores por grupo; Fig. 8B, Fig. 9C). Como ejemplo de cómo estos datos son útiles para definir nuevos marcadores de expresión génica específicos del desarrollo y del tipo de célula, consideremos a Sonic Hedgehog (*shh*), que se ha demostrado que desempeña un papel fundamental durante el desarrollo de muchos sistemas de órganos, incluidos los miembros, las estructuras de la línea media del cerebro, el tálamo, la médula espinal y los pulmones³¹. Detectamos la mayor expresión de *shh* en el grupo 30 (notocorda; datos no mostrados), junto con *ntn1*, *Slit1* y *Spon1*, todos se sabe que se expresan en las células de la notocorda y la placa del piso durante el desarrollo^{32,34}. Sin embargo, los genes *Tox2*, *Stxbp6*, *Schip1*, *Frmd4b*, no se habían descrito previamente como

marcadores de la notocorda, también eran altamente específicos del grupo 30.

Como era de esperar, observamos marcados cambios en las proporciones de los tipos de células durante la organogénesis. Si bien la mayoría de los 38 tipos de células principales proliferaron exponencialmente, algunas fueron transitorias y eventualmente desaparecieron en E13.5 (Fig. 11A-B). Por ejemplo, el linaje eritroide primitivo que se originó en el saco vitelino representado por el grupo 26 se caracterizó por la expresión de *Hbb-bh1*, mientras que el linaje eritroide definitivo originado en el hígado fetal estuvo marcado por la expresión de *Hbb-bs* en el grupo 22 (datos no mostrados). En E9.5, detectamos principalmente células correspondientes al linaje eritroide primitivo (Fig. 8A). Durante los siguientes 5 días, el linaje eritroide definitivo se convirtió en el tipo de célula predominante en la circulación fetal y, en última instancia, en el linaje exclusivo de glóbulos rojos de E13.5 (Fig. 8A). Los marcadores genéticos correspondientes exhibieron una dinámica similar (Fig. 11C).

Los 38 tipos de células principales identificados aquí tienen una media de 47.073 células, y el grupo más grande contiene 144.648 células (progenitores del tejido conectivo; 7,0 % del conjunto de datos general) y el grupo más pequeño solo 1.000 células (monocitos/granulocitos; 0,05 % del conjunto de datos general). Como la heterogeneidad del tipo de célula era evidente dentro de muchos de estos 38 grupos, adoptamos una estrategia iterativa, repitiendo la agrupación de Louvain en cada tipo de célula principal para identificar subgrupos (Figs. 12-13). Después de que se eliminaron los subgrupos dominados por uno o dos embriones y se fusionaron subgrupos muy similares (Métodos), se identificó un total de 655 subgrupos (denominados aquí "subtipos" para distinguirlos de los 38 tipos de células principales; Figs. 12-13). Es de destacar que nuestra sensibilidad para detectar tipos y subtipos de células en este estudio fue una función directa de la gran cantidad de células perfiladas. Por ejemplo, al repetir la agrupación de Louvain en el 2,5 % de nuestros datos (50.000 células) se identificó solo un subconjunto de tipos y subtipos de células (Fig. 14).

Los 655 subtipos consisten en una mediana de 1.869 células y varían desde 51 células (un subtipo de células de notocorda) hasta 65.894 células (un subtipo de células progenitoras del tejido conectivo) (Fig. 15A).

Casi todos los subtipos (99 %) se componen de contribuciones de múltiples embriones, sin que predomine ningún embrión (Fig. 15B-C). En apoyo de la opinión de que estos subtipos constituyen programas transcripcionales auténticos que son distintos de los subtipos relacionados, identificamos una mediana de 55 marcadores específicos por subtipo (Fig. 15D; tenga en cuenta que los marcadores específicos de subtipo se definen en virtud de ser específicos dentro del tipo de célula principal correspondiente, en lugar de en el conjunto de datos general). Más que los 38 tipos de células principales, los subtipos individuales exhibieron una dinámica variable entre E9.5 y E13.5. La mayoría de los subtipos (64%) aumentaron en el número de células estimado, mientras que el 12 % disminuyó y el 24 % mostró patrones más complejos (Fig. 16A-B). Curiosamente, pudimos separar fácilmente embriones de varias etapas de desarrollo basándose únicamente en las proporciones de células asignadas a cada uno de los subtipos (Fig. 16C).

Caracterización de las trayectorias de expresión génica durante el desarrollo de la cresta ectodérmica apical (AER) de las extremidades

Como ejemplo de lo que se puede lograr con una exploración y anotación detallada de subtipos, nos centramos en el epitelio (grupo 6) y, en particular, en la cresta ectodérmica apical (subgrupo 6.25). En base a genes marcadores específicos de subtipo, anotamos los 29 subtipos de epitelio (grupo 6; Fig. 17A; Fig. 18A, datos no mostrados). Por ejemplo, las células epiteliales del subtipo 6.10 estaban marcadas por *oc90*, un gen expresado exclusivamente en el epitelio de la vesícula ótica³⁵, mientras que las células epiteliales en el subtipo 6.25 mostraron una mayor expresión de los genes creadores bien caracterizados *Fgf8*, *Msx2* y *Rspo2*, específico de la cresta ectodérmica apical (AER), un epitelio altamente especializado involucrado en el desarrollo de los dedos³⁶. Para todos los subtipos epiteliales, identificamos genes que previamente no se sabía que fueran marcadores. Por ejemplo, la ARE también se distinguió por la expresión de *Fndc3a*, *Adamts3*, *Slc16a10*, *Snap91*, y *Pou6f2*. Montaje completo de la hibridación in situ (WISH) de *Fgf8* (un marcador conocido) y *Fndc3a* (un marcador nuevo) confirmó que ambos genes se expresan en la punta más distal de la yema de la extremidad que representa el AER en E10.5 (Fig. 17B-E).

A continuación, examinamos la dinámica de la proliferación celular y la expresión genética durante el desarrollo de AER. Identificamos un total de 1237 células AER, lo que representa solo el 0,06 % de nuestro conjunto de datos general, pero contribuyó a casi todos los embriones (45 de 61 con más de 5 células AER perfiladas). Aunque se detectan células AER en todos los puntos temporales, observamos que están en su punto máximo en términos de proporción celular por embrión en E9.5 y disminuyen posteriormente (Fig. 17F), de acuerdo con informes anteriores³⁷ y nuestros estudios de validación in situ (Fig. 17C). Para caracterizar la dinámica de la expresión génica dentro del AER durante el desarrollo, realizamos un ordenamiento pseudotemporal de las células AER en función de los 500 genes principales expresados diferencialmente entre las etapas del desarrollo, produciendo una trayectoria simple de temprano a tardío (Fig. 17G). 710 genes codificadores de proteínas se expresaron diferencialmente a lo largo del pseudotiempo de desarrollo (FDR del 5 %) (datos no mostrados). Por ejemplo, *Fgf9*, conocido por exhibir una expresión específica de AER en la yema de la extremidad³⁸, mostró una dinámica de activación retrasada en comparación con *Fgf8* y *Fndc3a* (Figura 17H). Los genes significativamente activados pueden desempeñar funciones importantes en la diferenciación de las células AER. Por ejemplo, los genes activados incluyen *Rspo2*, conocido por

ser fundamental para el mantenimiento de AER y para el crecimiento y el patrón de desarrollo de las extremidades.³⁹ (Figura 17H).

También identificamos genes cuya expresión disminuyó significativamente dentro de las células AER entre E9.5 y E13.5 (169 genes con un FDR del 1 %; Fig. 19A). Estos incluyen Ki67 (*Mki67*) y factor de crecimiento similar a la insulina 2 (*Igf2*), los cuales tienen funciones en la promoción de la proliferación celular^{40,41} (Figura 17H). De hecho, de acuerdo con el cese de la proliferación por parte del AER durante esta ventana de desarrollo, los análisis a nivel de ruta de genes significativamente decrecientes resaltan términos relacionados con la progresión del ciclo celular y el metabolismo de la glucosa, así como factores de transcripción asociados con la pluripotencia (*Isl1*, *Pou5f1*, *Nanog*) (Figuras 19B-C).

Caracterización de las trayectorias del destino celular durante el desarrollo del mesénquima de las extremidades

A continuación, intentamos investigar las trayectorias de desarrollo que atraviesan los tipos de células durante este período crítico del desarrollo de los mamíferos, incluidas las transiciones entre tipos y subtipos de células. La mayoría de los algoritmos contemporáneos para la reconstrucción de trayectorias pseudotemporales adolecen de dos limitaciones importantes. Primero, suponen que las células residen en una única variedad continua, *es decir*, sin discontinuidades entre subconjuntos de células. Sin embargo, debido a que nuestros primeros embriones derivan de E9.5, nuestro conjunto de datos no contiene células correspondientes al menos a algunos estados ancestrales. En segundo lugar, suponen que la trayectoria subyacente es un árbol en el que los puntos de ramificación corresponden a decisiones de destino. Sin embargo, se sabe que algunos tejidos contienen células transcripcionalmente indistinguibles aportadas por linajes transcripcionalmente distintos, *es decir*, la convergencia de trayectorias separadas por uno o varios eventos de ramificación.

Para abordar estas limitaciones, desarrollamos un nuevo algoritmo, incorporado en el paquete Monocle⁴², para resolver múltiples trayectorias disjuntas y al mismo tiempo permitir tanto la ramificación como la convergencia dentro de las trayectorias. Monocle 3 comienza proyectando las células en un espacio de baja dimensión que codifica el estado transcripcional mediante proyección y aproximación de colector uniforme (UMAP)⁴³. Luego, Monocle 3 detecta comunidades de células mutuamente similares utilizando la agrupación de Louvain y fusiona comunidades adyacentes usando una prueba estadística introducida en el algoritmo de abstracción de gráficos aproximados (AGA)⁴⁴. Es importante destacar que estos procedimientos permiten el mantenimiento de comunidades de células múltiples e inconexas. La etapa final en Monocle 3 tiene como objetivo resolver los caminos que las células individuales pueden tomar durante el desarrollo, señalando la ubicación no solo de las ramas sino también de las convergencias dentro del conjunto de células que componen cada comunidad, *es decir*, trayectorias. Anteriormente describimos un procedimiento llamado "gráfico L1" para incrustar un "gráfico principal" dentro de una proyección de perfiles de secuencia de ARN unicelular, de modo que cada célula esté cerca de algún punto del gráfico⁴⁵. Aunque el gráfico L1 pudo aprender trayectorias con bucles cerrados y ramas, solo pudo ejecutarse en conjuntos de datos con unos pocos cientos de células. Para permitir que el algoritmo procese miles o incluso millones de células, implementamos dos mejoras. Primero, lo ejecutamos en varios cientos de centroides de los datos en lugar de en las células mismas. En segundo lugar, restringimos el procedimiento de programación lineal del algoritmo para respetar los límites entre las trayectorias disjuntas definidas por la prueba AGA.

Primero intentamos aplicar este nuevo algoritmo a un único tipo de célula principal, el grupo 25, cuyas 26.559 células anotamos como mesénquima de yemas de extremidades sobre la base de la expresión de *Hoxd13*, *Fgf10* y *Lmx1b* (datos no mostrados). La visualización de la trayectoria de las células de este grupo con Monocle 3 ilustra la dramática expansión de las células mesenquimales de las extremidades a lo largo del tiempo de desarrollo, con el principal crecimiento entre E10.5 y E12.5 (Fig. 20A). La expresión genética es todo menos estática durante esta expansión, con los niveles de 4.763 genes codificadores de proteínas cambiando significativamente (FDR del 1 %; datos no mostrados). Las primeras etapas del desarrollo del mesénquima de las extremidades se caracterizan por algunos genes esperados, como *Tbx15*⁴⁶, y *Gpc3*⁴⁷ y las etapas posteriores por *Msx1*⁴⁸, *Efa4*⁴⁹ y *Dach1*⁵⁰ (Fig. 21A), pero la gran mayoría de los genes expresados dinámicamente son nuevos. Los factores de transcripción significativamente regulados positivamente durante el desarrollo del mesénquima de las extremidades incluyeron aquellos con funciones en la diferenciación de condrocitos. (*por ejemplo*, *Sox9*⁵¹ y *Yap1*⁵²), diferenciación muscular (*por ejemplo*, *Tead4*⁵³), y curación de heridas y regeneración de extremidades. (*por ejemplo*, *Smarcd1*⁵⁴) (Figura 21B).

Curiosamente, las células de las extremidades anteriores y posteriores no se separaron fácilmente mediante agrupamiento no supervisado (Fig. 21C) o análisis de trayectoria (Fig. 22A), pero pudieron distinguirse por la expresión mutuamente excluyente de *Tbx5* en las extremidades anteriores (2.085 células, 7,9 % de todas las células mesenquimales de las extremidades) y *Pitx1* en las extremidades posteriores (1.885 células, 7,1 % de todas las células mesenquimales de las extremidades) con solo 22 células que expresan ambos marcadores (0,08 % de todas las células mesenquimales de las extremidades frente a ~0,6 % esperado si fueran independientes; Fig. 20B)⁵⁵. Se expresaron diferencialmente 285 genes entre células asignadas a la extremidad anterior y posterior de esta manera (Fig. 20C, datos no mostrados). Genes marcadores conocidos como *Tbx4* y los genes del grupo Hoxc (*Hoxc4-10*)⁵⁶ se sobreexpresaron en las células de las extremidades traseras como se esperaba, pero también se identificaron nuevos marcadores. Por ejemplo, observamos que *Epha3* y *Hs3st3b1* están enriquecidos 5 veces en las extremidades anteriores, y *Pcdh17* e *Igf1* están enriquecidos 3 veces en las extremidades posteriores.

Aunque el tiempo de desarrollo es un eje importante de variación en la trayectoria del mesénquima de la extremidad de Monocle 3 (Fig. 20A), existe claramente una estructura adicional. Al menos algo de esto parece corresponder a los dos ejes espaciales principales del desarrollo de las extremidades: el eje proximal-distal (la dirección principal del crecimiento) y el eje anteroposterior (correspondiente a los cinco dedos)⁵⁵. Por ejemplo, *Sox6* y *Sox9* (proximal)^{57,58}, *Hoxd13* y *Tfap2b* (distal)³⁶, *Pax9* y *Alx4* (anterior), y *Shh* y *Hand2* (posterior), se distribuyeron diferencialmente a lo largo de la trayectoria del Monocle 3 (Fig. 20D; Fig. 22B). Hibridación in situ del montaje completo de *Hoxd13* (un marcador distal conocido) y *Cpa2* (un marcador novedoso cuya distribución en la trayectoria de Monocle 3 era similar a la de los marcadores distales conocidos), confirmó que ambos genes se expresan en el mesénquima de la extremidad distal entre E10.5 y E13.5 (Fig. 20F-H). La aplicación de la prueba de Moran para detectar la autocorrelación espacial con la trayectoria mesenquimal de las extremidades reveló 1.191 genes que variaban significativamente (FDR del 1 %; I de Moran > 10). Estos genes se agruparon en ocho patrones de expresión, varios de los cuales coincidían con las distribuciones de marcadores para los ejes proximal-distal y anteroposterior (Fig. 23, datos no mostrados).

En la Fig. 20I se muestra un resumen combinado de nuestros resultados para las trayectorias del AER y del mesénquima de las extremidades. Aunque el desarrollo de las extremidades se define por trayectorias relativamente simples, nuestros análisis ilustran cómo este atlas unicelular de organogénesis en ratones puede usarse para caracterizar la dinámica espaciotemporal de la expresión génica en sistemas específicos.

Delineación y caracterización de los principales linajes celulares de la organogénesis del ratón

A continuación, buscamos identificar los principales linajes de desarrollo y trayectorias celulares en todo el conjunto de datos. Monocle 3 organizó una muestra de 100.000 células de alta calidad (UMI > 400) en ocho linajes bien separados (Fig. 24A, Fig. 25A). Casi los 38 tipos de células principales pertenecen casi exclusivamente a uno de estos ocho grupos (Fig. 24B). Las excepciones son tres de los cuatro grupos más pequeños, monocitos/granulocitos (36 células), cristalino (125 células) y megacariocitos (287 células), probablemente como consecuencia de su bajo número. Las dos estructuras más complejas son claramente la *trayectoria mesenquimal*, que incluye todos los tipos de células mesenquimales y musculares (a la izquierda de la Fig. 24A y Fig. 25A), y el *tubo neural/trayectoria de notocorda*, que incluye la notocorda, el tubo neural, el progenitor y los tipos de células neuronales y gliales en desarrollo (derecha de la Fig. 24A y la Fig. 25A). Una primera trayectoria de la cresta neural ("*cresta neural 1*") incluye melanocitos y precursores de células de Schwann, mientras que una segunda trayectoria de la cresta neural ("*cresta neural 2*"), está comprendida por neuronas sensoriales. La *trayectoria hematopoyética* incluye megacariocitos, eritrocitos y linfocitos, mientras que las tres trayectorias restantes (*hepática*, *endotelial*, *epitelial*) cada uno corresponde a un único tipo de célula principal. Aunque el número estimado de células por embrión en cada uno de estos linajes aumenta exponencialmente de E9.5 a E13.5, sus proporciones permanecen relativamente estables, con la excepción de los hepatocitos que amplían su contribución casi diez veces durante esta ventana de desarrollo (0,3 % en E9.5 → 2,8 % en E13.5) (Fig. 25B-C).

UMAP proyecta células del mismo tipo en regiones definidas, pero a diferencia de t-SNE, también coloca tipos de células relacionadas cerca unas de otras. Por ejemplo, las primeras células mesenquimales parecían irradiar desde una región definida hacia los miocitos, el mesénquima de las extremidades, los condrocitos/osteoblastos y los tejidos conectivos (Fig. 24A, izquierda). De manera similar, los tipos de células que se encuentran en puntos temporales de desarrollo posteriores, como las neuronas glutamatérgicas, están separados de los precursores tempranos del SNC (p. ej., la glía radial) por un "puente" de células progenitoras neurales (Fig. 24A, derecha). Por otro lado, las discontinuidades (p. ej. entre los ocho linajes principales), probablemente refleje la falta de representación de estados intermedios o ancestrales entre estos grupos, como consecuencia de la restricción de nuestro estudio a E9.5-E13.5.

Cuando sometimos por separado cada uno de los ocho linajes principales al análisis de trayectoria como anteriormente, de manera análoga a la subagrupación iterativa, las trayectorias mesenquimales y del tubo neural/notocorda se organizaron nuevamente como se describió anteriormente (Fig. 26-27, fila superior), mientras que los otros linajes principales (epitelial, endotelial, etc.) exhibieron múltiples sublinajes discontinuos, lo que potencialmente revela trayectorias detalladas para los subtipos (Fig. 26-27, filas restantes). Por ejemplo, si anotamos la trayectoria epitelial por sus 29 subtipos (Fig. 17A), observamos varias subtrayectorias distintas, cada una de las cuales emana de una concentración focal de células derivadas de E9.5, subconjuntos subyacentes de subtipos epiteliales (Fig. 24C, Fig. 28). Por ejemplo, las células epiteliales de la cresta ectodérmica apical (Fig. 17G) forman una subtrayectoria lineal desde las células E9.5 a E13.5 que está bien separada de otras subtrayectorias epiteliales (Fig. 24C, centro inferior).

Reconstrucción de trayectorias celulares durante la miogénesis esquelética

Es necesario un trabajo adicional considerable para dilucidar completamente las relaciones entre los tipos y subtipos de células que comprenden las trayectorias representadas en la Fig. 24, particularmente las más complejas. Como ejemplo representativo de lo que podría ser posible, intentamos examinar con mayor detalle el tejido muscular en desarrollo, que se compone de distintos linajes mesodérmicos que se forman antes del inicio de la organogénesis. Por ejemplo, los músculos extraoculares son aportados por el mesodermo precordial, mientras que otros de la cara y la mandíbula son generados por el mesodermo faríngeo. La miogénesis esquelética está impulsada por un conjunto

central de factores reguladores miogénicos (MRF) que se activan mediante distintos conjuntos de genes situados más arriba⁵⁹. Por ejemplo, *Pax3* activa *Myod1* en los músculos del tronco, mientras que en la cabeza, *Pax3* es prescindible y los MRF se activan mediante *Pitx2* y *Tbx1*⁶⁰⁻⁶². *Myod1* o *Myf5* a su vez activan la miogenina, que impulsa la expresión de numerosos genes requeridos por el músculo esquelético contráctil. Nuestra hipótesis es que la trayectoria miogénica, vista a escala de un embrión completo, presentaría múltiples puntos de entrada que alimentarían a las células en un camino común correspondiente a la activación del programa de expresión génica central compartido por los miotubos.

Para probar esta hipótesis, aislamos los miocitos y sus supuestas células "ancestrales" de la trayectoria del mesénquima cuantificando primero la fracción de células en cada nodo gráfico principal que se clasificaron como miocitos (grupo 13). Recopilamos todos los nodos de la "mayoría de miocitos" y luego utilizamos los bordes del gráfico principal para expandir este conjunto de nodos a un "vecindario" más amplio de células (Fig. 29A). A continuación, volvimos a ejecutar Monocle 3 en este subconjunto de células para construir una trayectoria específica de miogénesis. Esta trayectoria presentó múltiples concentraciones focales de células de E9.5, con células de etapas posteriores distribuidas en varios caminos que irradian hacia afuera (Fig. 29B). *Pax3* y *Pax7*, que marcan los progenitores del músculo esquelético, se expresaron en células distribuidas en una amplia franja del gráfico principal (Fig. 29C). Dos segmentos lineales paralelos emanan de esta región del gráfico, en los cuales las células expresan *Myf5* o *Myod*. Ambos caminos alimentan una región común ocupada por células que expresan *Myog* o *Myh3*, marcadores de miocitos y miotubos, respectivamente. Un camino adicional cruzado por las células de E9.5, que expresaron *Lhx2*, *Tbx1*, y *Pitx2* pero niveles muy bajos de *Pax3*, alimenta la trayectoria justo más arriba de los segmentos *Myf5* y *Myod1*, posiblemente correspondientes al mesodermo faríngeo. Por lo tanto, la dinámica de los MRF y sus activadores situados más arriba a lo largo de la trayectoria fue coherente con la visión de que diferentes linajes mesodérmicos usan distintos factores para converger en el programa central de los genes musculares (Fig. 29D).

Discusión

En este estudio, buscamos caracterizar el desarrollo de los mamíferos perfilando los transcriptomas de células individuales a escala de todo el embrión de ratón, centrándonos en la ventana que corresponde a la organogénesis clásica. Al perfilar más de 2.000.000 de células de 61 embriones individuales en un solo experimento con sci-RNA-seq3, también proporcionamos el marco técnico para que los laboratorios pequeños generen conjuntos de datos de RNA-seq unicelulares con un rendimiento sin precedentes. Para resolver ramificaciones, convergencias y discontinuidades en las trayectorias de desarrollo, presentamos Monocle 3, un algoritmo novedoso para la inferencia de trayectorias que escala a millones de células.

En embriones de ratón en la mitad de la gestación, identificamos 38 tipos de células principales y más de 600 subtipos. Cada uno de estos tipos y subtipos se caracteriza por la expresión de conjuntos de genes marcadores, la gran mayoría de los cuales son nuevos, y ejemplos representativos de los cuales validamos por hibridación in situ de montaje completo. Como ilustración de la utilidad de la cobertura celular de escopeta profunda para caracterizar tipos de células raras, destacamos marcadores y genes expresados dinámicamente en la cresta ectodérmica apical (AER), un epitelio especializado con un papel crítico en el desarrollo de los dedos, pero que solo representa el 0,06% de las células perfiladas aquí. Los 38 tipos de células principales se resuelven en términos generales en 8 trayectorias, incluidas las trayectorias mesenquimal, del tubo neural/notocorda, hematopoyética, hepática, endotelial, epitelial y dos de la cresta neural. La discontinuidad entre estas ocho trayectorias es probablemente una consecuencia de la falta de representación de estados ancestrales o intermedios en nuestro conjunto de datos, que comienza en E9.5. El análisis de la trayectoria del mesénquima de las extremidades reveló correlatos de heterogeneidad del desarrollo correspondientes a ejes espaciales temporales y múltiples. Centrándonos en el subconjunto de la trayectoria mesenquimal correspondiente a los miocitos y sus progenitores, identificamos múltiples subtrayectorias que alimentan un punto final común correspondiente a los miotubos. Este ejemplo de "convergencia" de programas de expresión contrasta con la estructura de ramificación asumida por la mayoría de los algoritmos para la inferencia de trayectorias de desarrollo.

Nuestro estudio tiene varias limitaciones que deben tenerse en cuenta. En primer lugar, al igual que con otros atlas unicelulares, los datos del transcriptoma de células individuales son escasos. Sin embargo, investigaciones anteriores han demostrado que los programas transcripcionales se pueden distinguir fácilmente dentro de conjuntos de datos de transcriptomas unicelulares a profundidades de secuenciación sorprendentemente bajas⁶³. El hecho de que podamos definir 655 subtipos transcripcionalmente distintos con una mediana de 671 UMI por célula es coherente con esta visión, y la agregación de transcriptomas con cada tipo o subtipo de célula nos permite construir perfiles de expresión representativos. En segundo lugar, aunque tenemos una confianza razonable en la mayoría de las asignaciones de tipos de células realizadas aquí, deben considerarse, no obstante, preliminares. Un desafío clave es que el desarrollo del ratón en la mitad de la gestación (E9.5-E13.5) no se ha estudiado previamente con resolución unicelular ni a escala de organismo completo. Los atlas transcripcionales unicelulares existentes han perfilado órganos individuales de ratones adultos o etapas embrionarias tardías^{12,13}. Aunque hemos logrado avances significativos hasta la fecha, la anotación integral de estos 655 subtipos de células es un proyecto en curso y uno que anticipamos que se beneficiará de los aportes de la comunidad y la experiencia en el dominio para llegar a un consenso estable. Con ese fin, creamos una wiki para facilitar sus anotaciones por parte nuestra y de la comunidad (disponible en internet en atlas.gs.washington.edu/mouse-rna/). Una página única para cada subtipo incluye una matriz descargable de las

células que lo componen, una lista de los genes marcadores específicos de ese subtipo y una descripción de la dinámica de ese subtipo durante la ventana de desarrollo examinada aquí.

Un objetivo de larga duración en este campo, quizás por fin a la vista desde una perspectiva técnica, es producir un atlas molecular integral, resuelto espaciotemporalmente, del desarrollo de los mamíferos con resolución unicelular. Con este fin, centrarse en el ratón tiene varias ventajas, incluido su pequeño tamaño, la accesibilidad de los primeros momentos del desarrollo, un trasfondo genético endogámico y la manipulabilidad genética. Al perfilar un número de células correspondientes a un porcentaje sustancial del contenido celular de un embrión de ratón individual (3 a 80 % de "cobertura celular de escopeta" por etapa), estos datos constituyen un poderoso recurso para el campo de la biología del desarrollo y también pueden ayudar a avanzar aún más en el desarrollo de métodos computacionales para resolver e interpretar tipos de células o trayectorias de desarrollo. De cara al futuro, anticipamos que la medición integrada del transcriptoma, fenotipos moleculares adicionales⁶⁴, historia del linaje⁶⁵ y la información espacial dará forma a una visión global del desarrollo de los mamíferos.

Concluimos señalando que los atlas unicelulares del desarrollo de ratones de tipo salvaje también representan un primer etapa hacia la comprensión de los trastornos del desarrollo pleiotrópicos a escala del organismo, así como para investigaciones detalladas de las funciones sutiles de los genes y las secuencias reguladoras en el desarrollo. Por ejemplo, mientras que ~35 % de los genes inactivados en ratones son letales⁵, muchos desactivados, y en particular aquellos de secuencias reguladoras conservadas, no muestran ninguna anomalía con el fenotipado convencional⁶⁶. Anticipamos que sc-RNA-seq a escala de organismo potenciará la genética inversa, p. ej. potencialmente permitiendo el descubrimiento de fenotipos previamente pasados por alto con defectos sutiles en los programas moleculares o las proporciones relativas de tipos de células específicas⁶⁷.

Métodos

Disección embrionaria

Los ratones C57BL/6 se obtuvieron de The Jackson Laboratory (Bar Harbor, ME) y se establecieron apareamientos con tapón. El día de taponamiento se consideró día embrionario (E) 0.5. Las disecciones se realizaron como se describió anteriormente⁶⁹ y todos los embriones se congelaron inmediatamente en nitrógeno líquido. Todos los procedimientos con animales se realizaron de acuerdo con las regulaciones institucionales, estatales y gubernamentales (protocolo IACUC 4378-01).

Hibridación in situ de montaje completo

La expresión de ARNm en embriones de ratón E9.5-E11.5 se evaluó mediante hibridación in situ de montaje completo (WISH) usando una ribsonda antisentido marcada con digoxigenina transcrita a partir de sondas específicas de genes clonados (PCR DIG Probe Synthesis Kit, Roche). Los embriones completos se fijaron durante la noche en PFA/PBS al 4 %. Los embriones se lavaron en PBST (Tween al 0,1 %) y se deshidrataron paso a paso en metanol al 25 %, 50 % y 75 %/PBST y finalmente se almacenaron a -20 °C en metanol al 100 %. El protocolo WISH fue el siguiente: Día 1) Los embriones se rehidrataron en hielo en etapas inversas de metanol/PBST, se lavaron en PBST y se blanquearon en H₂O₂ al 6 %/PBST durante 1 hora y se lavaron en PBST. Luego, los embriones se trataron en 10 µg/ml de proteinasa K/PBST durante 3 minutos, se incubaron en glicina/PBST, se lavaron en PBST y finalmente se volvieron a fijar durante 20 minutos con 4 % de PFA/PBS, 0,2 % de glutaraldehído y 0,1 % de Tween 20. Después de más etapas de lavado con PBST, los embriones se incubaron a 68 °C en tampón L1 (formamida desionizada al 50 %, SSC 5x, SDS al 1 %, Tween 20 al 0,1 % en DEPC; pH 4,5) durante 10 minutos. A continuación, los embriones se incubaron durante 2 horas a 68 °C en tampón de hibridación 1 (L1 con 0,1 % de ARNt y 0,05 % de heparina). Posteriormente, los embriones se incubaron durante la noche a 68 °C en tampón de hibridación 2 (tampón de hibridación 1 con 0,1 % de ARNt y 0,05 % de heparina y sonda DIG 1:500). Día 2) La eliminación de la sonda no unida se realizó mediante una serie de pasos de lavado de 3x30 minutos cada uno a 68 °C: L1, L2 (formamida desionizada al 50 %, 2x SSC pH 4,5, Tween 20 al 0,1 % en DEPC; pH 4,5) y L3 (2x SSC pH 4,5, Tween 20 al 0,1 % en DEPC; pH 4,5). Posteriormente, los embriones se trataron durante 1 hora con disolución de RNasa (NaCl 0,1 M, Tris 0,01 M, pH 7,5, Tween 20 al 0,2 %, RNasa A 100 µg/ml en H₂O), seguido de lavado en TBST 1 (NaCl 140 mM, KCl 2,7 mM, Tris-HCl 25 mM, Tween 20 al 1 %; pH 7,5). A continuación, los embriones se bloquearon durante 2 horas a temperatura ambiente en una disolución de bloqueo (TBST 1 con 2 % de suero de ternera y 0,2 % de BSA), seguido de una incubación a 4 °C durante la noche en disolución de bloqueo que contiene Anti-Digoxigenina-AP 1:5000. Día 3) La eliminación del anticuerpo no unido se realizó mediante una serie de etapas de lavado 8 x 30 min a temperatura ambiente con TBST 2 (TBST con 0,1 % de Tween 20 y 0,05 % de levamisol/tetramisol) y se dejó durante la noche a 4 °C. Día 4) La tinción de los embriones se inició lavando a temperatura ambiente con tampón fosfato alcalino (NaCl 0,02 M, MgCl₂ 0,05 M, Tween 20 al 0,1 %, Tris-HCl 0,1 M y levamisol/tetramisol al 0,05 % en H₂O) 3x 20 minutos, seguido de tinción con BM Purple AP Substrate (Roche). Se tomaron imágenes de los embriones teñidos utilizando un microscopio Zeiss Discovery V.12 y una cámara digital Leica DFC420.

Cultivo de células de mamíferos

Todas las células de mamífero se cultivaron a 37 °C con 5 % de CO₂ y se mantuvieron en DMEM con alto contenido de glucosa (Gibco núm. de catálogo 11965) para células HEK293T y NIH/3T3, ambas suplementadas con FBS al 10 % y 1X Pen/Strep (Gibco núm. de catálogo 15140122; 100 U/ml de penicilina, 100 µg/ml de estreptomycin). Las células se tripsinizaron con tripsina-EDTA al 0,25 % (Gibco núm. de catálogo 25200-056) y se dividieron 1:10 tres veces por semana.

Extracción y fijación de núcleos de embriones de ratón.

Se procesaron juntos embriones de ratón de diferentes etapas de desarrollo para reducir el efecto por lotes. Cada embrión de ratón se picó en trozos pequeños con una cuchilla en 1 ml de tampón de lisis celular enfriado con hielo (Tris-HCl 10 mM, pH 7,4, NaCl 10 mM, MgCl₂ 3 mM y IGEPAL CA-630 al 0,1% de⁷⁰, modificado para incluir también 1 % de SUPERasa In y 1 % de BSA) y transferido a la parte superior de un colador de células de 40 µm (Falcon). Los tejidos se homogeneizaron con la punta de goma de un émbolo de jeringa (5 ml, BD) en 4 ml de tampón de lisis celular. Luego, los núcleos filtrados se transfirieron a un tubo nuevo de 15 ml (Falcon) y se granularon mediante centrifuga a 500xg durante 5 minutos y se lavaron una vez con 1 ml de tampón de lisis celular. Los núcleos se fijaron en 4 ml de paraformaldehído al 4% (EMS) enfriado con hielo durante 15 minutos en hielo. Después de la fijación, los núcleos se lavaron dos veces en 1 ml de tampón de lavado de núcleos (tampón de lisis celular sin IGEPAL) y se resuspendieron en 500 µl de tampón de lavado de núcleos. Las muestras se dividieron en dos tubos con 250 µl en cada tubo y se congelaron instantáneamente en nitrógeno líquido.

Como control de calidad, se tripsinizaron células HEK293T y NIH/3T3, se centrifugaron a 300 xg durante 5 min (4 °C) y se lavaron una vez en 1XPBS. Se combinaron y lisaron un número igual de células HEK293T y NIH/3T3 usando 1 ml de tampón de lisis celular enfriado con hielo, seguido de las mismas condiciones de fijación y almacenamiento que en el embrión de ratón.

Preparación y secuenciación de la biblioteca sci-RNA-seq3

Los núcleos descongelados se permeabilizan con tritonX-100 al 0,2 % (en tampón de lavado de núcleos) durante 3 minutos en hielo y se sonicen brevemente (Diagenode, 12 s en modo de baja potencia) para reducir la aglutinación de los núcleos. Luego, los núcleos se lavaron una vez con tampón de lavado de núcleos y se filtraron a través de un colador celular Flowmi de 1 ml (Flowmi). Los núcleos filtrados se centrifugaron a 500 xg durante 5 minutos y se resuspendieron en tampón de lavado de núcleos.

Luego se distribuyeron los núcleos de cada embrión de ratón en varios pocillos individuales en cuatro placas de 96 pocillos. Los vínculos entre la identificación del pocillo y el embrión de ratón se registraron para el procesamiento de datos posterior. Para cada pocillo, se mezclaron 80.000 núcleos (16 µl) con 8 µl de cebador oligo-dT anclado 25 µM (5'-5Phos/CAGAGCNNNNNNNN[código de barras de 10 pb]TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT-3' (SEQ ID NO:1), donde "N" es cualquier base; IDT) y 2 µl de mezcla de dNTP 10 mM (Thermo), se desnaturizó a 55 °C durante 5 min y se colocó inmediatamente en hielo. 14 µl de mezcla de reacción de primera hebra, que contiene 8 µl de tampón de primera hebra Superscript IV 5X (Invitrogen), 2 µl de DTT 100 mM (Invitrogen), 2 µl de transcriptasa inversa SuperScript IV (200 U/µl, Invitrogen), 2 µl de inhibidor de ribonucleasa recombinante RNaseOUT (Invitrogen), se añadió después a cada pocillo. La transcripción inversa se llevó a cabo incubando placas mediante gradiente de temperatura (4 °C 2 minutos, 10 °C 2 minutos, 20 °C 2 minutos, 30 °C 2 minutos, 40 °C 2 minutos, 50 °C 2 minutos y 55 °C 10 minutos).

Después de la reacción a temperatura ambiente, se añadieron a cada pocillo 60 µl de tampón de dilución de núcleos (Tris-HCl 10 mM, pH 7,4, NaCl 10 mM, MgCl₂ 3 mM y BSA al 1 %). Los núcleos de todos los pocillos se agruparon y se centrifugaron a 500 xg durante 10 minutos. Luego, los núcleos se resuspendieron en tampón de lavado de núcleos y se redistribuyeron en otras cuatro placas de 96 pocillos, cada pocillo incluía 4 µl de tampón de ligadura T4 (NEB), 2 µl de ADN T4 ligasa (NEB), 4 µl de disolución de betaína (5 M, Sigma-Aldrich), 6 µl de núcleos en tampón de lavado de núcleos, 8 µl de adaptador de ligadura con código de barras (100 µM, 5'-GCTCTG[código de barras A de 9 pb o 10 pb]/ideoxiU/ACGACGCTCTCCGATCT[complemento inverso del código de barras A]-3') (SEQ ID NO:2) y 16 µl de PEG 8000 al 40 % (Sigma-Aldrich). La reacción de ligadura se realizó a 16 °C durante 3 horas.

Después de la reacción a temperatura ambiente, se añadieron a cada pocillo 60 µl de tampón de dilución de núcleos (Tris-HCl 10 mM, pH 7,4, NaCl 10 mM, MgCl₂ 3 mM y BSA al 1 %). Los núcleos de todos los pocillos se agruparon y se centrifugaron a 600xg durante 10 minutos. Los núcleos se lavaron una vez con tampón de lavado de núcleos y se filtraron con un colador de células Flowmi de 1 ml (Flowmi) dos veces, se contaron y se redistribuyeron en ocho placas de 96 pocillos, cada pocillo incluía 2500 núcleos en 5 µl de tampón de lavado de núcleos y 5 µl de tampón de elución (Qiagen). Luego se añadieron a cada pocillo 1,33 µl de tampón de síntesis de la segunda hebra (NEB) y 0,66 µl de enzima de síntesis de la segunda hebra (NEB) de ARNm, y la síntesis de la segunda hebra se llevó a cabo a 16 °C durante 180 minutos.

Para la tagmentación, cada pocillo se mezcló con 11 µl de tampón Nextera TD (Illumina) y 1 µl de enzima TDE1 solo i7 (62,5 nM, Illumina) y luego se incubó a 55 °C durante 5 minutos para llevar a cabo la tagmentación. Luego se detuvo

la reacción añadiendo 24 µl de tampón de unión a ADN (Zymo) por pocillo e incubando a temperatura ambiente durante 5 minutos. Luego, cada pocillo se purificó utilizando perlas AMPure XP 1,5x (Beckman Coulter). En la etapa de elución, a cada pocillo se le añadieron 8 µl de agua libre de nucleasa, 1 µl de tampón USER 10X (NEB), 1 µl de enzima USER (NEB) y se incubó a 37 °C durante 15 min. Se añadieron otros 6,5 µl de tampón de elución a cada pocillo. Las perlas AMPure XP se retiraron mediante un soporte magnético y el producto de elución se transfirió a una nueva placa de 96 pocillos.

Para la amplificación por PCR, cada pocillo (16 µl de producto) se mezcló con 2 µl de cebador P5 indexado 10 µM (5'-AATGATACGGCGACCGAGATCTACAC[i5]ACACTCTTTCCCTACACGACGC TCTTCCGATCT-3'; IDT) (SEQ ID NO:3), 2 µl de cebador P7 de 10 µM (5'-CAAGCAGAAGACGGCATACGAGAT[i7]GTCTCGTGGGCTCGG-3', IDT) (SEQ ID NO:4) y 20 µl de mezcla maestra de PCR 2X de alta fidelidad NEBNext (NEB). La amplificación se llevó a cabo usando el siguiente programa: 72 °C durante 5 min, 98 °C durante 30 s, 12-14 ciclos de (98 °C durante 10 s, 66 °C durante 30 s, 72 °C durante 1 min) y un final de 72 °C durante 5 min.

Después de la PCR, las muestras se combinaron y purificaron usando 0,8 volúmenes de perlas AMPure XP. Las concentraciones de las bibliotecas se determinaron mediante Qubit (Invitrogen) y las bibliotecas se visualizaron mediante electroforesis en un gel TBE-PAGE al 6 %. Todas las bibliotecas se secuenciaron en una plataforma NovaSeq (Illumina) (Lectura 1: 34 ciclos, Lectura 2: 52 ciclos, Índice 1: 10 ciclos, Índice 2: 10 ciclos).

Procesamiento de lecturas de secuenciación

Las llamadas base se convirtieron al formato fastq usando bcl2fastq de Illumina y se desmultiplexaron en base a los códigos de barras de PCR i5 e i7 usando el paquete de desmultiplexación de máxima probabilidad deML⁷¹ con la configuración predeterminada. El procesamiento de secuencia posterior y la generación de matriz de expresión digital de células individuales fueron similares con sci-RNA-seq¹⁹ excepto que el índice RT se combinó con el índice del adaptador de horquilla y, por lo tanto, las lecturas mapeadas se dividieron en índices celulares constituyentes mediante lecturas de desmultiplexación usando tanto el índice RT como el índice de ligadura (ED <2, incluidas inserciones y eliminaciones). Brevemente, las lecturas desmultiplexadas se filtraron en base al índice RT y el índice de ligadura (ED <2, incluidas inserciones y eliminaciones) y el adaptador se recortó usando trim_galore/0.4.1 con la configuración predeterminada. Las lecturas recortadas se asignaron al genoma de referencia del ratón (mm10) para los núcleos de embriones de ratón, o a un genoma de referencia quimérico de hg19 humano y mm10 de ratón para los núcleos mixtos de HEK293T y NIH/3T3, usando STAR/v 2.5.2b⁷² con configuraciones predeterminadas y anotaciones génicas (GENCODE V19 para seres humanos; GENCODE VM11 para ratón). Se extrajeron lecturas de mapeo exclusivo y se eliminaron duplicados usando la secuencia de identificador molecular único (UMI), el índice de transcripción inversa (RT), el índice del adaptador de ligadura en horquilla y la coordenada final de lectura 2 (es decir las lecturas con UMI, índice RT, índice del adaptador de ligadura y sitio de tagmentación idénticos se consideraron duplicados). Finalmente, las lecturas mapeadas se dividieron en índices celulares constituyentes mediante lecturas de desmultiplexación adicionales usando el índice RT y la horquilla de ligadura (ED <2, incluidas inserciones y eliminaciones). Para el experimento de especies mixtas, se calculó el porcentaje de lecturas de mapeo exclusivo para los genomas de cada especie. Las células con más del 85 % de los UMI asignados a una especie se consideraron como células específicas de especie, y las células restantes se clasificaron como células mixtas o "colisiones". Para generar matrices de expresión digital, calculamos el número de UMI específicos de hebra para cada célula que se asigna a las regiones exónicas e intrónicas de cada gen con el paquete Python HTseq⁷³. Para las lecturas de mapeo múltiple, las lecturas se asignaron al gen más cercano, excepto en los casos donde otro gen interseccionado se encontraba dentro de los 100 pb del final del gen más cercano, en cuyo caso la lectura se descartó. Para la mayoría de los análisis, incluimos UMI intrónicos y exónicos de hebra esperada en matrices de expresión unicelular por gen.

Análisis completo de embriones de ratón

Después de generar la matriz de recuento de genes de una sola célula, cada célula se asignó a su embrión de ratón original en base al código de barras RT. Las lecturas mapeadas de cada embrión se agregaron para generar "RNA-seq masivo" para cada embrión. Para la separación por sexos de los embriones, contamos las lecturas de mapeo a genes de ARN no codificante (Xist) específicos de hembras o cromosoma Y (excepto el gen Erdr1 que se encuentra tanto en cromosoma X como en el cromosoma Y). Los embriones se separaron fácilmente en población de hembras (más lecturas de mapeo de genes Xist que de genes del cromosoma Y) y grupo de machos (más lecturas de mapeo de genes del cromosoma Y que de Xist).

Monocle 2 realizó el orden pseudotemporal de embriones de ratón completos⁷⁴. Brevemente, se construyó una matriz de expresión génica agregada como se describe anteriormente. Se identificaron genes expresados diferencialmente en diferentes condiciones de desarrollo con la función differentialGeneTest de Monocle 2⁷⁴. Los 2.000 genes principales con el valor q más bajo se usaron para construir la trayectoria del pseudotiempo utilizando Monocle 2⁷⁴. A cada embrión se le asignó un valor de pseudotiempo basado en su posición a lo largo del árbol de trayectoria.

Agrupación celular, visualización de t-SNE e identificación de genes marcadores

Se construyó una matriz de expresión génica digital a partir de los datos de secuenciación sin procesar como se

describe anteriormente. Se descartaron las células con menos de 200 UMI. Los análisis posteriores se realizaron con Monocle2⁷⁴ y paquete de Python scanpy⁷⁵. Brevemente, el mapeo del recuento de genes en los cromosomas sexuales se eliminó antes de la agrupación y la reducción de dimensiones. La etapa de preprocesamiento es similar al enfoque usado por Zheng *et al.*²² por la función "receta zheng17" (n_genes_principales = 2.000) en scanpy⁷⁵. La dimensión de los datos se redujo primero con PCA (30 componentes) y luego con t-SNE, seguido de la agrupación de Louvain realizada en los 30 componentes principales (resolución = 1,5). Se identificaron 40 grupos. Luego tomamos muestras de 1.000 células de cada grupo y se identificaron los genes expresados diferencialmente en diferentes grupos con la función differentialGeneTest de Monocle 2⁷⁴. Los genes específicos de cada grupo se identificaron de manera similar a antes⁷⁶. Los grupos se asignaron a tipos de células conocidos en base a marcadores específicos del grupo (Tabla 1). Un grupo tenía recuentos de UMI anormalmente altos, pero no tenía genes fuertemente específicos del grupo, lo que sugiere que puede ser un artefacto técnico de dobletes celulares y, por lo tanto, se elimina. Otros dos grupos parecían corresponder al linaje eritroide definitivo y están fusionados. Los perfiles de expresión de consenso para cada tipo de célula se construyeron como en ⁷⁶. Para identificar el marcador genético específico del tipo de célula, seleccionamos genes que se expresaron diferencialmente en diferentes tipos de células (FDR del 5 %, prueba de razón de verosimilitud) y que también tienen expresión máxima en cada tipo de célula con un aumento de al menos 2 veces en comparación con otros tipos de células. con la segunda expresión máxima.

Tabla 1

Tipo de celda de grupo	Marcadores usados para identificación del tipo de célula	Bibliografía
1 Progenitores del tejido conectivo	Il1rapl2, Meox2, Tgfb2, Adamts9, Postn, Ror1	(disponible en internet en www.ncbi.nlm.nih.gov/pubmed/25249460)
2 Condrocitos	Runx2, Twist2, Prrx1	(disponible en Internet en www.ncbi.nlm.nih.gov/pmc/articles/PMC4990491/)
3 Mesodermo intermedio	Wt1, Mylk, Ednra	(disponible en Internet en www.ncbi.nlm.nih.gov/pubmed/25480331)
4 Progenitores de mandíbula y dientes	Sox9, Foxp2, Col2a1, Col9a1, Col11a1, Pax9	(disponible en internet en www.ncbi.nlm.nih.gov/pubmed/26969076)
5 Neuronas del mesencéfalo	Car10, Mapt, Ntng1, Snhg11	(disponible en Internet en www.cell.com/cell/comments/S0092-8674(18)30116-8)
6 Células epiteliales	Epcam, Trp63, Grh12,	(disponible en Internet en www.cell.com/cell/comments/S0092-8674(18)30116-8)
7 Glía radial	Pth2r, Fabp7, Pax3, Fzd10, Hes5	(disponible en Internet en www.cell.com/cell/comments/S0092-8674(18)30116-8)
8 Mesénquima temprano	Gpc5, Smoc1	(disponible en internet en www.ncbi.nlm.nih.gov/pubmed/21194678)
9 Células progenitoras neuronales	Prmt8, Gadd45g, Cdkn1c, Btg2	(disponible en Internet en www.cell.com/cell/comments/S0092-8674(18)30116-8)
10 Neuronas posmitóticas	Tmem163, Slc17a6	(disponible en Internet en www.cell.com/cell/comments/S0092-8674(18)30116-8)
11 Progenitores de oligodendrocitos	Slc17a6, Sox1, Olig2, Nkx2-1	(disponible en Internet en www.ncbi.nlm.nih.gov/pmc/articles/PMC3472585/)

Tipo de celda de grupo	Marcadores usados para identificación del tipo de célula	Bibliografía
12 Progenitores del mesencéfalo/prosencéfalo	Egfem1, EN2, Fgf15	(disponible en Internet en www.ncbi.nlm.nih.gov/pubmed/10640701?dopt=Abstract)
13 Miocitos	Neb, Myh3, Tpm2, Acta2 (músculo liso)	(disponible en Internet en www.ncbi.nlm.nih.gov/pubmed/24509862)
14 Tubo neural	Foxb1, Scube2, Prtg	(disponible en internet en www.ncbi.nlm.nih.gov/pubmed/18064677)
15 Neuronas glutamatérgicas	Grm5, Nxph1, Ndn, Pcsk1n, Cdh23, Grin2b	(disponible en Internet en www.cell.com/cell/comments/S0092-8674(18)30116-8)
16 Células estromales	Il23a, Bmpr1a, Prtg	(disponible en Internet en www.ncbi.nlm.nih.gov/pmc/articles/PMC4307917/)
17 Osteoblastos	Col1a1, Camk1d, Rbm8a	(disponible en Internet en www.ncbi.nlm.nih.gov/pubmed/15042706?dopt=Abstract)
18 Neuronas del hipocampo	Pax2, Lhx1, Cdh22, Asic4	(disponible en Internet en www.cell.com/cell/comments/S0092-8674(18)30116-8)
19 Oligodendrocitos prematuros	Id4, Gpc1	(disponible en internet en www.ncbi.nlm.nih.gov/pmc/articles/PMC305683/) (disponible en internet en www.ncbi.nlm.nih.gov/pmc/articles/PMC5650988/)
20 Células endoteliales	Ptprb, Pecam1, Vwf, Klf14, Hbegf, Egr17 (Riñón)	(disponible en Internet en www.ncbi.nlm.nih.gov/pmc/articles/PMC305683/)
21 Progenitores de condrocitos	ITGA11, ATP1A2, Lamc3, EphA7	(disponible en internet en journals.plos.org/plosone/article?id=10.1371/journal.pone.0082035)
22 Linaje definitivo de eritrocitos	Snca, Hbb-bs, Abcb4, Slc4a1, Kel, HBB-y, Hba-x,	(disponible en Internet en www.cell.com/cell/comments/S0092-8674(18)30116-8)
23 Precursores de células de Schwann	Plp1, Cdh19	(disponible en Internet en www.ncbi.nlm.nih.gov/pmc/articles/PMC4484967/)
24 Neuronas sensoriales	Syt13, Shox2, Ptprr, Pcbp3	(disponible en Internet en www.cell.com/cell/comments/S0092-8674(18)30116-8)
25 Mesénquima de las extremidades	Msx1, Fgf10, Wnt5a, Lmx1b	(disponible en Internet en www.ncbi.nlm.nih.gov/pubmed/9106164?dopt=Abstract)
26 Linaje eritroide primitivo	Hba-a1, HBB-y, Hba-x, Hbb-bh1	(disponible en Internet en www.cell.com/cell/comments/S0092-8674(18)30116-8)
27 Neuronas basales del prosencéfalo	Arx, DLX1, DLX2	(disponible en Internet en www.ncbi.nlm.nih.gov/pubmed/9256348?dopt=Abstract)
28 Células neuronales	Neurod2, Neurod6,	(disponible en Internet en www.cell.com/cell/comments/S0092-8674(18)30116-8)

Tipo de celda de grupo	Marcadores usados para identificación del tipo de célula	Bibliografía
piramidales	Tiam2, Kcnq3	8674(18)30116-8)
29 Hepatocitos	Afp, Alb, Apoa2, Afp29, Pik3c2g, Hoga1, TFEB	(disponible en Internet en www.cell.com/cell/comments/S0092-8674(18)30116-8)
30 Células de notocorda	Shh, Slit1, Ntn1	(disponible en internet en www.cell.com/trends/genetics/pdf/S0168-9525(04)00016-2.pdf)
31 Linfocitos	Apoe, Lyz2, Selenop, Ptprc, Ly86, CTSS	(disponible en Internet en www.cell.com/cell/comments/S0092-8674(18)30116-8)
32 Células endodermias	Sostdc1, Htr2c, Kcne2, Ttr	(disponible en Internet en www.cell.com/cell/comments/S0092-8674(18)30116-8)
33 Neuronas colinérgicas	Slit2, Slit3, Chat	(disponible en Internet en www.ncbi.nlm.nih.gov/pubmed/10349621?dopt=Abstract)
34 Linaje del músculo cardíaco	Myl2, Myocd, Hcn4, Ctnna3, Ryr2, Tbx20	(disponible en Internet en onlinelibrary.wiley.com/doi/10.1002/dvg.22819/epdf)
35 Megacariocitos	Pf4, Itgb3, Itga2b, Pbbp, Cd226	(disponible en Internet en www.cell.com/cell/comments/S0092-8674(18)30116-8)
36 Melanocitos	Tyr, TRPM1, Pmel	(disponible en Internet en www.ncbi.nlm.nih.gov/pubmed/1557387?dopt=Abstract)
37 Cristalino	GJA8, Cryba1, CRYAA	(disponible en Internet en ghr.nlm.nih.gov/gene/GJA8)
38 Monocitos/granulocitos	Ngp, S100A8	(disponible en Internet en www.cell.com/cell/comments/S0092-8674(18)30116-8)

Para la identificación de subgrupos, seleccionamos células de alta calidad (UMI > 400) en cada tipo de célula principal y aplicamos PCA, t-SNE y la agrupación Louvain de manera similar al análisis de grupos generales. Los subgrupos altamente sesgados se filtraron si la mayoría de las células (>50 %) del grupo eran de un solo embrión. Se fusionaron subgrupos muy similares si sus transcriptomas agregados estaban altamente correlacionados (coeficiente de correlación de Pearson > 0,95) y los dos grupos estaban cerca entre sí en el espacio t-SNE. Se identificaron genes expresados diferencialmente en subgrupos para cada tipo de célula principal como se describe anteriormente.

Para estimar el número de células de cada tipo de célula (o subtipos de células), primero calculamos la proporción de cada tipo de célula en un embrión individual y luego multiplicamos la proporción por el número total estimado de células para cada embrión (E9.5: 200.000, E10.5: 1.100.000; E11.5: 2.600.000; E12.5: 6.100.000; E13.5: 13.000.000).

Para identificar tipos de células (o subtipos de células) específicos del sexo, primero calculamos el número de células en cada tipo de célula (tipo de subcélula) para machos y hembras en cinco etapas de desarrollo. La relación específica del tipo de célula entre machos y hembras se comparó con la relación general del número de células entre machos y hembras en cada etapa de desarrollo. Luego aplicamos la prueba binomial en R para identificar tipos de células o subtipos de células con diferencias significativas entre machos y hembras en cada tipo de célula (x y n son el número de células hembras y el total de células en cada tipo de células de cada etapa de desarrollo, p es la relación de células hembras en cada etapa de desarrollo). El valor p se convierte en valor q ajustado mediante el método de Benjamini y Hochberg con la función p.adjust en R.

Análisis de pseudotiempo del mesénquima de las extremidades y AER.

Monocle 2 realizó el orden pseudotemporal de las células AER, las extremidades anteriores o las extremidades posteriores.⁷⁴. Brevemente, se identificaron genes expresados diferencialmente en cinco etapas de desarrollo con la función differentialGeneTest de Monocle 2⁷⁴. Los 500 genes principales con el valor q más bajo se usaron para construir la trayectoria del pseudotiempo usando Monocle 2⁷⁴, con el recuento de UMI por célula como covariable en la construcción del árbol. A cada célula se le asignó un valor de pseudotiempo en base a su posición a lo largo del árbol de trayectoria. El cambio suavizado en la expresión del marcador génico a lo largo del pseudotiempo se generó mediante la función representación_genes_en_pseudotiempo en Monocle 2⁷⁴. Las células en la trayectoria se agruparon con el mismo método que⁷⁷. Brevemente, las células se agruparon primero en posiciones similares en pseudotiempo mediante agrupación de k-medias a lo largo del eje del pseudotiempo ($k = 10$). Estos grupos se subdividieron en grupos que contenían al menos 50 y no más de 100 células. Luego agregamos los perfiles transcriptomas de las células dentro de cada grupo. La expresión génica a lo largo del pseudotiempo se calculó con el mismo enfoque que⁷⁷. Brevemente, se seleccionaron genes que pasaron una prueba significativa (FDR del 5 %) en diferentes condiciones de tratamiento y se usó una tira natural para ajustar la expresión génica a lo largo del pseudotiempo, con el número_medio_genes incluido como covariable. La expresión génica de cada gen se restó por la expresión más baja y luego se dividió por la expresión más alta. Los genes con expresión máxima dentro del 20 % inicial del pseudotiempo se marcaron como genes activados. Los genes con expresión máxima en el último 20 % del pseudotiempo se marcaron como genes reprimidos. Otros genes se marcaron como genes transitorios. Los términos del reactoma enriquecido (Reactome_2016) y los factores de transcripción (ChEA_2016) se identificaron usando el paquete EnrichR⁷⁸.

Inferencia de trayectoria con Monocle 3

El flujo de trabajo de Monocle 3 consiste en 3 etapas principales para organizar las células en trayectorias potencialmente discontinuas, seguidos de pruebas estadísticas opcionales para encontrar genes cuya expresión varía a lo largo de esas trayectorias. Monocle 3 también incluye herramientas de visualización para ayudar a explorar trayectorias en tres dimensiones.

Reducción de dimensionalidad con aproximación y proyección de colector uniforme (UMAP)

Monocle 3 primero proyecta los datos en un espacio de baja dimensión, lo que facilita el aprendizaje de un gráfico principal que describe cómo las células transitan entre estados transcriptómicos. Monocle 3 lo hace con UMAP, un algoritmo propuesto recientemente basado en geometría de Riemann y topología algebraica para realizar reducción de dimensiones y visualización de datos⁷⁹. Su calidad de visualización es competitiva con el popular método t-SNE (incrustación de vecino t-estocástico) usado ampliamente en transcriptómica unicelular. Sin embargo, mientras que t-SNE tiene como objetivo principal colocar células muy similares en las mismas regiones de un espacio de baja dimensión, UMAP también preserva relaciones de distancia de mayor alcance. El algoritmo UMAP en sí mismo también es más eficaz (la complejidad del algoritmo de UMAP es $O(N)$ frente a $O(N \log(N))$ para t-SNE). Brevemente, UMAP primero construye una representación topológica de los datos de alta dimensión con aproximaciones y parches múltiples locales y junto a sus representaciones locales difusas de conjuntos simpliciales. Luego, UMAP optimiza la integración de dimensiones inferiores, minimizando la entropía cruzada entre la representación de dimensiones bajas y la de dimensiones altas.

La eficiencia computacional de UMAP aceleró drásticamente el análisis de los datos de embriones de ratón. Descubrimos que UMAP termina de analizar un conjunto de datos de dos millones de células en 3 horas, mientras que t-SNE tarda más de 10 horas con 10 núcleos (se usa el bh-t-SNE multinúcleo). Unos pocos detalles de implementación conducen a la eficacia de UMAP. Tanto en el algoritmo UMAP como en t-SNE están implicados dos pasos principales: en primer lugar, se construye una estructura intermedia del espacio de alta dimensión (normalmente el espacio reducido PCA superior) y luego se encuentra una integración de baja dimensión que representa la estructura intermedia. Para la segunda etapa, ambos métodos usaron un enfoque de descenso de cuadrícula estocástico con diferentes funciones de pérdida para integrar los datos en un espacio de baja dimensión. Mientras que t-SNE necesita una función de pérdida para la normalización global, UMAP utiliza una función objetivo diferente que evita esa necesidad. Básicamente, esta etapa permite que UMAP escale linealmente con el número de muestras de datos. En Monocle 3, interactuamos con la implementación de Python UMAP (disponible en internet en atgithub.com/lmcinnes/umap) de Leland McInnes y John Healy a través de paquete reticulado (disponible en la internet atcran.r-project.org/web/packages/reticulate/index.html).

Reparto de células en trayectorias discontinuas.

Recientemente, Wolf y sus colegas propusieron la idea de organizar los datos del transcriptoma unicelular en un "gráfico de reparto abstracto" (AGA) que relaciona grupos de células que podrían estar relacionadas entre sí desde el punto de vista del desarrollo. Brevemente, su algoritmo construye un gráfico de k vecinos más cercanos en las células y luego identifica "comunidades" de células mediante el método de Louvain, similar a los métodos anteriores para analizar CyTOF o datos de RNA-seq unicelulares⁸⁰. Luego, AGA construye un gráfico en el que los vértices son comunidades de Louvain. Dos vértices están unidos con un borde en el gráfico AGA cuando las células de las

respectivas comunidades son vecinas en el gráfico kNN con más frecuencia de lo que se esperaría en un modelo binomial simple⁸¹. También se desarrollaron y aplicaron recientemente métodos similares para analizar conjuntos de datos de atlas de células de pez cebra y xenopus^{82,83}.

Monocle 3 se basa en estas ideas: primero construye un gráfico kNN en las células en el espacio UMAP, luego las agrupa en comunidades de Louvain y prueba cada par de comunidades para detectar una cantidad significativa de vínculos entre sus respectivas células. Aquellas comunidades que tienen más vínculos de los esperados según la hipótesis nula de unión espuria (FDR <10 %) permanecen conectadas en el gráfico AGA, y aquellos vínculos que no superan esta prueba se cortan. El gráfico AGA resultante tendrá uno o más componentes, cada uno de los cuales pasa a la siguiente etapa (gráfico L1) como un grupo separado de células que se organizarán en una trayectoria. El algoritmo AGA esencialmente se detiene en esta etapa, presentando el gráfico AGA como una especie de trayectoria de grano grueso en cada comunidad que refleja un estado diferente que las células pueden adoptar a medida que se desarrollan. Por el contrario, como se describe en la siguiente sección, Monocle 3 usa el gráfico AGA para restringir el espacio de los gráficos principales que pueden formar la trayectoria final. Es decir, Monocle 3 usa el gráfico AGA de grano grueso para aprender una trayectoria de grano fino.

La implementación de Monocle 3 de los procedimientos anteriores se escala a millones de células. Brevemente, usa la función agrupación louvain del paquete igraph para realizar la detección de la comunidad. A continuación, los cálculos básicos de AGA de Wolf. *et al* se calculan mediante una serie de operaciones matriciales dispersas. Deja ser a X una matriz (escasa) que representa la membresía comunitaria de las células. Cada columna de X representa una comunidad de Louvain y cada fila de X corresponde a una célula en particular. $X_{ij} = 1$ si celda i pertenece a la comunidad de Louvain j , de lo contrario 0. Podemos obtener además la matriz de adyacencia A del gráfico kNN usado para realizar la agrupación de louvain donde $A_{ij} = 1$ si la célula i se conecta a j en el gráfico kNN. Entonces la matriz de conexión M entre cada grupo se calcula como,

$$M = X^T A X$$

Una vez que M está construido, entonces podemos seguir *Nota complementaria 3.1* de la referencia⁸¹ para calcular la importancia de la conexión entre cada agrupación de Louvain y considerar cualquier agrupación con un valor p mayor que 0,05 de forma predeterminada como no desconectada.

Aprendiendo el gráfico principal

Monocle 3 aprende un gráfico principal que reside en el mismo espacio de baja dimensión que los datos para representar los posibles caminos que pueden tomar las células a medida que se desarrollan. Monocle 3 usa una implementación mejorada del algoritmo de gráfico L1⁸⁴ para aprender el gráfico principal. Mao *et al.* describió dos versiones del enfoque del gráfico L1⁸⁴. En el primero ("Algoritmo 1"), optimizan con respecto a todos los puntos de datos individuales del conjunto de datos. Anteriormente, demostramos que, aunque el gráfico L1 se puede aplicar a datos de RNA-seq unicelulares, tiende a aprender gráficos muy ruidosos que no son resistentes a la reducción de resolución y el enfoque no se escala de manera efectiva a conjuntos de datos más allá de unos pocos cientos de celdas⁸⁵. En Qiu *et al.*, no exploramos el "Algoritmo 2", que primero selecciona un conjunto de puntos de datos "históricos" utilizando el algoritmo de agrupamiento K-means. Luego, el algoritmo optimiza frente a esta muestra mucho más pequeña de datos. Monocle 3 utiliza este enfoque, que cuando se aplica a células en el espacio UMAP, es robusto y, con algunas modificaciones clave, puede escalar a millones de células.

Nuestra implementación del Lgráfico L1 tiene algunas características clave que admiten el análisis de grandes conjuntos de datos y una recuperación sólida del gráfico principal. Primero, aprendemos el gráfico L1 en el espacio UMAP (por defecto, tridimensional). Usamos la agrupación de K-medioes para seleccionar celdas de referencia para acelerar la optimización. El número de células de puntos de referencia elegidas tiene un impacto en el tiempo de ejecución del algoritmo y en la calidad de la disolución: demasiados puntos de referencia conducirán a un problema de programación lineal inviable. Por lo tanto, determinamos el número de puntos de referencia de manera dependiente de los datos estableciendo que K sea tres veces el número de comunidades de Louvain detectadas entre las células, lo que en la práctica conduce a soluciones rápidas y estables.

La segunda optimización importante del gráfico L1 es que imponemos restricciones al espacio "factible" de todos los gráficos posibles W considerados por la optimización. Mao *et al.* Consideró todos los bordes posibles entre puntos de datos emblemáticos. Sin embargo, incluso con tan solo mil células de referencia, el problema de programación lineal puede volverse rápidamente inviable, porque el número de variables es función del número de bordes en el gráfico. En Monocle 3, solo admitimos bordes en el espacio factible que están en el árbol de expansión mínima (MST) construido en los puntos de referencia, o que están en el gráfico kNN (por defecto $k=3$) construido en los vértices que tienen valores impares en el MST. Finalmente, excluimos los bordes que unirían células en diferentes componentes conectados del gráfico AGA construido como se describe en la sección anterior.

Identificación de genes con expresión dependiente de la trayectoria

Para identificar genes cuya expresión varía a lo largo de una trayectoria de desarrollo, tomamos prestada una prueba

estadística comúnmente usada en el análisis de datos espaciales. La estadística de Moran I es una medida de autocorrelación espacial multidireccional y multidimensional. La estadística codifica las relaciones espaciales entre puntos de datos a través de un gráfico vecino más cercano, lo que la hace particularmente adecuada para analizar grandes conjuntos de datos de RNA-seq unicelulares.

5 La prueba de Moran I^{86} se define como

$$I = \frac{N \sum_i \sum_j w_{ij} (x_i - \bar{x})(x_j - \bar{x})}{W \sum_i (x_i - \bar{x})^2}$$

10 donde N es el número de celdas indexadas por i y j ; X es el valor de expresión del gen de interés; $\bar{x}_i$ ($\bar{x}_j$) es la media de la expresión génica de la célula i (o j 's) vecinos más cercanos; w_{ij} es una matriz de pesos definida por un gráfico vecino más cercano con cero en la diagonal (es decir, $w_{ii} = 0$) y $w_{ij} = 1/k_i$ donde k_i es el número de vecinos más cercanos; y W es la suma de todos w_{ij} .

Para identificar los vecinos más cercanos usados para crear la matriz de peso W , primero construimos un gráfico vecino más cercano (kNN) k (el valor predeterminado es 25) para todas las células en el espacio UMAP. También proyectamos cada célula a su nodo más cercano en el gráfico principal. Luego eliminamos todos los bordes del gráfico kNN que conectan las células que se proyectan en los nodos del gráfico principal y no comparten un borde.

15 En Monocle 3, implementamos la *prueba múltiple* función para identificar múltiples genes correlacionados que dependen en versiones modificadas de rutinas de paquete *spdep* para realizar la prueba I de Moran.

Referencias

1. Fogarty, NME et al. Genome editing reveals a role for OCT4 in human embryogenesis. *Nature* 550, 67-73 (2017).
2. Kojima, Y., Tam, O. H. y Tam, P. P. L. Timing of developmental events in the early mouse embryo. *Semin. Cell. Dev. Biol.* 34, 65-75 (2014).
3. Tam, P. P. L. y Loebel, D. A. F. Gene function in mouse embryogenesis: get set for gastrulation. *Nat. Rev. Genet.* 8, 368-381 (2007).
4. Rivera-Pérez, J. A. y Hadjantonakis, A.-K. The Dynamics of Morphogenesis in the Early Mouse Embryo. *Cold Spring Harb. Perspect. Biol.* 7, a015867 (2014).
5. Dickinson, M. E. et al. High-throughput Discovery of novel developmental phenotypes. *Nature* 537, 508-514 (2016).
6. Meehan, T.F. et al. Disease model Discovery from 3,328 gene knockouts by The International Mouse Phenotyping Consortium. *Nat. Genet.* 49, 1231-1238 (2017).
7. Shyer, A. E., Huycke, T. R., Lee, C., Mahadevan, L. y Tabin, C. J. Bending gradients: how the intestinal stem cell gets its home. *Cell* 161, 569-580 (2015).
8. Uygur, A. et al. Scaling Pattern to Variations in Size during Development of the Vertebrate Neural Tube. *Dev. Cell* 37, 127-135 (2016).
9. Gorkin, D. et al. Systematic mapping of chromatin state landscapes during mouse development. (2017). doi:10.1101/166652
10. Mayer, C. et al. Developmental diversification of cortical inhibitory interneurons. *Nature* 555, 457-462 (2018).
11. Lescroart, F. et al. Defining the earliest step of cardiovascular lineage segregation by single-cell RNA-seq. *Science* (2018). doi:10.1126/science.aao4174
12. Han, X. et al. Mapping the Mouse Cell Atlas by Microwell-Seq. *Cell* 172, 1091-1107.e17 (2018).
13. El Consorcio Tabula Muris, Quake, S. R., Wyss-Coray, T. y Darmanis, S. Transcriptomic characterization of 20 organs and tissues from mouse at single cell resolution creates a Tabula Muris. (2017). doi:10.1101/237446
14. Amini, S. et al. Haplotype-resolved whole-genome sequencing by contiguity-preserving transposition and combinatorial indexing. *Nat. Genet.* 46, 1343-1349 (2014).
15. Adey, A. et al. In vitro, long-range sequence information for de novo genome assembly via transposase contiguity. *Genome Res.* 24, 2041-2049 (2014).
16. Cusanovich, D. A. et al. Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* 348, 910-914 (2015).

17. Vitak, S. A. et al. Sequencing thousands of single-cell genomes with combinatorial indexing. *Nat. Methods* 14, 302-308 (2017).
18. Ramani, V. et al. Massively multiplex single-cell Hi-C. *Nat. Methods* 14, 263-266 (2017).
19. Cao, J. et al. Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science* 357, 661-667 (2017).
20. Mulqueen, RM et al. Scalable and efficient single-cell DNA methylation sequencing by combinatorial indexing. (2017). doi:10.1101/157230
21. Rosenberg, AB et al. Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science* (2018). doi:10.1126/science.aam8999
22. Zheng, G. X. Y. et al. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* 8, 14049 (2017).
23. Qiu, X. et al. Reversed graph embedding resolves complex single-cell developmental trajectories. (2017). doi:10.1101/110668
24. Fernández, T. et al. Disruption of contactin 4 (CNTN4) results in developmental delay and other features of 3p deletion syndrome. *Am. J. hum. Genet.* 74, 1286-1293 (2004).
25. Olson, J. M. et al. NeuroD2 is necessary for development and survival of central nervous system neurons. *Dev. Biol.* 234, 174-187 (2001).
26. Uittenbogaard, M., Baxter, K. K. y Chiaramello, A. NeuroD6 Genomic Signature Bridging Neuronal Differentiation to Survival via the Molecular Chaperone Network. *J. Neurosci. Res.* 88, 33 (2010).
27. Yang, A. et al. p63 is essential for regenerative proliferation in limb, craniofacial and epithelial development. *Nature* 398, 714-718 (1999).
28. McQualter, J. L., Yuen, K., Williams, B. y Bertoncello, I. Evidence of an epithelial stem/progenitor cell hierarchy in the adult mouse lung. *Proc. Natl. Acad. Sci. U.S.A.* 107, 1414-1419 (2010).
29. Cichorek, M., Wachulska, M., Stasiewicz, A. y Tyminińska, A. Skin melanocytes: biology and development. *Advances in Dermatology and Allergology* 1, 30-41 (2013).
30. Tomihari, M., Hwang, S.-H., Chung, J.-S., Cruz, P. D., Jr. y Ariizumi, K. Gpnmb is a melanosome-associated glycoprotein that contributes to melanocyte/keratinocyte adhesion in a RGD-dependent fashion. *Exp. Dermatol.* 18, 586-595 (2009).
31. Varjosalo, M. y Taipale, J. Hedgehog: functions and mechanisms. *Genes Dev.* 22, 2454-2472 (2008).
32. Strähle, U., Lam, C. S., Ertzer, R. y Rastegar, S. Vertebrate floor-plate specification: variations on common themes. *Trends Genet.* 20, 155-162 (2004).
33. Holmes, G. P. et al. Distinct but overlapping expression patterns of two vertebrate slit homologs implies functional roles in CNS development and organogénesis. *Mech. Dev.* 79, 57-72 (1998).
34. Akle, V. et al. F-spondin/spn1b expression patterns in developing and adult zebrafish. *PLoS One* 7, e37593 (2012).
35. Hartman, B. H., Durruthy-Durruthy, R., Laske, R. D., Losorelli, S. y Heller, S. Identification and characterization of mouse otic sensory lineage genes. *Front. Cell. Neurosci.* 9, 79 (2015).
36. Petit, F., Sears, K. E. y Ahituv, N. Limb development: a paradigm of gene regulation. *Nat. Rev. Genet.* 18, 245-258 (2017).
37. Guo, Q., Loomis, C. y Joyner, A. L. Fate map of mouse ventral limb ectoderm and the apical ectodermal ridge. *Dev. Biol.* 264, 166-178 (2003).
38. Lewandoski M, E. al. Fgf8 signalling from the AER is essential for normal limb development. - PubMed - NCBI. Disponible en: <https://www.ncbi.nlm.nih.gov/pubmed/11101846>. (Consultado: 22 de abril de 2018)
39. Aoki M, E. al. R-spondin2 expression in the apical ectodermal ridge is essential for outgrowth and patterning in mouse limb development. - PubMed - NCBI. Disponible en: <https://www.ncbi.nlm.nih.gov/pubmed/18067586>. (Consultado: 22 de abril de 2018)
40. Gerdes, J., Schwab, U., Lemke, H. y Stein, H. Production of a mouse monoclonal antibody reactive with a human

nuclear antigen associated with cell proliferation. *Int. J. Cancer* 31, 13-20 (1983).

41. Bergman, D., Halje, M., Nordin, M. y Engström, W. Insulin-like growth factor 2 in development and disease: a mini-review. *Gerontology* 59, 240-249 (2013).

42. Trapnell, C. et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* 32, 381-386 (2014).

43. McInnes, L. y Healy, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. (2018).

44. Alexander Wolf, F. et al. Graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *bioRxiv* 208819 (2017). doi:10.1101/208819

45. Qiu, X. et al. Reversed graph embedding resolves complex single-cell trajectories. *Nat. Methods* 14, 979-982 (2017).

46. Singh, M. K. et al. The T-box transcription factor Tbx15 is required for skeletal development. *Mech. Dev.* 122, 131-144 (2005).

47. Paine-Saunders, S., Viviano, B. L., Zupicich, J., Skarnes, W. C. y Saunders, S. glypican-3 controls cellular responses to Bmp4 in limb patterning and skeletal development. *Dev. Biol.* 225, 179-187 (2000).

48. Hara, K. y Ide, H. Msx1 expressing mesoderm is important for the apical ectodermal ridge (AER)-signal transfer in chick limb development. *Dev. Growth Differ.* 39, 705-714 (1997).

49. Lupiáñez, D. G. et al. Disruptions of Topological Chromatin Domains Cause Pathogenic Rewiring of Gene-Enhancer Interactions. *Cell* 161, 1012-1025 (2015).

50. Davis, R. J. et al. Dach1 mutant mice bear no gross abnormalities in eye, limb, and brain development and exhibit postnatal lethality. *Mol. Cell. Biol.* 21, 1484-1490 (2001).

51. Akiyama, H., Chaboissier, M.-C., Martin, J. F., Schedl, A. y de Crombrughe, B. The transcription factor Sox9 has essential roles in successive steps of the chondrocyte differentiation pathway and is required for expression of Sox5 and Sox6. *Genes Dev.* 16, 2813-2828 (2002).

52. Deng, Y. et al. Yap1 Regulates Multiple Steps of Chondrocyte Differentiation during Skeletal Development and Bone Repair. *Cell Rep.* 14, 2224-2237 (2016).

53. Joshi, S. et al. TEAD transcription factors are required for normal primary myoblast differentiation in vitro and muscle regeneration in vivo. *PLoS Genet.* 13, e1006600 (2017).

54. Knapp, D. et al. Comparative transcriptional profiling of the axolotl limb identifies a tripartite regeneration-specific gene program. *PLoS One* 8, e61352 (2013).

55. Zeller, R., López-Ríos, J. y Zuniga, A. Vertebrate limb bud development: moving towards integrative analysis of organogenesis. *Nat. Rev. Genet.* 10, 845-858 (2009).

56. Nishimoto, S., Minguillon, C., Wood, S. y Logan, M. P. O. A combination of activation and repression by a colinear Hox code controls forelimb-restricted expression of Tbx5 and reveals Hox protein specificity. *PLoS Genet.* 10, e1004245 (2014).

57. Vargesson, N., Luria, V., Messina, I., Erskine, L. y Laufer, E. Expression patterns of Slit and Robo family members during vertebrate limb development. *Mech. Dev.* 106, 175-180 (2001).

58. Chimal-Monroy, J. et al. Analysis of the molecular cascade responsible for mesodermal limb chondrogenesis: Sox genes and BMP signaling. *Dev. Biol.* 257, 292-301 (2003).

59. Braun, T. y Gautel, M. Transcriptional mechanisms regulating skeletal muscle differentiation, growth and homeostasis. *Nat. Rev. Mol. Cell. Biol.* 12, 349-361 (2011).

60. Tajbakhsh, S., Rocancourt, D., Cossu, G. y Buckingham, M. Redefining the genetic hierarchies controlling skeletal myogenesis: Pax-3 and Myf-5 act upstream of MyoD. *Cell* 89, 127-138 (1997).

61. Harel, I. et al. Distinct origins and genetic programs of head muscle satellite cells. *Dev. Cell* 16, 822-832 (2009).

62. Sambasivan, R. et al. Distinct regulatory cascades govern extraocular and pharyngeal arch muscle progenitor cell fates. *Dev. Cell* 16, 810-821 (2009).

63. Heimberg, G., Bhatnagar, R., El-Samad, H. y Thomson, M. Low Dimensionality in Gene Expression Data Enables

the Accurate Extraction of Transcriptional Programs from Shallow Sequencing. *Cell Syst* 2, 239-250 (2016).

64. Cusanovich, D. A. et al. The cis-regulatory dynamics of embryonic development at single cell resolution. (2017). doi:10.1101/166066

65. McKenna, A. et al. Whole-organism lineage tracing by combinatorial and cumulative genome editing. *Science* 353, aaf7907 (2016).

66. Osterwalder, M. et al. Enhancer redundancy provides phenotypic robustness in mammalian development. *Nature* 554, 239-243 (2018).

67. Dickel, D. E. et al. Ultraconserved Enhancers Are Required for Normal Development. *Cell* 172, 491-499.e15 (2018).

68. Li, D. et al. Formation of proximal and anterior limb skeleton requires early function of *Ir3* and *Ir5* and is negatively regulated by *Shh* signaling. *Dev. Cell* 29, 233-240 (2014).

69. Kraft, K. et al. Deletions, Inversions, Duplications: Engineering of Structural Variants using CRISPR/Cas in Mice. *Cell Rep.* (2015). doi:10.1016/j.celrep.2015.01.016

70. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. y Greenleaf, W. J. Transposition of native chromatin form fast and sensitive epigenomic profiling of open chromatin DNA-binding proteins and nucleosome position. *Nat. Methods* 10, 1213-1218 (2013).

71. Renaud, G., Stenzel, U., Maricic, T., Wiebe, V. y Kelso, J. deML: robust demultiplexing of Illumina sequences using a likelihood-based approach. *Bioinformatics* 31, 770-772 (2015).

72. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15-21 (2013).

73. Anders, S., Pyl, P. T. y Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* 30, 1693-1694 (2014).

74. Qiu, X. et al. Reversed graph embedding resolves complex single-cell developmental trajectories. (2017). doi:10.1101/110668

75. Wolf, F. A., Angerer, P. y Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* 19, 15 (2018).

76. Cao, J. et al. Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science* 357, 661-667 (2017).

77. Pliner, H. et al. Chromatin accessibility dynamics of myogenesis at single cell resolution. (2017). doi:10.1101/155473

78. Kuleshov, M. V. et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* 44, W90-7 (2016).

79. McInnes, L. y Healy, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. (2018).

80. Levine, J. H. et al. Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. *Cell* 162, 184-197 (2015).

81. Wolf, F. A. et al. Graph abstraction reconcile clustering with trajectory inference through a topology preserving map of single cells. (2017). doi: 10.1101/208819

82. Wagner, D. E. et al. Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science* 359, 473-478 (2018).

83. Briggs, J. A. et al. The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution. *Science* 359, 473-478 (2018).

84. Mao, Q., Wang, L., Tsang, I. y Sun, Y. Principal Graph and Structure Learning Based on Reversed Graph Embedding. *IEEE Trans. Pattern Anal. Mach. Intell.* (2016). doi:10.1109/TPAMI.2016.2635657

85. Qiu, X. et al. Reversed graph embedding resolves complex single-cell trajectories. *Nat. Methods* 14, 979-982 (2017).

86. Moran, P. A. P. Notes on continuous stochastic phenomena. *Biometrika* 37, 17-23 (1950).

Ejemplo 2

Una nueva técnica para la extracción y fijación de núcleos tisulares (sc-RNA-seq)

Reactivos. BSA (grado de biología molecular, NEB, núm. B9000S); inhibidor de superRnasa (Thermo, núm. AM2696); EMS 157-4-100 disolución acuosa de paraformaldehído (formaldehído) al 4 %, grado EM, 100 ml (Amazon).

- 5 Tampones. Tampón de núcleos (almacenado a 4 °C): Tris-HCl 10 mM, pH 7,4, NaCl 10 mM, MgCl₂ 3 mM. IGEPAL CA-630 al 10 % (almacenado a 4 °C). Tampón de lavado de núcleos (recién preparado cada vez): 980 ul de tampón de núcleos con 10 ul de BSA y 10 ul de SuperRnasal, mezclar bien y almacenar en hielo. Tampón de lisis de núcleos (recién hecho cada vez): Tampón de lavado de núcleos con IGEPAL CA-630 al 0,1 %.

Extracción de núcleos directamente del tejido

- 10 Los tejidos se pican en trozos pequeños con una cuchilla en 1 ml de tampón de lisis celular enfriado con hielo (Tris-HCl 10 mM, pH 7,4, NaCl 10 mM, MgCl₂ 3 mM y IGEPAL CA-630 al 0,1 %, SUPERasa In al 1 % y BSA al 1 %) y se transfirió a la parte superior de un colador de células de 40 um (Falcon).

Los tejidos se homogeneizaron con la punta de goma de un émbolo de jeringa (5 ml, BD) en 4 ml de tampón de lisis celular.

- 15 Luego, los núcleos filtrados se transfirieron a un tubo nuevo de 15 ml (Falcon) y se granularon mediante centrifuga a 500 xg durante 5 minutos y se lavaron una vez con 1 ml de tampón de lisis celular.

Fijación de núcleos

Los núcleos se fijaron en 4 ml de paraformaldehído al 4% (EMS) enfriado con hielo durante 15 minutos en hielo.

Después de la fijación, los núcleos se lavaron dos veces en 1 ml de tampón de lavado de núcleos (tampón de lisis celular sin IGEPAL) y se resuspendieron en 500 ul de tampón de lavado de núcleos.

- 20 Las muestras se dividieron en varias y se congelaron instantáneamente en nitrógeno líquido. Las muestras congeladas se pueden transportar en hielo seco.

Ejemplo 3

Caracterización de la dinámica de transición del estado de una sola célula por sci-fate

- 25 La belleza del desarrollo reside en la generación de diversos estados celulares en un orden temporal estrictamente organizado. A pesar de la proliferación de técnicas genómicas unicelulares, sigue siendo un desafío determinar cuantitativamente la dinámica de transición del estado celular. Aquí presentamos sci-fate, un ensayo combinatorio de alto rendimiento basado en indexación para perfilar tanto el transcriptoma completo como el recién sintetizado en cada una de miles de células individuales. Como prueba de concepto, aplicamos sci-fate a un sistema modelo de respuesta de cortisol y caracterizamos más de 6.000 eventos de transición de estado de una sola célula, coherente con la
- 30 dinámica conocida del ciclo celular tras la activación del receptor de glucocorticoides. A partir del análisis, mostramos que la dirección de transición del estado celular y las probabilidades están reguladas por las distancias entre estados y el panorama de inestabilidad del estado. La técnica y los enfoques computacionales son fácilmente aplicables a otros sistemas biológicos para caracterizar cuantitativamente la dinámica del estado celular y descifrar el mecanismo interno para la determinación del destino celular.

- 35 Las células transitan a través de estados funcionales y molecularmente distintos durante el desarrollo del organismo multicelular. Caracterizar la ruta de transición del estado celular, o destino celular, es fundamental para comprender el desarrollo y aplicaciones como la ingeniería celular. Si bien han proliferado los métodos para técnicas genómicas unicelulares, solo capturan una instantánea del estado celular, por lo que no pueden proporcionar información sobre la dinámica de transición celular (1). Aunque el rastreo de células individuales basado en microscopía de lapso de
- 40 tiempo se puede usar para caracterizar las transiciones del estado celular (2, 3), su rendimiento es limitado y solo pueden rastrear los cambios de varios genes y, por lo tanto, tienen poca capacidad para descifrar sistemas complejos.

- Aquí describimos una estrategia novedosa para inferir la dinámica de transición cuantitativa del estado celular a nivel de transcriptoma completo. Esta estrategia depende de una nueva técnica de RNA-seq unicelular basada en indexación combinatoria, sci-fate. Al marcar el ARNm recién sintetizado con 4-tiouridina (4, 5), que generará
- 45 mutaciones puntuales C > T durante la transcripción inversa, sci-fate captura tanto el transcriptoma completo como el transcriptoma recién sintetizado a nivel unicelular, junto con la información degradada del transcriptoma de su estado pasado (memoria de estado pasado). Luego, la memoria del estado pasado de cada célula se corrige mediante la velocidad de degradación del ARNm (técnica de corrección de la memoria), de modo que cada célula puede caracterizarse por la dinámica del transcriptoma entre dos puntos temporales.

- 50 Para caracterizar la dinámica de transición del estado celular regulada por factores intrínsecos y extrínsecos, aplicamos sci-fate a un sistema modelo de respuesta de cortisol, en el que el destino celular fue impulsado por dos fuerzas principales: el programa del ciclo celular intrínseco y la activación del receptor de glucocorticoides (GR)

extrínseco inducido por fármacos. La activación de GR influye en la actividad de casi todas las células del cuerpo y regula los genes que controlan el desarrollo, el metabolismo y la respuesta inmune (6). Con sci-fate, perfilamos la dinámica del transcriptoma completo de más de 6.000 células individuales. En base a la similitud entre los estados del transcriptoma pasados y actuales, construimos miles de trayectorias de transición de estados celulares que abarcan cinco puntos temporales, que pueden agruparse en tres tipos de destinos celulares coherentes con patrones conocidos de progreso del ciclo celular en la activación de GR. Además, caracterizamos los estados ocultos celulares mediante la actividad de los módulos TF funcionales e inferimos una red de transición celular para la predicción del estado celular. Finalmente, mostramos que la dirección y la probabilidad de la transición del estado celular están reguladas por la similitud del transcriptoma y el panorama de inestabilidad de sus estados cercanos. Los enfoques teóricos, computacionales y experimentales desarrollados aquí deberían ser fácilmente aplicables a otros sistemas biológicos en los que aún se desconoce la dinámica de transición celular.

Descripción general de sci-fate

sci-fate se basa en las siguientes etapas (Fig. 30A): (i) las células se incubaron primero con 4-tiouridina (S4U), un análogo de timidina ampliamente usado para marcar el ARN recién sintetizado (7-13). (ii) Las células se cosechan, se fijan con paraformaldehído al 4 %, seguido de una reacción de alquilación ligada a tiol (SH) que une covalentemente un grupo carboxiamidometilo a S4U mediante sustitución nucleófila (4). (iii) Las células se distribuyeron en masa en cada pocillo de placas de 4x96 pocillos. El primer índice molecular RNA-seq se introduce en el ARNm de las células en cada pocillo a través de transcripción inversa *in situ* (RT) con un cebador poli(T) que lleva tanto un código de barras específico del pocillo como un identificador molecular único (UMI) degenerado. Durante la síntesis de ADNc, el ARNm marcado con S4U modificado imita las conversiones de timina a citosina (T > C) y da como resultado un ADNc de primera hebra mutado. (iv) Las células de todos los pocillos se agrupan y luego se redistribuyen mediante clasificación de células activadas por fluorescencia (FACS) en múltiples placas de 96 pocillos. Las células se clasifican mediante tinción DAPI (4',6-diamidino-2-fenilindol) para discriminar las células individuales de dobletes durante la clasificación. El ADNc bicatenario se genera mediante la degradación del ARN y síntesis de la segunda hebra, y se somete a transposición con Tn5. Luego, el ADNc se amplifica mediante la reacción en cadena de la polimerasa (PCR) con una combinación de cebadores que reconocen el adaptador Tn5 en el extremo 5' y el cebador RT en el extremo 3'. Estos cebadores también llevan un código de barras específico del pocillo que introduce el segundo índice molecular RNA-seq. (v) Los amplicones de la PCR se agrupan y se someten a una secuenciación paralela masiva. Al igual que con otros protocolos "sci-" (14-21), la mayoría de los núcleos pasan a través de una combinación única de pocillos y, por lo tanto, el contenido de cada célula está marcado por una combinación única de códigos de barras que se pueden usar para agrupar lecturas que se derivan de la misma célula. El ARNm recién sintetizado a partir del transcriptoma completo se identifica mediante conversiones "T > C" corregidas por error de fondo (método).

Como control de calidad, primero probamos la técnica en una mezcla de células HEK293T (humana) y NIH/3T3 (ratón) en cuatro condiciones: con o sin etiquetado S4U (200 nM, 6 h) y con o sin tratamiento con IAA (Fig. 31A-D). Con el etiquetado S4U y el tratamiento con IAA (condición de sci-fate), los transcriptomas de células humanas/de ratón fueron abrumadoramente coherentes con las especies (> 99 % de pureza tanto para células humanas como de ratón, 2,6 % de colisiones) con una alta relación de lecturas mutadas T > C detectadas (46 % para células humanas y 31 % para células de ratón en condiciones de sci-fate frente a 0,8 % para células humanas y 0,8 % para células de ratón sin condiciones de tratamiento). Obtuvimos una pureza celular aproximadamente equivalente en cuatro condiciones, aunque se detectaron UMI ligeramente más bajos en los grupos de tratamiento con IAA. Los transcriptomas agregados de sci-fate frente a sci-RNA-seq normal estaban altamente correlacionados (correlación de Spearman $r = 0,99$; Fig. 31E-F), lo que sugiere que el proceso de conversión y etiquetado a corto plazo tiene un efecto mínimo sobre el estado celular.

Perfil conjunto del transcriptoma total y recién sintetizado en células A549 tratadas con dexametasona

Luego aplicamos sci-fate a un modelo de respuesta del cortisol, en donde la dexametasona (DEX), un imitador sintético del cortisol, activa el receptor de glucocorticoides (GR), que se une a miles de posiciones en todo el genoma y altera significativamente el estado celular en un corto plazo (22-25). Tratamos células A549 derivadas de adenocarcinoma de pulmón durante 0, 2, 4, 6, 8 o 10 h con DEX 100 nM. En cada condición, las células se incubaron con S4U (200 nM) durante las últimas dos horas antes de la recolección para sci-fate de 384 x 192 pocillos (Fig. 30B). Cada una de las seis condiciones se representó en 64 pocillos durante la primera ronda de indexación para que la condición de tratamiento pudiera recuperarse en función del primer índice de cada célula.

Después de filtrar células de baja calidad, posibles dobletes y un pequeño subgrupo de células diferenciadas (Método), obtuvimos perfiles de células individuales para 6.680 células (mediana de 26.176 ARNm detectados por célula) con una mediana de 20 % de UMI marcados por célula (Fig. 30C, Figuras 32A-B). Las lecturas intrónicas muestran una tasa de nueva síntesis significativamente mayor que las lecturas exónicas (65 % en lecturas intrónicas frente a 13 % en lecturas exónicas, valor de $p < 2,2e-16$, prueba de rango con signo de Wilcoxon; Fig. 30D), coherente con la expectativa de que las lecturas intrónicas se enriquecen en el transcriptoma recién sintetizado.

Primero preguntamos si el transcriptoma completo y el transcriptoma recién sintetizado transmiten información diferente en la caracterización del estado celular. Agregamos el transcriptoma completo y el transcriptoma recién

sintetizado para cada condición de tratamiento y verificamos sus correlaciones. A diferencia del transcriptoma completo, el transcriptoma recién sintetizado mostró una marcada diferencia entre los grupos sin tratamiento con DEX (0 h) y los grupos tratados (Fig. 32C). En coherencia con esto, la reducción de dimensiones con Aproximación y proyección de variedad uniforme (UMAP) (26) en un transcriptoma completo o recién sintetizado da resultados diferentes (Fig. 30E): el transcriptoma completo no puede separar las células sin tratamiento con DEX (0 h) y con tratamiento temprano con DEX (2 h) mientras que el transcriptoma recién sintetizado agrega todas las células tratadas con DEX en un solo grupo. Los grupos de células identificados por el transcriptoma completo o recién sintetizado no coinciden completamente entre sí (Fig. 30F, Fig. 32D-E). Esto es de esperar ya que el transcriptoma recién sintetizado refleja directamente la actividad del promotor del gen o la respuesta epigenética al entorno externo, mientras que el transcriptoma completo está determinado principalmente por el ARNm sobrante de su estado anterior.

Para caracterizar los estados celulares con información conjunta, combinamos los principales componentes principales (PC) del transcriptoma completo y recién sintetizado para el análisis UMAP. La información conjunta separa las células en ningún tratamiento con DEX (0 h), tratamiento temprano (2 h) y tratamiento tardío (>2 h) (Fig. 30E). Curiosamente, dos grupos (grupos 1 y 4) caracterizados por un transcriptoma completo se dividieron en cuatro grupos separados mediante información conjunta (Fig. 30F). Verificamos el nivel de expresión y la nueva velocidad de síntesis de marcadores genéticos relacionados con el ciclo celular (27) (Fig. 30G, Fig. 32F-G): los grupos recién separados mediante información conjunta corresponden a la fase G2/M (alta expresión y alta velocidad de síntesis de marcadores G2/M) y células de fase temprana G0/G1 (alta expresión y baja velocidad de síntesis de marcadores G2/M). Esto sugiere que el transcriptoma recién sintetizado transmite información sobre el estado celular diferente en comparación con el transcriptoma completo, y la información conjunta permite potencialmente una mayor resolución en la caracterización del estado celular.

Caracterización de módulos TF funcionales que impulsan la determinación del destino celular

A continuación, buscamos caracterizar los módulos TF que impulsan la transición del estado celular. Los vínculos entre los factores de transcripción (TF) y sus genes regulados se identificaron mediante dos etapas: para cada gen, calculamos correlaciones entre la velocidad de síntesis de ARNm durante las últimas dos horas y el nivel de expresión de TF en más de 6.000 células usando LASSO (operador de contracción y selección mínima absoluta). Estos vínculos identificados se filtraron aún más mediante datos CHIP-seq publicados (28) o análisis de enriquecimiento de motivos (29) (Método). En total, identificamos 986 vínculos entre 29 TF y 532 genes (Fig. 33A, Tabla S1), según la covarianza del gen TF y validados mediante datos de unión al ADN. Para evaluar la posibilidad de que los vínculos fueran artefactos de regresión regularizada, permutamos los ID de muestra de la matriz de expresión de TF y realizamos el mismo análisis. No se identificaron vínculos después de esta permutación.

Se identifican los módulos TF que impulsan la respuesta de GR, incluidos los efectores de respuesta de GR conocidos como CEBPB(30) (Fig. 34A-B), FOXO1(31), y JUNB(32) (Fig. 33A). También encontramos varios módulos nuevos de TF relacionados con la respuesta de GR, incluidos YOD1 y GTF2IRD1, tanto con expresión como actividad sobreexpresadas en células tratadas con DEX (Fig. 34C-D). Se identifican los principales módulos TF que impulsan la progresión del ciclo celular, y estos incluyen E2F1, E2F2, E2F7, BRCA1 y MYBL2 (33). En comparación con el nivel de expresión total, la nueva velocidad de síntesis de ARN de genes regulados por los módulos TF del ciclo celular muestra una mayor correlación con la expresión de TF diana (Fig. 34E). Además, también encontramos módulos TF relacionados con la diferenciación celular como GATA3, expresado principalmente en un grupo de poblaciones de células inactivas (34), y módulos TF relacionados con la respuesta al estrés oxidativo como NRF1 (35) y NFE2L2 (NRF2) (36).

A continuación, caracterizamos la actividad de TF agregando la nueva velocidad de síntesis de ARN de genes dentro de cada módulo de TF y calculamos el coeficiente de correlación absoluta entre cada par de TF (Fig. 34F). La actividad TF altamente correlacionada sugiere que pueden funcionar en un proceso vinculado. La agrupación jerárquica segrega estos 29 módulos TF en cinco módulos principales (Fig. 34F): el primer módulo son todos los módulos TF relacionados con el ciclo celular, como E2F1 y FOXM1 (33), y representa la fuerza impulsora para la progresión del ciclo celular. El tercer módulo son todos los módulos TF relacionados con la respuesta de GR, como FOXO1, CEBPB, JUNB y RARB (30)(31)(32). Los otros grupos de módulos TF incluyen tres TF (KLF6, TEAD1 y YOD1) coregulados tanto por el ciclo celular como por la respuesta GR (módulo 2), una vía de diferenciación interna que incluye GATA3 y AR (módulo 3) y TF relacionados con la respuesta al estrés, como NRF1 y NFE2L2 (módulo 5).

Para identificar diferentes estados del ciclo celular, primero ordenamos las células por actividad del módulo TF vinculado al ciclo celular. Las células se ordenan en una trayectoria suave del ciclo celular, se validan por la velocidad de síntesis de marcadores conocidos del ciclo celular (27) (Fig. 33B). Observamos un hueco entre la fase G2/M y la fase G0/G1, coherente con el dramático cambio de estado celular durante la división celular. Mediante agrupamiento no supervisado, identificamos nueve estados del ciclo celular que abarcan las fases del ciclo celular G0/G1, S y G2/M en base a la expresión del marcador del ciclo celular (Fig. 33B). Las células se pueden ordenar en otra trayectoria suave mediante módulos TF vinculados a respuesta GR. La trayectoria se correlaciona bien con el tiempo de tratamiento con DEX y la dinámica de la actividad TF regulada por activación de GR conocida (Fig. 33C). Mediante análisis de agrupamiento no supervisado, identificamos tres grupos de células a lo largo de la respuesta de GR, correspondientes a un estado de respuesta de GR nulo/bajo/alto (Fig. 33C).

A continuación, buscamos caracterizar cuantitativamente los estados celulares ocultos en el sistema (Fig. 35A). En la Fig. 33B-C se identificaron nueve estados del ciclo celular y tres estados de respuesta de GR. Se identificaron todos los estados combinatorios posibles, y el grupo más pequeño incluía el 1,1 % (74) de todas las células (Fig. 33D). La proporción del estado celular observada es cercana a la proporción esperada suponiendo una distribución independiente. Esto es coherente con el bajo coeficiente de correlación (correlación de Pearson $r = 0,004$) entre la actividad de estos dos módulos TF funcionales en más de 6.000 células. A modo de comparación, mediante reducción de dimensiones y análisis de agrupamiento en el transcriptoma completo y recién sintetizado, identificamos 6 grupos principales (Fig. 35B). Estos grupos principales pueden definirse fácilmente mediante grupos combinados de estos 27 estados celulares (Fig. 33E).

Caracterización de la trayectoria de transición de una sola célula y la red de transición de estado

Tanto con el transcriptoma completo como con el transcriptoma recién sintetizado caracterizados para cada célula, podemos inferir el estado del transcriptoma de una sola célula antes del etiquetado con S4U (Fig. 36A). La recuperación del transcriptoma celular pasado depende de dos parámetros: la velocidad de detección de lecturas recién sintetizadas en sci-fate y la velocidad de degradación (o tiempo medio) de cada ARNm (método). Ambos parámetros pueden estimarse a partir del mismo experimento de sci-fate.

Primero estimamos la velocidad de detección de sci-fate. Suponemos que la vida media del ARNm es estable en diferentes condiciones de tratamiento con DEX. Esta suposición se valida aún más mediante una verificación de autocorrelación posterior. Bajo esta suposición, el transcriptoma en masa parcialmente degradado antes del etiquetado S4U de 2 horas debe ser el mismo entre células sin tratar con DEX y células tratadas con DEX de 2 horas. Por lo tanto, sus diferencias en el transcriptoma completo (en masa) deberían ser iguales a sus diferencias en el transcriptoma recién sintetizado (en masa) corregido por la velocidad de detección de la técnica. Como en nuestro experimento se perfilan tanto el transcriptoma completo como el recién sintetizado, podemos calcular directamente la velocidad de detección de sci-fate. Las diferencias en el ARNm recién sintetizado se correlacionan bien con las diferencias en el nivel de expresión del ARNm (r de Pearson = 0,93, Fig. 37A), lo que sugiere que la velocidad de detección de ARN nuevo es bastante estable en todos los genes. Por lo tanto, utilizamos la mediana de la velocidad de captura de ARN nuevo (82 %) para el análisis posterior.

Luego calculamos la velocidad de degradación del ARNm en 2 horas. Como la población de células A549 puede considerarse estable sin perturbaciones externas, para las células después de 2 horas de tratamiento con DEX, su estado anterior (antes del etiquetado con S4U de 2 horas) debe ser el mismo que el de las células tratadas con DEX de 0 horas. De manera similar, el estado anterior (antes del etiquetado con S4U) para las células tratadas con DEX con $T = 0/2/4/6/8/10$ horas debe ser similar a las células perfiladas con $T = 0/0/2/4/6/8$ horas. Con el transcriptoma completo y el transcriptoma recién sintetizado perfilados para todas las condiciones de tratamiento, se puede estimar la velocidad de degradación del ARNm en miles de genes en cada intervalo de tiempo de 2 horas. Como verificación de autocorrelación mencionada anteriormente, las velocidades de degradación génica están altamente correlacionadas entre diferentes tiempos de tratamiento con DEX (Fig. 37B). Luego utilizamos la velocidad promedio de degradación de genes para el análisis posterior. Tanto con la nueva velocidad de detección de ARNm como con la velocidad de degradación de genes disponibles, estimamos el estado del transcriptoma pasado de una sola célula para que cada célula pueda caracterizarse por la dinámica del transcriptoma en un intervalo de dos horas.

Para recuperar la dinámica del estado celular durante un intervalo más largo (es decir, 10 horas), desarrollamos un canal de unión celular para vincular las células madre e hija en la misma trayectoria de transición del estado celular (Fig. 36A): para cada célula A (p. ej., células tratadas con DEX 2 horas), identificamos una célula B perfilada en el punto de tiempo anterior (p. ej., células no tratadas con DEX) y B tenía su estado actual similar al estado pasado de A, en base a una estrategia de alineación desarrollada recientemente para identificar estados de la célula comunes entre dos conjuntos de datos (27). B puede considerarse como el estado parental de A. De manera similar, también identificamos otra célula C perfilada en el punto temporal posterior (p. ej., células tratadas con DEX 4 horas) y C tenía su estado pasado similar al estado actual de A. La célula C puede considerarse como el estado futuro de A. Al extender la misma estrategia a todos los estados pasados y futuros identificados para cada célula, construimos 6.680 trayectorias de transición unicelular a lo largo de 10 horas y cinco puntos temporales (Fig. 36A-B). Es de destacar que este análisis se basa en la suposición de que el estado pasado y actual de cada célula (excepto las células en los puntos de tiempo inicial y final) se detectan de manera integral, lo que es cierto en nuestros conjuntos de datos, ya que se perfilan más de 6000 celdas (más de 1.000 células por condición), o una célula durante menos de un minuto durante el ciclo celular. Se perfilan múltiples celdas (>50) en cada estado celular, por lo que también se puede capturar el proceso estocástico de transición del estado celular.

Para validar el resultado, aplicamos reducción de dimensiones y análisis de agrupamiento no supervisado a estas 6.680 trayectorias de células individuales, que se agruparon en tres grupos de trayectorias. Verificamos la dinámica de los estados celulares caracterizados en la Fig. 36C. Como se esperaba, las tres trayectorias mostraron una transición del estado celular desde ninguna respuesta de GR a estados de respuesta de GR baja/alta a lo largo del tiempo (Fig. 36D). Observamos dinámicas distintas del ciclo celular a lo largo de estas tres trayectorias (Fig. 36D): la trayectoria 1 mostró una fase G2/M disminuida y una fase G0/G1 constantemente aumentada, y representó la transición del estado celular de los estados intermedios G2/M y G1 a la fase G1. La trayectoria 2 mostró una transición

del estado celular desde los estados intermedios S y G2/M a la fase G2/M. En la trayectoria 3, observamos la transición del estado celular de la fase intermedia G1 y S a la fase S temprana durante el tratamiento temprano con DEX (0-2 horas), pero la transición se inhibe en las condiciones tardías del tratamiento con DEX (>2 horas de tratamiento con DEX), lo que sugiere que el tratamiento con DEX a largo plazo da por resultado la detención de la fase G1. Esto es coherente con los cambios en la proporción del estado celular a lo largo del tiempo de tratamiento y con investigaciones previas (37, 38) (Fig. 36D). Esto sugiere que las rutas de transición de una sola célula caracterizadas por sci-fate pueden recuperar direcciones generales de transición del estado celular.

Con múltiples células (>70) perfiladas en cada estado, calculamos la probabilidad de transición del estado celular en los 27 estados ocultos. Las transiciones del estado celular con bajas probabilidades de transición ($< 0,1$) se deben potencialmente a eventos raros o ruido y, por lo tanto, se filtran. La red de transición de estado celular puede definirse mediante 27 estados celulares como nodos y los vínculos muestran las posibles rutas de transición (Fig. 36E). La dirección de la progresión del ciclo celular se caracteriza fácilmente por al menos tres etapas de transición con direcciones de transición irreversibles a lo largo del ciclo celular (Fig. 36E). En la fase G1 tardía y en la fase G2/M tardía, también encontramos varios estados que muestran dinámicas de transiciones reversibles, que potencialmente reflejan dos puntos de control del ciclo celular en las fases G1/S y G2/M (33). Como se esperaba, las células con un ciclo celular similar pero con diferentes estados de respuesta de GR mostraron dinámicas de transición dramáticamente diferentes, y las células con un estado de respuesta de GR alto tienden a detenerse en la fase G1 o G2/M.

Como verificación de coherencia para validar si la red de transición del estado celular captura la dinámica de transición del estado celular, evaluamos si las probabilidades de transición pueden recuperar las distribuciones reales del estado celular en diferentes puntos temporales. De hecho, aunque las proporciones del estado celular cambian dinámicamente a lo largo de 10 horas (Fig. 36F), la red de transición de estado predice con precisión las 27 relaciones del estado celular en los cinco puntos temporales posteriores a partir de la proporción del estado celular en las células tratadas con DEX de 0 horas (Fig. 36G, Fig. 38A). También calculamos la red de transición del estado celular con solo parte de los datos (0 horas a 6 horas), lo que proporcionó probabilidades de transición altamente correlacionadas con los datos completos y predijo con precisión los estados celulares a las 10 horas (Fig. 36H, Fig. 38B).

Factores característicos que regulan las direcciones de transición del estado celular

Para caracterizar los factores que regulan la probabilidad de transición del estado celular, primero calculamos la distancia del estado celular, mediante la distancia de Pearson del transcriptoma agregado (completo y recién sintetizado) entre cada par de estados. Como se esperaba, la probabilidad de transición del estado celular se correlaciona negativamente con la distancia de transición (coeficiente de correlación de Spearman = $-0,38$, Fig. 39A). También calculamos la inestabilidad del estado, definida por la proporción de células que salen del estado en dos horas (Fig. 39B). El panorama de inestabilidad de estado coincide bien con las direcciones de transición celular (Fig. 39B): los estados sin respuesta de GR muestran una mayor inestabilidad en comparación con los estados de alta respuesta de GR. En estados de respuesta GR alta, las células en la fase G1 temprana tienen la inestabilidad más baja, mientras que las células en los estados intermedios G1/S mostraron un pico alto de inestabilidad, coherente con la detención de la fase G1 en el tratamiento tardío con DEX.

Los cambios en la proporción del estado celular después de 10 horas se correlacionan bien con la inestabilidad del estado celular (coeficiente de correlación de Spearman = $-0,88$, Fig. 39C), lo que sugiere que las dinámicas del estado celular están reguladas por el panorama de inestabilidad del estado celular. La inestabilidad del estado también se correlaciona bien con la entropía de probabilidad de transición de estado, que refleja la diversidad de las dianas de transición de estado (correlación de Pearson $r = 0,73$, Fig. 39D). Para validar si la probabilidad de transición entre estados puede ser inferida por la inestabilidad del estado cercano, ajustamos la inestabilidad y la distancia del estado cercano en un modelo de red neuronal, para predecir la probabilidad de transición de estado de cada estado a los otros estados. La combinación tanto de la inestabilidad de los estados cercanos como las distancias logró un rendimiento diez veces mayor en la predicción de la probabilidad de transición entre estados, en comparación con el uso solo de distancias de estados (la mediana del r cuadrado validado en cruz es $0,58$ al usar ambas informaciones frente a $0,046$ al usar solo la distancia de estados, $\text{valor } p = 4,5e-10$, prueba de suma de rangos de Wilcoxon bilateral, Fig. 39E), lo que sugiere que las direcciones y probabilidades de transición del estado celular están reguladas por el panorama de estabilidad del estado cercano. Y las células prefieren moverse a un estado cercano más estable que a la posición más cercana.

Discusión

Aquí desarrollamos la primera estrategia para caracterizar la dinámica de transición del estado celular a nivel del transcriptoma completo. La estrategia depende de sci-fate, una novedosa técnica de RNA-seq unicelular de alto rendimiento basada en indexación combinatoria, capaz de perfilar el transcriptoma tanto completo como recién sintetizado en miles de células. Al igual que otras técnicas de "sci-", sci-fate se escala fácilmente a millones de células (39) y es potencialmente compatible con la elaboración de perfiles tanto del transcriptoma como del epigenoma (40). Esto permite a sci-fate caracterizar la dinámica del estado celular en un sistema muy complejo (es decir, el desarrollo completo del embrión) donde aún se desconoce la ruta real de transición celular a cientos de tipos de células. Además,

desarrollamos un proceso de cálculo para estimar la velocidad de captura de ARN recién sintetizado y la velocidad de degradación génica a partir de datos de sci-fate (corrección de memoria), e inferir miles de trayectorias diferenciales para cada célula individual, vinculadas por el estado compartido del transcriptoma pasado y actual en cada punto temporal.

- 5 Para validar las técnicas y examinar cómo la dinámica del estado celular está regulada por factores internos y externos, aplicamos la estrategia a un sistema modelo de respuesta de cortisol, en el que el destino celular estaba regulado dinámicamente por el ciclo celular interno y la activación de GR inducida por fármacos extrínsecos. Mostramos que el transcriptoma recién sintetizado se vincula directamente con la respuesta del epigenoma a los estímulos ambientales, y el análisis conjunto tanto del transcriptoma completo como del recién sintetizado permite una mayor resolución en la separación del estado celular. Mediante la covarianza entre la expresión de TF y la velocidad de síntesis de ARN nuevo en miles de células, identificamos hasta mil vínculos entre TF y genes regulados, validados por datos de unión de ADN. Además, identificamos 27 "estados celulares ocultos" caracterizados por el estado combinatorio de los módulos TF funcionales en la progresión del ciclo celular y la respuesta de GR, en comparación con solo 6 estados según el análisis de agrupamiento convencional.
- 10
- 15 Mediante la corrección de la memoria y el análisis de vinculación celular, construimos más de 6000 trayectorias de transición de una sola célula que abarcaron 10 horas, siendo las trayectorias principales coherentes con la dinámica conocida del estado celular en el ciclo celular y la respuesta GR. La red de transición de estado celular se caracteriza por la probabilidad de transición en todos los estados celulares, validada por la recuperación de la dinámica de 27 estados celulares en los cinco puntos temporales. Finalmente, encontramos que las probabilidades de transición del estado celular están reguladas por dos características clave de la red de transición del estado celular: la distancia entre estados y el panorama de inestabilidad del estado, los cuales pueden estimarse potencialmente mediante técnicas convencionales de RNA-seq unicelular.
- 20

- Si bien es poderosa, esta estrategia tiene varias limitaciones. En primer lugar, para construir fielmente la trayectoria de una sola célula, necesitamos una caracterización completa del estado celular en cada punto temporal. También se necesitan múltiples observaciones para cada estado para estimar de manera sólida la probabilidad de transición. Estas limitaciones pueden resolverse fácilmente mediante la estrategia combinatoria de sci-fate, que es capaz de perfilar millones de células en un solo experimento. Otra advertencia es que la mayoría de los experimentos de etiquetado de S4U se aplican a sistemas *in vitro*. Sin embargo, una investigación reciente ha demostrado que S4U puede marcar de forma estable la transcripción de ARN específica de un tipo de célula en múltiples tejidos de ratón (es decir, cerebro, intestino y tejido adiposo) (41, 42), lo que sugiere que sci-fate, con optimizaciones adicionales para mejorar la incorporación de S4U y la velocidad de detección, se puede aplicar para perfilar la dinámica del transcriptoma unicelular *in vivo*.
- 25
- 30

- sci-fate abre una nueva vía para aplicar técnicas genómicas unicelulares "estáticas" para caracterizar sistemas dinámicos. En comparación con las técnicas tradicionales basadas en imágenes, Sci-Fate perfila la dinámica del estado celular a nivel del transcriptoma completo y permite una caracterización integral del estado celular sin selección de marcadores y el descubrimiento de la fuerza impulsora clave en la diferenciación celular. Finalmente, anticipamos que Sci-Fate se puede combinar fácilmente con técnicas alternativas de rastreo de linaje (43-45), para decodificar la dinámica detallada de transición del estado celular a cada estado celular final dentro de cientos de linajes de desarrollo.
- 35

Materiales y métodos:

- 40 Cultivo de células de mamíferos

- Todas las células de mamíferos se cultivaron a 37 °C con 5 % de CO₂, y se mantuvieron en DMEM con alto contenido de glucosa (Gibco núm. de catálogo 11965) para células HEK293T y NIH/3T3 o medio DMEM/F12 para células A549, ambos suplementados con 10 % de FBS y 1X Pen/Strep (Gibco núm. de catálogo 15140122; 100 U/ml de penicilina, 100 µg/ml de estreptomycin). Las células se tripsinizaron con tripsina-EDTA al 0,25 % (Gibco núm. de catálogo 25200-056) y se dividieron 1:10 tres veces por semana.
- 45

Procesamiento de muestras para sci-fate

- Se trataron células A549 con DEX 100 nM durante 0 h, 2 h, 4 h, 6 h, 8 h y 10 h. Las células en todas las condiciones de tratamiento se incubaron con S4U 200 µM durante las últimas dos horas antes de la recolección celular. Para las células HEK293T y NIH/3T3, las células se incubaron con S4U 200 µM durante 6 horas antes de la recolección celular.
- 50 Todas las líneas celulares (células A549, HEK293T y NIH/3T3) se tripsinizaron, se centrifugaron a 300 xg durante 5 min (4 °C) y se lavaron una vez en PBS 1X enfriado con hielo. Todas las células se fijaron con 4 ml de paraformaldehído al 4 % (EMS) enfriado con hielo durante 15 minutos en hielo. Después de la fijación, las células se granularon a 500 xg durante 3 min (4 °C) y se lavaron una vez con 1 ml de PBSR (1 x PBS, pH 7,4, 1 % de BSA, 1 % de SuperRnasalN, 1 % de DTT 10 mM). Después del lavado, las células se resuspendieron en PBSR a 10 millones de células por ml, se congelaron instantáneamente y se almacenaron en nitrógeno líquido. Las células fijadas con paraformaldehído se descongelaron en un baño de agua a 37 grados, se centrifugaron a 500 xg durante 5 minutos y se incubaron con 500 µl de PBSR que incluía Triton X-100 al 0,2 % durante 3 minutos en hielo. Las células se granularon y se resuspendieron
- 55

en 500 µl de agua libre de nucleasas, incluida SuperARNasIn al 1 %. Se añadieron 3 ml de HCl 0,1 N a las células durante 5 minutos de incubación en hielo (21). Se añadieron a las células 3,5 ml de Tris-HCl (pH = 8,0) y 35 µl de Triton X-100 al 10 % para neutralizar el HCl. Las células se granularon y se lavaron con 1 ml de PBSR. Las células se resuspendieron en 100 µl de PBSR. Se incubaron 100 µl de PBSR con células fijadas con una mezcla que incluía 40 µl de yodoacetamida (IAA, 100 mM), 40 µl de tampón fosfato sódico (500 mM, pH = 8,0), 200 µl de DMSO y 20 µl de H₂O, a 50 °C durante 15 minutos. La reacción se desactivó mediante 8 µl de DTT (1 M) y 8,5 ml de PBS (47). Las células se granularon y se resuspendieron en 100 µl de PBSI (1 x PBS, pH 7,4, 1 % de BSA, 1 % de SuperRNasIn). Para todos los lavados posteriores, los núcleos se granularon mediante centrifugación a 500 xg durante 5 minutos (4 °C).

Las siguientes etapas son similares al protocolo sci-RNA-seq con núcleos fijados con paraformaldehído (15, 16). Brevemente, las células se distribuyeron en cuatro placas de 96 pocillos. Para cada pocillo, se mezclaron 5.000 núcleos (2 µl) con 1 µl de cebador oligo-dT anclado 25 µM (5'-ACGACGCTCTTCCGATCTNNNNNNNN[índice de 10 pb]TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN-3') (SEQ ID NO:5), donde "N" es cualquier base y "V" es "A", "C" o "G"; IDT) y 0,25 µl de mezcla de dNTP 10 mM (Thermo), se desnaturalizó a 55 °C durante 5 min y se colocó inmediatamente en hielo. Luego se añadió a cada pocillo 1,75 µl de mezcla de reacción de primera hebra, que contiene 1 µl de tampón de primera hebra Superscript IV 5X (Invitrogen), 0,25 µl de DTT 100 mM (Invitrogen), 0,25 µl de transcriptasa inversa SuperScript IV (200 U/µl, Invitrogen), 0,25 µl de RNasaOUT inhibidor de ribonucleasa recombinante (Invitrogen). La transcripción inversa se llevó a cabo incubando placas al siguiente gradiente de temperatura: 4 °C 2 minutos, 10 °C 2 minutos, 20 °C 2 minutos, 30 °C 2 minutos, 40 °C 2 minutos, 50 °C 2 minutos y 55 °C 10 minutos. Luego se agruparon todas las células (o núcleos), se tiñeron con 4',6-diamidino-2-fenilindol (DAPI, Invitrogen) a una concentración final de 3 µM y se clasificaron en 25 núcleos por pocillo en 5 µl de tampón EB. Las células se clasificaron en base a la tinción DAPI de modo que los singletes se discriminaran de los dobletes y se clasificaran en cada pocillo. Luego se añadieron a cada pocillo 0,66 µl de tampón de síntesis de la segunda hebra (NEB) de ARNm y 0,34 µl de enzima de síntesis de la segunda hebra (NEB) de ARNm, y la síntesis de la segunda hebra se llevó a cabo a 16 °C durante 180 minutos. Luego cada pocillo se mezcló con 5 µl de tampón Nextera TD (Illumina) y 1 µl de enzima TDE1 solo i7 (25 nM, Illumina, diluida en tampón Nextera TD) y luego se incubó a 55 °C durante 5 minutos para llevar a cabo la tagmentación. La reacción se detuvo añadiendo 10 µl de tampón de unión a ADN (Zymo) e incubando a temperatura ambiente durante 5 minutos. Luego, cada pocillo se purificó usando 30 µl de perlas AMPure XP (Beckman Coulter), se eluyó en 16 µl de tampón EB (Qiagen) y luego se transfirió a una placa multipocillo nueva.

Para las reacciones de PCR, cada pocillo se mezcló con 2 µl de cebador P5 10 µM (5'-AATGATACGGCGACCACCGAGATCTACAC[i5]ACACTCTTTCCCTACACGACGCTCTTCCGATCT-3; IDT) (SEQ ID NO:6), 2 µl de cebador P7 10 µM (5'-CAAGCAGAAGACGGCATACGAGAT[i7]GTCTCGTGGGCTCGG-3; IDT) (SEQ ID NO:7) y 20 µl de mezcla maestra de PCR 2X de alta fidelidad NEBNext (NEB). La amplificación se llevó a cabo usando el siguiente programa: 72 °C durante 5 min, 98 °C durante 30 s, 18-22 ciclos de (98 °C durante 10 s, 66 °C durante 30 s, 72 °C durante 1 min) y un final de 72 °C durante 5 min. Después de la PCR, las muestras se combinaron y purificaron usando 0,8 volúmenes de perlas AMPure XP. Las concentraciones de las bibliotecas se determinaron mediante Qubit (Invitrogen) y las bibliotecas se visualizaron mediante electroforesis en un gel TBE-PAGE al 6 %. Las bibliotecas se secuenciaron en la plataforma NextSeq 500 (Illumina) utilizando un kit de 150 ciclos V2 (Lectura 1: 18 ciclos, Lectura 2: 130 ciclos, Índice 1: 10 ciclos, Índice 2: 10 ciclos).

Leer alineaciones y procesamiento posterior

La generación de la matriz de alineación de lectura y recuento de genes para RNA-seq unicelular se realizó usando el proceso que desarrollamos para sci-RNA-seq (48) con modificaciones menores. Las lecturas se asignaron por primera vez a un genoma de referencia con STAR/v2.5.2b (49), con anotaciones genéticas de GENCODE V19 para humanos y GENCODE VM11 para ratón. Para experimentos con células HEK293T y NIH/3T3, usamos un índice que combina cromosomas tanto humanos (hg19) como de ratón (mm10). Para el experimento A549, usamos la construcción del genoma humano hg19.

Los archivos sam unicelulares se convirtieron primero en un archivo tsv de alineación usando la función sam2tsv en jvarkit (50). A continuación, para cada archivo de alineación de células individuales, se filtraron las mutaciones que coincidían con los SNP de fondo. Para la referencia de SNP de fondo de las células A549, descargamos los datos de RNA-seq en masa de extremos emparejados para las células A549 de ENCODE (28) (nombre de muestra: ENCFF542FVG, ENCFF538ZTA, ENCFF214JEZ, ENCFF629LOL, ENCFF149CJD, ENCFF006WNO, ENCFF828WTU, ENCFF380VGD). Cada archivo fastq de extremo emparejado se recortó primero con un adaptador usando trim_galore/0.4.1(51) con la configuración predeterminada, alineada con la construcción del genoma humano hg19 con STAR/v2.5.2b (49). Samtools/v1.3 (52) eliminó las lecturas no mapeadas y múltiples mapeadas. Las lecturas duplicadas fueron filtradas por la función MarkDuplicates en picard/1.105(53). Las lecturas no duplicadas de todas las muestras se combinaron y ordenaron con samtools/v1.3 (52). Los SNP de fondo fueron llamados mediante la función mpileup en samtools/v1.3 (52) y función mipilup2snp en VarScan/2.3.9(54). Para el experimento de prueba de HEK293T y NIH/3T3, la referencia de SNP de fondo se generó en un proceso similar anterior, con los datos sam de células individuales agregados de la condición de control (sin etiquetado S4U y sin condición de tratamiento IAA).

Para cada archivo de alineación de células individuales, se eliminaron todas las mutaciones con un puntaje de calidad ≤ 13 . Las mutaciones en ambos extremos de cada lectura se debieron principalmente a errores de secuenciación y, por lo tanto, también se filtraron. Para cada lectura, verificamos si hay mutaciones T > C (para la hebra sentido) o mutaciones A > G (para la hebra antisentido) y etiquetamos estas lecturas mutadas como lecturas recién sintetizadas.

- 5 Cada célula se caracterizó por dos matrices de expresión génica digitales a partir de los datos de secuenciación completos y los datos de ARN recién sintetizado como se describe anteriormente. Se filtraron los genes con expresión en igual o menos de 5 células. Se descartaron las células con menos de 2.000 UMI o más de 80.000 UMI. Las células con puntuación de doblete $> 0,2$ mediante el proceso de análisis de doblete Scrublet/0,2(55) fueron eliminadas.

- 10 La dimensionalidad de los datos se redujo primero con PCA (después de seleccionar los 2000 genes principales con la mayor variación) en matrices de expresión génica digitales, ya sea en datos de expresión génica completos o en los datos de expresión génica recién sintetizados por Monocle 3 (56, 57). Las 10 mejores PC fueron seleccionadas para el análisis de reducción de dimensiones con aproximación y proyección múltiple uniforme (UMAP/0.3.2), un algoritmo propuesto recientemente basado en geometría de Riemann y topología algebraica para realizar reducción de dimensiones y visualización de datos (26). Para el análisis conjunto, combinamos las 10 PC principales calculadas en todo el transcriptoma y las 10 PC principales en el transcriptoma recién sintetizado para cada célula individual antes de la reducción de dimensiones con UMAP. Los grupos de células se realizaron mediante el algoritmo de densidadPeak implementado en Monocle 3 (56, 57). Primero realizamos un análisis UMAP de la información conjunta de todas las células procesadas e identificamos un grupo atípico (724 de 7.404 células). Estas células estaban marcadas por un alto nivel de expresión de GATA3, un marcador de células diferenciadas (34), y se filtraron antes del análisis posterior.

Análisis para vincular el factor de transcripción (TF) con genes regulados

- Nuestro objetivo era identificar vínculos entre los TF y los genes regulados en función de su covarianza. Se seleccionaron células con más de 10.000 UMI detectadas y genes con lecturas de síntesis recientes detectadas en más del 10 % de todas las células. La expresión génica completa y el recuento de genes recién sintetizados por célula se normalizaron mediante factores de tamaño de biblioteca específicos de cada célula calculados en la matriz de expresión génica completa mediante estimaSizeFactors en Monocle 3 (56, 57), se transformaron logarítmicamente, se centraron y luego se escalaron mediante la función escalar() en R. Para cada gen detectado, se construyó un modelo de regresión LASSO con el paquete glmnet (58) para predecir los niveles de expresión normalizados, basándose en la expresión normalizada de 853 TF anotados en los datos "motifAnnotations_hgnc" del paquete RcisTarget(29), ajustando el siguiente modelo:

$$G_i = \beta_0 + \beta_t T_i$$

- donde G_i es el valor de expresión génica ajustado para el gen i . Se calcula mediante el recuento de ARNm recién sintetizado para cada célula, se normaliza por la estimación del factor de tamaño específico de la célula (SG_i) por los FactoresTamañoEstimados en Monocle 3 (56, 57) en la matriz de expresión completa de cada célula, y se transforma logarítmicamente:

$$G_i = \ln\left(\frac{g_i}{SG_i} + 0,1\right)$$

Para simplificar la comparación posterior entre genes, estandarizamos la respuesta G_i antes de ajustar el modelo para cada gen i con la función escalar() en R.

- Similar con G_i , t_i es el valor de expresión de TF ajustado para cada célula. Se calcula mediante el recuento completo de expresión de TF para cada célula, se normaliza por la estimación del factor de tamaño específico de la célula (SG_i) por FactoresTamaño en Monocle 3 (56, 57) en la matriz de expresión completa de cada celda, y log transformado:

$$T_i = \ln\left(\frac{t_i}{SG_i} + 0,1\right)$$

Antes de la instalación, t_i están estandarizados con la función escalar() en R.

- Nuestro enfoque apunta a los TF que pueden regular cada gen, encontrando el subconjunto que puede usarse para predecir su expresión en un modelo de regresión. Sin embargo, un TF con expresión correlacionada con la expresión de un gen no garantiza que esté regulando ese gen: si el gen A se expresa específicamente en el estado celular 1 y el TF B se expresa específicamente en el tipo de célula 2. Aunque existen correlaciones negativas entre la expresión de un TF y una nueva velocidad de síntesis del gen podría reflejar la actividad de un represor transcripcional, sentimos que la explicación más probable para los vínculos negativos presentados por glmnet eran patrones mutuamente excluyentes de expresión específica del estado celular y actividad TF. Por lo tanto, durante la predicción, excluimos los TF con expresión correlacionada negativamente con la velocidad de síntesis del gen y también con uniones de coeficiente de correlación bajo ($\leq 0,03$). Identificamos un total de 6.103 vínculos entre TF y genes regulados.

Para identificar supuestas dianas de enlace directo, cruzamos los vínculos con los TF perfilados en el experimento ENCODE Chip-seq (28). De 1.086 vínculos con TF caracterizados en ENCODE, 807 vínculos fueron validados por sitios de unión de TF cerca de promotores génicos (59), un enriquecimiento de 4,3 veces en la relación impar (número de vínculos validados sobre vínculos no validados) en comparación con el fondo (relación impar = 2,89 en los vínculos identificados en la regresión LASSO frente a 0,67 en el fondo, valor de $p < 2,2e-16$, de Prueba exacta de Fisher). Solo se conservan los conjuntos de genes con un enriquecimiento significativo de los sitios de unión correctos de TF Chip-seq (prueba exacta de Fish, tasa de descubrimiento falso del 5 %) y se podan para eliminar genes diana indirectos sin soporte de datos de unión de TF. En este enfoque se retuvieron 591 vínculos.

Para ampliar los vínculos validados del gen TF, aplicamos adicionalmente el paquete SCENIC (29), un proceso para construir redes reguladoras de genes basadas en el enriquecimiento de motivos TF diana alrededor de los promotores de genes (10 kb). Cada módulo de coexpresión identificado mediante regresión LASSO se analizó mediante un análisis de motivo regulador cis usando RcisTarget (29). Solo se retuvieron los módulos con un enriquecimiento significativo de motivos del regulador TF correcto y se podaron para eliminar genes diana indirectos sin soporte de motivos. Filtramos los vínculos del gen TF mediante tres umbrales de coeficientes de correlación (0,3, 0,4 y 0,5) y combinamos todos los vínculos validados por RcisTarget (29). En total, hubo 509 vínculos validados mediante un enfoque de análisis de motivos. Combinando ambos enfoques, identificamos un total de 986 vínculos reguladores del gen TF mediante la covarianza entre la expresión de TF y la velocidad de síntesis génica, validados mediante datos de unión de ADN o análisis de motivos. Para evaluar la posibilidad de que los enlaces fueran artefactos de regresión regularizada, permutamos los ID de muestra de la matriz de expresión de TF y realizamos el mismo análisis. No se identificaron vínculos después de esta permutación.

Ordenar células por módulos TF funcionales

Para calcular la actividad de TF en cada célula, los recuentos de UMI recién sintetizados para genes dentro del módulo de TF diana se escalan según el tamaño de la biblioteca, se transformaron logarítmicamente, se agregaron y luego se asignaron a puntuaciones Z. Como los TF con actividad altamente correlacionada o anticorrelacionada sugieren que pueden funcionar en procesos biológicos vinculados, calculamos el coeficiente de correlación absoluta de Pearson entre cada par de actividad de TF y, en base a esto, agrupamos los TF mediante el método de agrupación ward.d2 en el paquete pheatmap/1.0.12(60). Se identificaron y anotaron cinco módulos TF funcionales en función de sus funciones.

Para caracterizar los estados celulares en la dimensión de cada módulo funcional de TF, las células se ordenaron por la actividad de los TF relacionados con el ciclo celular (módulo de TF 1) o los TF relacionados con la respuesta de GR (módulo de TF 3) con UMAP (métrica = "coseno", $n_vecinos = 30$, $dist_min = 0,01$). La trayectoria de progresión del ciclo celular fue validada por marcadores génicos del ciclo celular en Seurat/2.3.4(27). Se identificaron tres fases del ciclo celular mediante el algoritmo de Picosdensidad implementado en Monocle 3 (56, 57), sobre las coordenadas UMAP ordenadas por módulos TF del ciclo celular. Como cada fase principal del ciclo celular todavía mostraba actividad TF variable y expresión de marcadores del ciclo celular, segmentamos cada fase en estados temprano/medio/tardío mediante agrupación de k-medias ($k = 3$) y recuperamos un total de nueve estados del ciclo celular. Se identificaron tres estados de respuesta de GR mediante el algoritmo de Picodensidad implementado en Monocle 3 (56, 57).

Recuperación del estado del transcriptoma pasado del sci-fate

Para identificar el estado pasado del transcriptoma (el estado celular antes del etiquetado con S4U), asumimos que la vida media del ARNm es estable en diferentes condiciones de tratamiento con DEX. Esta suposición se valida aún más mediante una verificación de autoconsistencia posterior. Bajo esta suposición, el transcriptoma en masa parcialmente degradado antes del etiquetado S4U de 2 horas debe ser el mismo entre células sin DEX y células tratadas con DEX de 2 horas. Por lo tanto, sus diferencias en el transcriptoma completo (en masa) deberían ser iguales a sus diferencias en el transcriptoma recién sintetizado (en masa) corregido por la velocidad de detección de la técnica:

$$A_{0h} / S_{0h} - (N_{0h} / S_{0h}) / \alpha = A_{2h} / S_{2h} - (N_{2h} / S_{2h}) / \alpha$$

A_{0h} es el recuento UMI agregado para todas las células en ningún grupo de tratamiento con DEX; S_{0h} es el tamaño de la biblioteca (recuento total de células UMI) sin tratamiento con DEX; N_{0h} es el recuento UMI recién sintetizado agregado para todas las células en el grupo sin tratamiento con DEX; A_{2h} es el recuento UMI agregado para todas las células en el grupo de tratamiento DEX de 2 horas; S_{2h} es el tamaño de la biblioteca (recuento total de células UMI) en el grupo de tratamiento con DEX de 2 horas; N_{2h} es el recuento agregado de UMI recién sintetizado para todas las células en el grupo de tratamiento con DEX de 2 horas; α es la velocidad de detección de sci-fate. En teoría, se puede calcular una velocidad de detección para cada gen. Sin embargo, para genes con diferencias menores en la velocidad de nueva síntesis entre dos condiciones, el valor estimado α está dominado por el ruido. Por lo tanto, seleccionamos genes que mostraban mayores diferencias en la velocidad de nueva síntesis normalizada entre dos condiciones: primero probamos una serie de umbrales para el filtrado de genes y calculamos el α para cada gen. Luego trazamos la relación entre el umbral y la relación de genes con valores α fuera de intervalo. (< 0 o > 1). Seleccionamos el umbral que estaba en el punto culminante de la parcela con 186 genes seleccionados. Las diferencias en el ARNm recién

sinetizado de estos genes se correlacionan altamente con las diferencias en el nivel de expresión del ARNm (r de Pearson = 0,93, Fig. 35A), lo que sugiere que la tasa de detección de ARN nuevo es bastante estable en todos los genes. Hay una media del 82 % del ARN recién sintetizado capturado por sci-fate.

- 5 A continuación, calculamos la velocidad de degradación del ARNm cada 2 horas. Como la población de células A549 puede considerarse estable sin perturbaciones externas, para las células tratadas con DEX de 2 horas, su estado anterior (antes del etiquetado con S4U de 2 horas) debe ser el mismo que el de las células tratadas con DEX de 0 horas. De manera similar, el estado anterior (antes del etiquetado S4U) para las células tratadas con DEX $T = 0/2/4/6/8/10$ horas debe ser similar a las células perfiladas $T = 0/0/2/4/6/8$ horas:

$$A_{t1} / S_{t1} - (N_{t1} / S_{t1}) / \alpha = A_{t0} / S_{t0} * \beta$$

- 10 A_{t1} es el recuento UMI agregado para todas las células en $t1$; S_{t1} es el tamaño de la biblioteca (el recuento total de células UMI) en $t1$; N_{t1} es el recuento UMI recién sintetizado agregado para todas las células en $t1$; α es la velocidad de detección estimada de sci-fate; A_{t0} es el recuento UMI agregado para todas las celdas en $t0$; S_{t0} es el tamaño de la biblioteca (el recuento total de células del UMI) en $t0$; β es $1 -$ velocidad de degradación específica del gen entre $t0$ y $t1$, y está relacionada con la vida media del ARNm y por:

15
$$\beta = 1 - (1 / 2)^{(t1 - t0) / \gamma}$$

La velocidad de degradación génica β se puede calcular en cada intervalo de 2 horas de tratamiento con DEX. Como verificación de autocoherencia mencionada anteriormente, las velocidades de degradación génica están altamente correlacionadas entre diferentes tiempos de tratamiento con DEX (Fig. 35B). Luego utilizamos la velocidad promedio de degradación génica para el análisis posterior.

- 20 Con la velocidad de detección y la velocidad de degradación génica estimadas, el estado pasado del transcriptoma de cada célula se puede estimar mediante:

$$a_{t1} - n_{t1} / \alpha = a_{t0} * \beta$$

- a_{t1} es el recuento de UMI unicelular en $t1$; n_{t1} es el recuento de UMI de células individuales recién sintetizadas en $t1$; α es la velocidad de detección estimada de sci-fate; β es $1 -$ velocidad de degradación específica del gen entre $t0$ y $t1$.
25 a_{t0} es el recuento estimado de UMI unicelular en un punto de tiempo pasado $t0$, con todos los valores negativos convertidos a 0.

Análisis de vinculación para construir una trayectoria de estado unicelular

- Mediante el análisis de vinculación, pretendemos identificar células madres e hijas vinculadas en la misma trayectoria celular. Técnicamente, para las células en $t1$, combinamos su estado de transcriptoma pasado (antes del etiquetado S4U, 2 horas antes de $t1$ en nuestro experimento) como un grupo 1, y el estado de transcriptoma completo de $t0$ (2 horas antes de $t1$) como otro grupo 2. Suponiendo que no haya apoptosis celular evidente, estos dos grupos deberían tener una distribución del estado celular similar. Aplicamos una estrategia de alineación múltiple para identificar estados celulares comunes entre dos conjuntos de datos, basados en fuentes comunes de variación (27). Este análisis se basa en otra suposición de que el estado pasado y actual de cada célula (excepto las células en los puntos de tiempo inicial y final) se detectan de manera integral, lo que es válido en nuestros conjuntos de datos, ya que se perfilan más de 6000 células (más de 1.000 celdas por condición) o una célula durante menos de un minuto durante el ciclo celular. Como resultado del proceso, los estados celulares de $t0$ y los estados de celda anteriores de $t1$ están alineados en el mismo espacio UMAP. La violación de los supuestos anteriores puede ser detectada por valores atípicos durante la alineación de los dos conjuntos de datos. Para cada célula A en $t1$, seleccionamos a su vecino más cercano en $t0$ como su estado parental en el espacio UMAP de alineación. De manera similar, para cada célula en $t0$, seleccionamos a su vecino más cercano en $t1$ como su estado celular hijo. Es de destacar que no es necesario que el vínculo sea bidireccional: el estado parental de una célula puede estar vinculado a una célula hija diferente. Como se identificaron el estado parental y el estado hijo para cada célula (excepto las células en las 0 horas y las 10 horas), luego identificamos la célula parental vinculada de cada célula principal y, de manera similar, la célula hija vinculada del hijo de cada célula. Por lo tanto, cada célula se puede caracterizar por una ruta de transición de estado de celda única a través de los cinco puntos de tiempo que abarcan 10 horas. Como se perfilan múltiples celdas (>50) en cada estado celular, también se puede capturar el proceso de transición de estado celular estocástico.

Análisis de agrupamiento y reducción de dimensiones para la dinámica del transcriptoma unicelular

- Para la reducción de dimensiones en la dinámica del transcriptoma unicelular, se seleccionaron para cada estado las 5 PC principales para el transcriptoma completo y las 5 PC principales para el transcriptoma recién sintetizado, y se combinaron en orden temporal a lo largo de la trayectoria del estado de una sola célula para el análisis UMAP. Los principales tipos de trayectorias celulares se identificaron mediante el algoritmo de agrupación de picos de densidad (61).

Con la proporción del estado celular en el punto temporal inicial (tratamiento de 0 horas) y las probabilidades de transición del estado celular estimadas a partir de los datos, primero predijimos la distribución del estado celular después de 2 horas, suponiendo que el proceso de transición del estado celular en el tratamiento con DEX es una dinámica markoviana, independiente del tiempo, autónoma para la célula. De manera similar, la distribución del estado de las células en un punto temporal posterior se puede calcular en función de la distribución del estado de las células predicha 2 horas antes.

Predicción de la probabilidad de transición entre estados por inestabilidad del estado

La inestabilidad del estado celular se define como la probabilidad de que cada estado pase a otros estados después de 2 horas. Para calcular la distancia del estado celular, primero tomamos muestras de un número igual ($n = 50$) de células en cada estado y agregamos el transcriptoma completo y el transcriptoma recién sintetizado de todas las células dentro del estado. Cada estado celular puede definirse mediante la información conjunta que combina el transcriptoma completo y el recién sintetizado. La distancia del estado celular se calcula como el coeficiente de correlación de Pearson de la información conjunta entre dos estados.

Para predecir la probabilidad de transición entre estados, construimos una red neuronal de 3 capas (número de unidades: 128, 128, 26 con activación relu en cada capa; función de pérdida: coseno_proximidad, tamaño de lote: 128, épocas: 80) con Keras/2.2.4(62). Como entrada, usamos la inestabilidad del estado actual, la inestabilidad del estado normalizado de los otros 26 estados (escalada por la inestabilidad del estado actual) y la distancia de transición (al cuadrado) del estado actual a los otros 26 estados (en el mismo orden de estados en vector de inestabilidad del estado). Para evitar un ajuste excesivo, permutamos los órdenes de los estados en la inestabilidad del estado 200 veces para cada entrada, manteniendo al mismo tiempo el orden de los estados de la distancia de transición del estado igual con la inestabilidad del estado. Para evaluar el rendimiento del modelo, aplicamos la validación de dejar uno afuera entrenando el modelo en 26 estados y validamos el modelo en el estado izquierdo al predecir las probabilidades de transición del estado a los otros 26 estados. Para predecir la probabilidad entre estados solo con la distancia de transición de estado, se usa el mismo modelo para el entrenamiento y la validación con todas las inestabilidades de estado de entrada reemplazadas por 1.

Referencias

1. N. Moris, C. Pina, A. M. Arias, Transition states and cell fate decisions in epigenetic landscapes. *Nat. Rev. Genet.* 17, 693-703 (2016).
2. A. Filipczyk et al., Network plasticity of pluripotency transcription factors in embryonic stem cells. *Nat. Cell Biol.* 17, 1235-1246 (2015).
3. S. Hormoz et al., Inferring Cell-State Transition Dynamics from Lineage Trees and Endpoint Single-Cell Measurements. *Cell Syst.* 3, 419-433.e8 (2016).
4. V. A. Herzog et al., Thiol-linked alkylation of RNA to assess expression dynamics. *Nat. Methods.* 14, 1198-1204 (2017).
5. J. A. Schofield, E. E. Duffy, L. Kiefer, M. C. Sullivan, M. D. Simon, TimeLapse-seq: adding a temporal dimension to RNA sequencing through nucleoside recording. *Nat. Methods.* 15, 221-225 (2018).
6. J. C. Buckingham, Glucocorticoids: exemplars of multi-tasking. *Br. J. Pharmacol.* 147, S258 (2006).
7. M. D. Cleary, C. D. Meiering, E. Jan, R. Guymon, J. C. Boothroyd, Biosynthetic labeling of RNA with uracil phosphoribosyltransferase allows cell-specific microarray analysis of mRNA synthesis and decay. *Nat. Biotechnol.* 23, 232-237 (2005).
8. L. Dolken et al., High-resolution gene expression profiling for simultaneous kinetic parameter analysis of RNA synthesis and decay. *RNA.* 14, 1959-1972 (2008).
9. C. Miller et al., Dynamic transcriptome analysis measures rates of mRNA synthesis and decay in yeast. *Mol. Syst. Biol.* 7, 458-458 (2014).
10. E. E. Duffy et al., Tracking Distinct RNA Populations Using Efficient and Reversible Covalent Chemistry. *Mol. Cell.* 59, 858-866 (2015).
11. B. Schwalb et al., TT-seq maps the human transient transcriptome. *Science.* 352, 1225-1228 (2016).
12. M. Rabani et al., Metabolic labeling of RNA uncovers principles of RNA production and degradation dynamics in mammalian cells. *Nat. Biotechnol.* 29, 436-442 (2011).
13. M. R. Miller, K. J. Robinson, M. D. Cleary, C. Q. Doe, TU-tagging: cell type-specific RNA isolation from intact complex tissues. *Nat. Methods.* 6, 439-441 (2009).

14. D. A. Cusanovich et al., Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science*. 348, 910-914 (2015).
15. J. Cao et al., Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science*. 357, 661-667 (2017).
- 5 16. J. Cao et al., Joint profiling of chromatin accessibility and gene expression in thousand of single cells. *Science*. 361, 1380-1385 (2018).
17. V. Ramani et al., Massively multiplex single-cell Hi-C (2016), doi:10.1101/065052.
18. R. M. Mulqueen et al., Highly scalable generation of DNA methylation profiles in single cells. *Nat. Biotechnol.* 36, 428-431 (2018).
- 10 19. S. A. Vitak et al., Sequencing thousands of single-cell genomes with combinatorial indexing. *Nat. Methods*. 14, 302-308 (2017).
20. Y. Yin et al., High-throughput mapping of meiotic crossover and chromosome mis-segregation events in interspecific hybrid mice (2018), doi:10.1101/338053.
21. A. B. Rosenberg et al., Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science*. 360, 176-182 (2018).
- 15 22. T. E. Reddy et al., Genomic determination of the glucocorticoid response reveals unexpected mechanisms of gene regulation. *Genome Res.* 19, 2163-2171 (2009).
23. S. John et al., Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nat. Genet.* 43, 264-268 (2011).
- 20 24. T. E. Reddy, J. Gertz, G. E. Crawford, M. J. Garabedian, R. M. Myers, Hypersensitive Glucocorticoid Response Specifically Regulates Period 1 and Expression of Circadian Genes. *Mol. Cell. Biol.* 32, 3756-3767 (2012).
25. C. M. Vockley et al., Direct GR Binding Sites Potentiate Clusters of TF Binding across the Human Genome. *Cell*. 166, 1269-1281.e19 (2016).
26. L. McInnes, J. Healy, N. Saul, L. Großberger, UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software*. 3, 861 (2018).
27. A. Butler, P. Hoffman, P. Smibert, E. Papalexi, R. Satija, Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* 36, 411-420 (2018).
28. El Consorcio del Proyecto ENCODE, The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science*. 306, 636-640 (2004).
- 30 29. S. Aibar et al., SCENIC: single-cell regulatory network inference and clustering. *Nat. Methods*. 14, 1083-1086 (2017).
30. M. Boruk, J. G. A. Savory, R. J. G. Haché, AF-2-Dependent Potentiation of CCAAT Enhancer Binding Protein β -Mediated Transcriptional Activation by Glucocorticoid Receptor. *Mol. Endocrinol.* 12, 1749-1763 (1998).
- 35 31. W. Qin et al., Identification of functional glucocorticoid response elements in the mouse FoxO1 promoter. *Biochem. Biophys. Res. Commun.* 450, 979-983 (2014).
32. CS Sheela Rani, N. Elango, S.-S. Wang, K. Kobayashi, R. Strong, Identification of an Activator Protein-1-Like Sequence as the Glucocorticoid Response Element in the Rat Tyrosine Hydroxylase Gene. *Mol. Pharmacol.* 75, 589 (2009).
33. M. Fischer, G. A. Müller, Cell cycle transcription control:DREAM/MuvB and RB-E2F complexes. *Crit. Rev. Biochem. Mol. Biol.* 52, 638-662 (2017).
- 40 34. J. Chou, S. Provot, Z. Werb, GATA3 in development and cancer differentiation: cells GATA have it! *J. Cell. Physiol.* 222, 42-49 (2010).
35. J. Y. C. Madhurima Biswas, Role of Nrf1 in antioxidant response element-mediated gene expression and beyond. *Toxicol. Appl. Pharmacol.* 244, 16 (2010).
- 45 36. I. -G. Ryoo, M.-K. Kwak, Regulatory crosstalk between the oxidative stress-related transcription factor Nfe212/Nrf2 and mitochondria. *Toxicol. Appl. Pharmacol.* 359, 24-33 (2018).

37. J. M. Harmon, M. R. Norman, B. J. Fowlkes, E. B. Thompson, Dexamethasone induces irreversible G1 arrest and death of a human lymphoid cell line. *J. Cell. Physiol.* 98, 267-278 (1979).
38. A. K. Greenberg et al., Glucocorticoids inhibit lung cancer cell growth through both the extracellular signal-related kinase pathway and cell cycle regulators. *Am. J. Respir. Cell. Mol. Biol.* 27, 320-328 (2002).
- 5 39. J. Cao et al., Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science*. 357, 661-667 (2017).
40. J. Cao et al., Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science*. 361, 1380-1385 (2018).
- 10 41. W. Matsushima et al., SLAM-ITseq: sequencing cell type-specific transcriptomes without cell sorting. *Development*. 145 (2018), doi:10.1242/dev.164640.
42. U. Sharma et al., Small RNAs are trafficked from the epididymis to developing mammalian sperm (2017), doi:10.1101/194522.
43. A. McKenna et al., Whole-organism lineage tracing by combinatorial and cumulative genome editing. *Science*. 353, aaf7907 (2016).
- 15 44. B. Raj et al., Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain. *Nat. Biotechnol.* 36, 442-450 (2018).
45. K. L. Frieda et al., Synthetic recording and in situ readout of lineage information in single cells. *Nature*. 541, 107-111 (2017).
46. H. Wickham, *ggplot2: Elegant Graphics for data Analysis* (Springer, 2016).
- 20 47. M. Muhar et al., SLAM-seq defines direct gene-regulatory functions of the BRD4-MYC axis. *Science*. 360, 800-805 (2018).
48. J. Cao et al., Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science*. 357, 661-667 (2017).
49. A. Dobin et al., STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. 29, 15-21 (2013).
- 25 50. P. Lindenbaum, *JVarkit: java-based utilities for Bioinformatics*. figshare (2015)).
51. FelixKrueger, *FelixKrueger/TrimGalore*. GitHub, (disponible en <https://github.com/FelixKrueger/TrimGalore>).
52. H. Li et al., The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 25, 2078-2079 (2009).
53. Herramientas Picard: de Broad Institute, (disponible en <http://broadinstitute.github.io/picard/>).
54. D. C. Koboldt et al., VarScan 2: somatic mutation and copy number alteration Discovery in cancer by exome sequencing. *Genome Res*. 22, 568-576 (2012).
- 30 55. S. L. Wolock, R. López, A. M. Klein, Scrublet: computational identification of cell doublets in single-cell transcriptomic data (2018), doi:10.1101/357368.
56. X. Qiu et al., Reversed graph embedding resolves complex single-cell trajectories. *Nat. Methods*. 14, 979-982 (2017).
- 35 57. laboratorio-cole-trapnell, laboratorio-cole-trapnell/liberación-monocle. *GitHub*, (disponible en <https://github.com/cole-trapnell-lab/monocle-release>).
58. J. Friedman, T. Hastie, R. Tibshirani, Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Softw.* 33 (2010), doi:10.18637/jss.v033.i01.
59. Dataset – ENCODE Transcription Factor Binding Site Profiles, (disponible en <http://amp.pharm.mssm.edu/Harmonizome/dataset/ENCODE+Transcription+Factor+Binding+Site+Profiles>).
- 40 60. raivokolde, raivokolde/pheatmap. *GitHub*, (disponible en <https://github.com/raivokolde/pheatmap>).
61. A. Rodríguez, A. Laio, Clustering by fast search and find of density peaks. *Science*. 344, 1492-1496 (2014).
62. equipo-keras, equipo-keras/keras. *GitHub*, (disponible en <https://github.com/keras-team/keras>).

Captura de transcrito múltiple

La mayoría de los métodos de secuenciación de ARN unicelular se saturan con una cobertura de 15.000 a 50.000 lecturas únicas por célula (Ziegenhain et al. 2017), mientras que el contenido total de ARNm de células individuales puede oscilar de 50.000 a 300.000 moléculas (Marinov et al. 2014). Además, la mayoría de estos métodos usan cebado de oligo(dT) para la transcripción inversa (RT), que centra la secuenciación en el extremo 3' de los ARN. Esto significa que estos métodos tienen un poder limitado para detectar cambios en la abundancia de cualquier transcrito determinado. Estudios recientes que perfilaron una gran cantidad de células (Gasperini et al. 2019; Cao et al. 2019) han requerido una profundidad de secuenciación muy alta: las ejecuciones de Illumina NovaSeq utilizadas en estos estudios cuestan 30.000 dólares cada una, lo que coloca estos experimentos firmemente fuera del alcance de la mayoría de los grupos.

Sin embargo, en ambos casos, la cantidad de lecturas necesarias para obtener información biológica a partir de los datos es relativamente pequeña. En lecturas de células individuales de perturbaciones no codificantes, solo se analizan los genes cis del elemento regulador que está alterado para detectar cambios en la expresión (Xie et al. 2017; Gasperini et al. 2018). En los experimentos de atlas celulares, mientras que se usan patrones de expresión global para agrupar células similares, la asignación del tipo de célula se realizó usando una pequeña cantidad de genes de factores de transcripción clave. Por lo tanto, la capacidad de enfocar la lectura en los transcritos de genes que son más informativos en estos experimentos daría como resultado una gran reducción en la profundidad de secuenciación requerida y un aumento en el poder para detectar diferencias sutiles entre las células.

Nos centramos en la secuenciación unicelular de los ARNm de interés mediante el uso de cebadores RT específicos más que en el cebado oligo(dT). Recientemente se usó un método similar en masa para secuenciar específicamente todas las uniones de empalme conocidas en la levadura, lo que dio por resultado un enriquecimiento de 100 veces para las regiones diana en comparación con las que no eran diana (Xu et al., 2018). Un conjunto de cebadores RT agrupados en transcritos de interés permitirá la reducción de la lectura de una biblioteca de transcriptomas (sciRNA-seq) a cientos de transcritos capturados por experimento.

Esta manipulación de sciRNA-seq tiene múltiples ventajas sobre el cebado con oligo(dT). Primero, dirigirá la secuenciación a regiones del genoma que hemos determinado que son más informativas para cada experimento. En segundo lugar, permite que cada molécula de ARN tenga múltiples oportunidades de transcripción inversa en ADNc, lo que aumenta la probabilidad de detección por molécula de ARN. En tercer lugar, este enfoque nos permite apuntar solo a amplicones que son mapeables de manera única y podrían reducir el fondo de lecturas ribosómicas más que las alternativas de cebado aleatorio con hexámero u oligo(dT). En cuarto lugar, nos permite apuntar a regiones informativas de ARNm, como uniones de empalme y exones resultantes de eventos de sitios de inicio de transcripción alternativos, proporcionando así información de isoformas que no se detecta fácilmente con sciRNA-seq convencional.

sciRNA-seq es especialmente adecuado para la modificación con múltiples cebadores RT. La mayoría de los métodos de RNA-seq unicelular usan perlas unidas con oligos identificadores únicos para añadir códigos de barras de identificación celular al transcriptoma de cada célula, generalmente capturando ARNm mediante hibridación con su cola poli(A). Si bien dichas perlas se han modificado para añadir un puñado de cebadores RT específicos para aumentar la cobertura de algunos transcritos (Saikia et al. 2018), esta estrategia sería difícil de escalar a cientos de transcritos específicos o cambiar rápidamente entre experimentos. Por lo tanto, la adaptabilidad de la indexación combinatoria unicelular será útil en el desarrollo de RNA-seq unicelulares de RT múltiple.

El flujo de trabajo para este aspecto es similar al protocolo sciRNA-seq de tres niveles descrito en los Ejemplos 1 y 3, pero en algunas versiones no incluye la etapa RT.

1. Diseñar un grupo de cebadores RT. En un aspecto, estos se sintetizarán individualmente y se agruparán. Para apuntar a >384 amplicones, se puede sintetizar una biblioteca de cebadores, propagarla como ADN bicatenario y procesarla para producir cebadores monocatenarios como se describe (Xu et al. 2018). Esta segunda estrategia permite la adición de muchos índices únicos a los cebadores de RT (lo que permite la indexación de sciRNA-seq a RT y PCR final).

2. RT múltiple, usando el conjunto de cebadores. Esta será una reacción única con miles de celdas (si no se realiza ninguna indexación en esta etapa) o muchas reacciones paralelas que añaden un índice específico del pocillo al realizar la transcripción inversa.

3. Ligar un adaptador de horquilla para añadir un índice específico del pocillo.

4. Reunir todas las células y realizar la síntesis de la segunda hebra.

5. Distribuir las células entre muchos pocillos y realizar la tagmentación para añadir un segundo mango de PCR constante.

6. Amplificación por PCR, añadiendo un índice final específico del pocillo.

7. Secuencia.

Flujo de trabajo de diseño del cebador:

1. Recopilar la secuencia para todos los exones de los genes a los que se dirige.

2. Analizar todos los cebadores RT posibles de 25 pb.

5 3. Filtrar los cebadores RT candidatos por:

a. Contenido de GC entre 40-60 %, correspondiente a temperaturas de fusión que están aproximadamente entre 55 y 70 grados.

b. Al menos 2 G o C en los últimos 5 nt del cebador, lo que aumenta la probabilidad de que el cebador RT apareado sea un buen sustrato para la extensión mediante la enzima transcriptasa inversa.

10 c. Probabilidad de cebado fuera del objetivo. En nuestro primer experimento, descubrimos que, si bien nuestros genes diana estaban muy enriquecidos, una gran fracción de las lecturas todavía derivaban de otros ARN que abundan dentro de las células. La mayoría de estos eventos de cebado fuera del objetivo fueron el resultado de ~5-8 pb de complementariedad entre el extremo 3' del cebador y el ARN fuera del objetivo. Por lo tanto, nuestro último proceso de diseño de cebadores tiene en cuenta la abundancia del hexámero final del cebador RT dentro del ARN celular total. Solo incluimos cebadores RT donde este último hexámero está:

i. No está presente en el ARN ribosómico. Del conjunto de hexámeros 'no tan aleatorios' o NSR descritos anteriormente (Armour et al. 2009). Es mucho menos probable que los cebadores que pasen este filtro tengan un cebado fuera de objetivo dentro del ARN ribosómico.

20 ii. Poco representado dentro del ARN celular total. Contamos la abundancia de los 4.096 hexámeros posibles dentro de las lecturas de PRO-seq asignadas al genoma humano (Core et al. 2014). PRO-seq mide toda la transcripción naciente dentro de las células, incluida la transcripción ribosómica. Solo usamos cebadores RT que terminan en hexámeros que se encuentran dentro del cuartil inferior de abundancia en este conjunto de datos. Esto rescata algunos hexámeros que, aunque presentes dentro del ARN ribosómico, no abundan como ARN dentro de las células.

25 Este filtro de abundancia cambia drásticamente la elección del cebador. Solo hay ~17 % de superposición entre los cebadores elegidos por nuestro proceso con o sin este filtro. Las versiones futuras de nuestro proceso de diseño perfeccionarán este filtro fuera de objetivo. A medida que recopilemos datos para más cebadores, deberíamos poder evaluar más eventos de cebado fuera del objetivo.

4. Filtrar candidatos por capacidad de mapeado. Alineamos a cada candidato con hg19 usando bowtie, permitiendo 3 desajustes. Esta etapa asegura que cada cebador tendrá solo un sitio diana en el genoma.

30 5. De los posibles cebadores que han pasado por estos filtros, elegir el conjunto que se distribuya de manera más uniforme en todo el gen.

Para cada gen al que nos dirigimos, decidimos cuántos cebadores diseñar por exón. Incluimos el primer y el último cebador que pasa los filtros para cada exón y luego seleccionamos cebadores internos que cubren el exón de manera más uniforme minimizando la distancia desde las posiciones de los cebadores que dividirían exactamente el exón en n pedazos.

35 Por ejemplo, para un exón de 300 pb, donde estamos buscando 3 cebadores, tomamos los cebadores más cercanos a las posiciones 1, 150 y 300 que pasaron todos los filtros hasta este punto.

40 6. Para nuestro experimento piloto, los cebadores RT se ordenaron en placas de 384 pocillos y se agruparon para crear una mezcla equimolar de todos los cebadores. Luego, esta mezcla se fosforiló con polinucleótido quinasa T4, para permitir la ligadura de un oligo de horquilla indexado durante la generación de la biblioteca sciRNA-seq (Cao et al. 2019). Esto es mucho más rentable que pedir oligos fosforilados. Los cebadores RT de 25 pb también añaden un identificador molecular único (UMI) de 8 pb y un mango de 6 pb para el apareamiento de un oligo de horquilla que añadirá un índice específico del pocillo (para indexación combinatoria) y un mango de PCR.

45 Este proceso puede ser iterativo cuando cada cebador RT se solicita por separado: en experimentos posteriores se logró una relación menor fuera del objetivo al reagrupar selectivamente cebadores que se encontró que tenían velocidades de captura favorables en el primer experimento. Cada lectura de secuenciación de Illumina abarca el cebador RT de 25 pb y la molécula de ARN capturada, lo que nos permite mapear los cebadores RT y las moléculas capturadas por separado para calcular una tasa objetivo para cada cebador.

50 Rondas posteriores podrían incorporar más cebadores RT al sintetizarlos en matriz. La biblioteca de cebadores se puede propagar mediante PCR y convertirse en monocatenaria mediante degradación exonucleolítica selectiva de la hebra que no incluye un grupo de bloqueo en el cebador de PCR (Xu et al. 2018). Se podría usar una matriz grande

para sintetizar múltiples grupos de cebadores: si cada grupo tiene un mango de PCR específico, se podría usar una matriz para generar docenas de grupos de miles de cebadores cada uno de los cuales podría amplificarse selectivamente.

Transcripción inversa múltiple:

- 5 Es posible que la captura múltiple de la diana se pueda realizar en varias etapas durante el protocolo de generación de la biblioteca RNA-seq. Sin embargo, creemos que la transcripción inversa es la más fácil de paralelar. Las reacciones de PCR altamente multiplexadas son muy difíciles de llevar a cabo con éxito. Las reacciones de PCR incluyen muchos (10-20) ciclos. Esto significa que los problemas con el apareamiento fuera del objetivo se exacerban después de un crecimiento exponencial a través de estos ciclos que a menudo supera al del objetivo deseado. En la PCR múltiple, a cada diana se le asignan dos cebadores de PCR específicos. El objetivo es que estos dos cebadores amplifiquen específicamente su diana únicamente. Sin embargo, en un grupo grande de cebadores, habrá varias combinaciones que se aparean con otros cebadores del grupo. Debido a que la concentración de cebadores es mucho mayor que la de las moléculas plantilla, estos dímeros de cebadores dominarán el grupo al final de la PCR. La inviabilidad de la PCR altamente multiplexada es la razón por la cual muchos protocolos de amplificación dirigida, como la secuenciación del exoma, a menudo utilizan sondas de inversión molecular para capturar dianas (Hiatt et al. 2013). En dichos protocolos, la especificidad de la diana se logra mediante una única etapa de apareamiento entre la sonda y la diana. Las sondas específicas para la diana añaden mangos de PCR, que luego se usan en una amplificación por PCR genérica diana. Los métodos de indexación combinatoria de células individuales se basan en la indexación en varias etapas durante la generación de la biblioteca: un método de sonda de inversión para capturar dianas del ADNc no permitiría suficientes etapas de indexación.

Para la captura de dianas múltiples, usamos un cebador de transcripción inversa específico, seguido de una reacción de PCR que amplifica todas las moléculas que transcribimos de forma inversa. Por lo tanto, nuestra estrategia es análoga al uso de sondas de inversión molecular para la amplificación de ADN dirigida: un solo etapa(transcripción inversa) se dirige selectivamente a las transcripciones de interés y agrega un controlador de PCR general que se puede usar para amplificar todas las moléculas diana durante la PCR. Por tanto, es fundamental una alta especificidad durante la transcripción inversa. Mantener una temperatura alta después del apareamiento de los cebadores RT es útil para el cebado específico múltiple. Los protocolos normales de transcripción inversa desnaturalizan una mezcla de ARN y cebador de transcripción inversa y se enfrían a 4 grados para permitir el apareamiento. Esta baja temperatura de apareamiento es demasiado permisiva para los eventos de apareamiento fuera de objetivo. Necesitamos asegurarnos de que los únicos eventos de apareamiento que puedan extenderse sean aquellos donde la totalidad de los cebadores RT altamente específicos que hemos diseñado hayan encontrado sus objetivos. Por lo tanto, mantenemos una temperatura alta durante todo el protocolo, inspirado en otros métodos de transcripción inversa específicos múltiples (Xu et al. 2018). Desnaturalizamos una mezcla de células fijas, grupo de cebadores RT y dNTP a 65 °C, apareamos a 53 °C y luego añadimos una mezcla de enzima de transcripción inversa/tampón que está preequilibrada a 53 °C a la reacción de apareamiento, y extendemos a 53 °C durante 20 minutos. Por lo tanto, los cebadores de RT no tienen la oportunidad de aparearse a baja temperatura entre las etapas de desnaturalización y extensión.

El resto del método sigue los métodos descritos en los Ejemplos 1 y 3. Se liga un adaptador de horquilla in situ, añadiendo un índice celular. Las células se agrupan, se lavan y se dividen en pocillos nuevos para la última etapa de indexación. En estos pocillos se lleva a cabo la síntesis de la segunda hebra. Luego se tagmenta el ADNc bicatenario para añadir un segundo mango de PCR general (el primer mango es de ligadura, el segundo es de tagmentación). El ADN se purifica de las células mediante la unión de perlas Ampure y luego se lleva a cabo la PCR, añadiendo un segundo índice.

Resultados preliminares:

- 45 Todos los resultados, que se muestran en las figuras 40-42, provienen de una biblioteca masiva (sin indexación combinatoria de células individuales) in situ (todos los pasos se llevan a cabo en núcleos fijados con paraformaldehído) realizada usando un grupo de genes dirigidos a cebadores RT en el locus LMO2 en células K562.

Referencias:

- 50 Armor, Christopher D., John C. Castle, Ronghua Chen, Tomas Babak, Patrick Loerch, Stuart Jackson, Jyoti K. Shah, et al. 2009. "Digital Transcriptome Profiling Using Selective Hexamer Priming for cDNA Synthesis". *Nature Methods* 6 (9): 647-49.
- Cao, Junyue, Malte Spielmann, Xiaojie Qiu, Xingfan Huang, Daniel M. Ibrahim, Andrew J. Hill, Fan Zhang, et al. 2019. "The Single-Cell Transcriptional Landscape of Mammalian Organogenesis". *Nature* 566 (7745): 496-502.
- 55 Core, Leighton J., André L. Martins, Charles G. Danko, Colin T. Waters, Adam Siepel y John T. Lis. 2014. "Analysis of Nascent RNA Identifies a Unified Architecture of Initiation Regions at Mammalian Promoters and Enhancers". *Nature Genetics* 46 (12): 1311-20.

Gasperi, Molly, Andrew J. Hill, José L. McFaline-Figueroa, Beth Martin, Seungsoo Kim, Melissa D. Zhang, Dana Jackson, et al. 2019. "A Genome-Wide Framework for Mapping Gene Regulation via Cellular Genetic Screens". *Cell* 176 (6): 1516.

5 Gasperi, Molly, Andrew Hill, José L. McFaline-Figueroa, Beth Martin, Cole Trapnell, Nadav Ahituv y Jay Shendure. 2018. "crisprQTL Mapping as a Genome-Wide Association Framework for Cellular Genetic Screens". *bioRxiv*. <https://doi.org/10.1101/314344>.

Hiatt, Joseph B., Colin C. Pritchard, Stephen J. Salipante, Brian J. O'Roak y Jay Shendure. 2013. "Single Molecule Molecular Inversion Probes for Targeted, High-Accuracy Detection of Low-Frequency Variation". *Genome Research* 23 (5): 843-54.

10 Marinov, Georgi K., Brian A. Williams, Ken McCue, Gary P. Schroth, Jason Gertz, Richard M. Myers y Barbara J. Wold. 2014. "From Single-Cell to Cell-Pool Transcriptomes: Stochasticity in Gene Expression and RNA Splicing". *Genome Research* 24 (3): 496-510.

15 Saikia, Mridusmita, Philip Burnham, Sara H. Keshavjee, Michael F. Z. Wang, Pablo Moral-López, Meleana M. Hinchman, Charles G. Danko, John S. L. Parker e Iwijn De Vlaminc. 2018. "Simultaneous Multiplexed Amplicon Sequencing and Transcriptome Profiling in Single Cells". *bioRxiv*. <https://doi.org/10.1101/328328>.

Xie, Shiqi, Jialei Duan, Boxun Li, Pei Zhou y Gary C. Hon. 2017. "Multiplexed Engineering and Analysis of Combinatorial Enhancer Activity in Single Cells". *Molecular Cell* 66 (2): 285-99.e5.

Xu, Hansen, Benjamin J. Fair, Zach Dwyer, Michael Gildea y Jeffrey A. Pleiss. 2018. "Multiplexed Primer Extension Sequencing Enables High Precision Detection of Rare Splice Isoforms". *bioRxiv*. <https://doi.org/10.1101/331629>.

20 Ziegenhain, Christoph, Beate Vieth, Swati Parekh, Björn Reinius, Amy Guillaumet-Adkins, Martha Smets, Heinrich Leonhardt, Holger Heyn, Ines Hellmann y Wolfgang Enard. 2017. "Comparative Analysis of Single-Cell RNA Sequencing Methods". *Molecular Cell* 65 (4): 631-43.e4.

25 La descripción detallada y los ejemplos anteriores se han proporcionado únicamente para mayor claridad de comprensión. De ello no se desprenden limitaciones innecesarias. La descripción no se limita a los detalles exactos mostrados y descritos, ya que se incluirán variaciones obvias para un experto en la técnica dentro de la descripción definida por las reivindicaciones.

30 A menos que se indique lo contrario, se debe entender que todos los números que expresan cantidades de componentes, pesos moleculares, etc. usados en la memoria y las reivindicaciones están modificados en todos los casos por el término "aproximadamente". En consecuencia, a menos que se indique lo contrario, los parámetros numéricos descritos en la memoria y las reivindicaciones son aproximaciones que pueden variar dependiendo de las propiedades deseadas que se buscan obtener mediante la presente descripción. Como mínimo, y no como un intento de limitar la doctrina de los equivalentes al alcance de las reivindicaciones, cada parámetro numérico debería al menos interpretarse a la luz del número de dígitos significativos informados y aplicando técnicas ordinarias de redondeo.

35 Sin perjuicio de que los intervalos numéricos y los parámetros que describen el amplio alcance de la descripción sean aproximaciones, los valores numéricos descritos en los ejemplos específicos se presentan con la mayor precisión posible. Sin embargo, todos los valores numéricos contienen inherentemente un intervalo que resulta necesariamente de la desviación estándar encontrada en sus respectivas mediciones de prueba.

Todos los títulos son para comodidad del lector y no deben usarse para limitar el significado del texto que sigue al título, a menos que así se especifique.

40

REIVINDICACIONES

1. Un método para preparar una biblioteca de secuenciación que comprende ácidos nucleicos de una pluralidad de núcleos o células individuales, comprendiendo el método:
 - 5 (a) proporcionar una pluralidad de núcleos o células en una primera pluralidad de compartimentos, en donde cada compartimento comprende un subconjunto de núcleos o células;
 - (b) marcar el ARN recién sintetizado en los subconjuntos de células o núcleos obtenidos de las células;
 - (c) procesar moléculas de ARN en cada subconjunto de núcleos o células para generar núcleos o células indexados, en donde el procesamiento comprende añadir a los ácidos nucleicos de ARN presentes en cada subconjunto de núcleos o células una primera secuencia índice específica del compartimento para dar como resultado ácidos nucleicos indexados presentes en núcleos o células indexados,10 en donde el procesamiento comprende ligadura, extensión del cebador, hibridación o amplificación; y
 - (d) combinar los núcleos o células indexados para generar núcleos o células indexados agrupados.
2. El método según la reivindicación 1, en donde el procesamiento comprende:
 - 15 poner en contacto subconjuntos con transcriptasa inversa y un cebador que se aparea con ácidos nucleicos de ARN, dando como resultado ácidos nucleicos bicatenarios que comprenden el cebador y la secuencia de nucleótidos de ADN correspondiente de las moléculas de ARN.
3. El método según la reivindicación 2, en donde el cebador comprende una secuencia de nucleótidos poli-T que se aparea con una cola poli(A) de ARNm.
4. El método según la reivindicación 3, en donde el procesamiento comprende además poner en contacto subconjuntos con un segundo cebador, en donde el segundo cebador comprende una secuencia que se aparea con un ácido nucleico de ADN predeterminado; y opcionalmente, el segundo cebador comprende un índice específico del20 compartimento.
5. El método según la reivindicación 2, en donde el cebador comprende una secuencia que se aparea con un ácido nucleico de ARN predeterminado, el método comprende cebadores en diferentes compartimentos que se aparean con diferentes nucleótidos del mismo ácido nucleico de ARN predeterminado.
6. El método según la reivindicación 1, en donde el procesamiento para añadir la secuencia índice específica del primer compartimento comprende un proceso de dos etapas de añadir una secuencia de nucleótidos que comprende una secuencia universal a los ácidos nucleicos de ARN para dar como resultado ácidos nucleicos que comprenden la25 secuencia universal, y luego añadir la secuencia índice específica del primer compartimento a los ácidos nucleicos que comprenden la secuencia universal.
7. El método según la reivindicación 1, donde los ácidos nucleicos de ARN preexistentes y los ácidos nucleicos de ARN recién sintetizados se marcan con el mismo índice en el mismo compartimento.
8. El método según la reivindicación 1, en donde el etiquetado comprende incubar la pluralidad de núcleos o células en una composición que comprende una etiqueta de nucleótidos, en donde la etiqueta de nucleótidos se incorpora al35 ARN recién sintetizado.
9. El método según la reivindicación 1, que comprende además exponer subconjuntos de núcleos o células a una condición predeterminada antes del etiquetado, en donde preferiblemente la condición predeterminada comprende la exposición a un agente que comprende una proteína, una proteína no ribosómica, un policétido, una molécula orgánica, un molécula inorgánica, una molécula de ARN o ARNi, un carbohidrato, una glicoproteína, un ácido nucleico o una combinación de los mismos; pero lo más preferiblemente el agente comprende un fármaco terapéutico.
- 40 10. El método según la reivindicación 9, en donde la condición predeterminada de dos o más compartimentos es diferente cuando el método comprende exponer subconjuntos de núcleos o células a una condición predeterminada antes del etiquetado.
11. El método según la reivindicación 1, que comprende además:
 - 45 distribuir subconjuntos de los núcleos o células indexados agrupados en una segunda pluralidad de compartimentos y añadir a los ácidos nucleicos indexados presentes en subconjuntos de núcleos o células una segunda secuencia índice para generar núcleos o células con doble índice que comprenden fragmentos de ácido nucleico con doble índice, en donde la adición comprende ligadura, extensión del cebador, hibridación, amplificación o transposición;

combinar los núcleos o células de doble índice para generar núcleos o células de doble índice agrupados.

12. El método según la reivindicación 11, que comprende además:

distribuir subconjuntos de núcleos o células de índice dual agrupados en una tercera pluralidad de compartimentos y añadir a los ácidos nucleicos indexados presentes en subconjuntos de núcleos o células una tercera secuencia índice para generar núcleos o células de índice triple que comprenden fragmentos de ácido nucleico de índice triple, en donde la adición comprende ligadura, hibridación, extensión del cebador, amplificación o transposición;

combinar los núcleos o células de triple índice para generar núcleos o células de triple índice agrupados.

13. El método según la reivindicación 1, en donde la adición comprende la ligadura de la secuencia índice específica del primer compartimento, que comprende además añadir una segunda secuencia índice para generar núcleos o células con doble índice que comprenden fragmentos de ácido nucleico con doble índice, en donde la adición comprende transposición.

14. El método según la reivindicación 11, en donde la adición comprende la ligadura de la secuencia índice específica del segundo compartimento, que comprende además añadir una tercera secuencia índice para generar núcleos o células con índice dual que comprenden fragmentos de ácido nucleico con índice triple, en donde la adición comprende transposición.

15. El método según la reivindicación 1, que comprende además obtener los ácidos nucleicos indexados a partir de los núcleos o células indexados agrupados, produciendo así una biblioteca de secuenciación a partir de la pluralidad de núcleos o células.

16. El método según la reivindicación 11, que comprende además obtener los ácidos nucleicos de doble índice a partir de los núcleos o células de doble índice agrupados, produciendo así una biblioteca de secuenciación a partir de la pluralidad de núcleos o células.

17. El método según la reivindicación 12, que comprende además obtener los ácidos nucleicos de triple índice a partir de los núcleos o células de triple índice agrupados, produciendo así una biblioteca de secuenciación a partir de la pluralidad de núcleos o células.

18. El método según la reivindicación 2, en donde el cebador comprende una secuencia que se aparea con un ácido nucleico de ARN predeterminado, y preferiblemente en donde el ácido nucleico de ARN predeterminado es un ARNm.

19. El método según la reivindicación 2, en donde el cebador comprende un cebador de cambio de plantilla.

20. El método según la reivindicación 8, en donde la etiqueta de nucleótido comprende un análogo de nucleótido, un nucleótido marcado con hapteno, un nucleótido mutagénico o un nucleótido que puede modificarse mediante una reacción química.

21. El método según la reivindicación 8, en donde se incorpora más de una etiqueta de nucleótidos en el ARN recién sintetizado, y preferiblemente la relación de la etiqueta o etiquetas de nucleótidos es diferente para diferentes compartimentos o puntos temporales.

22. El método según la reivindicación 9, en donde la exposición y el etiquetado ocurren al mismo tiempo o la exposición ocurre antes del etiquetado.

FIG. 1

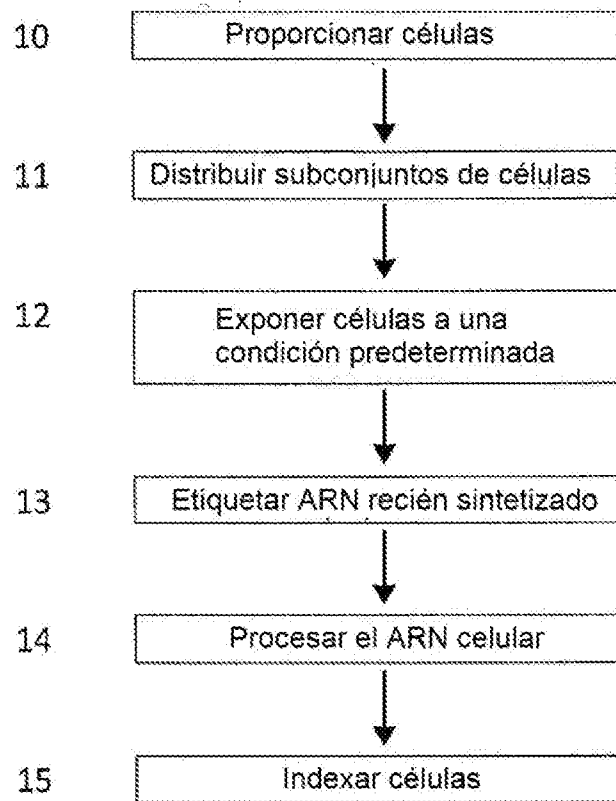


FIG. 2

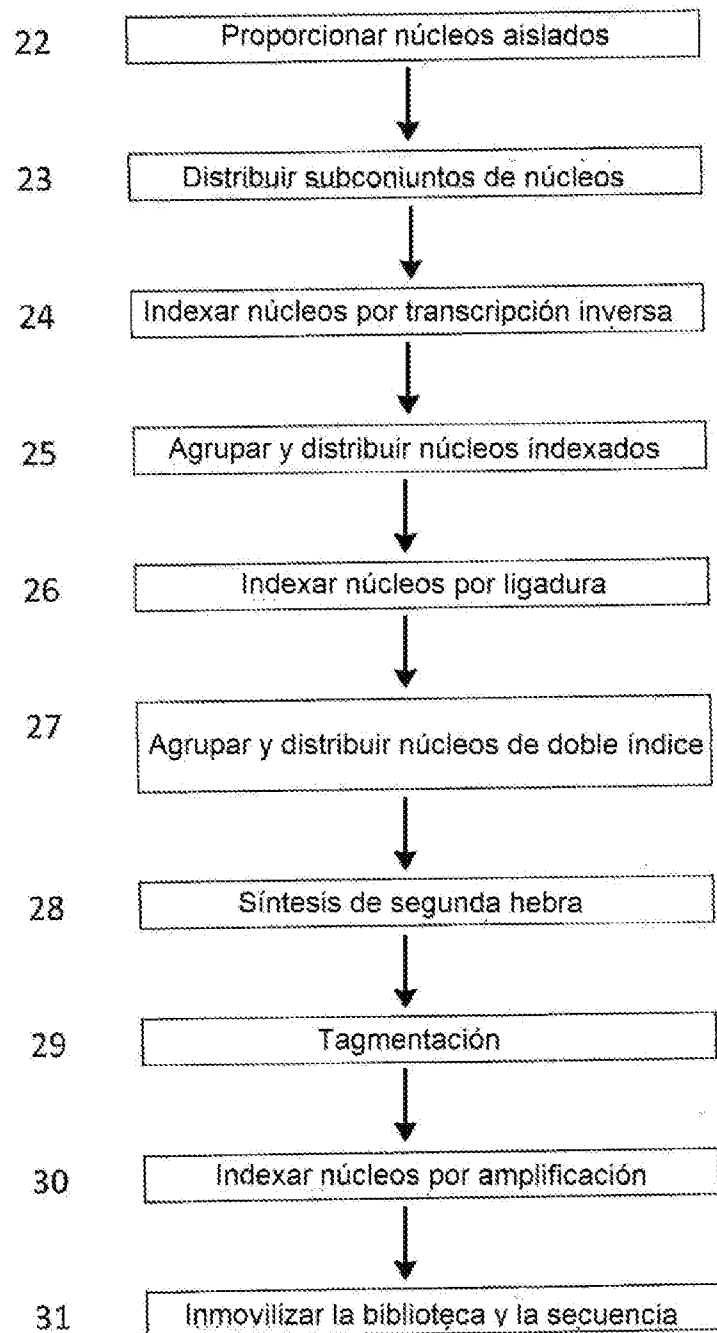
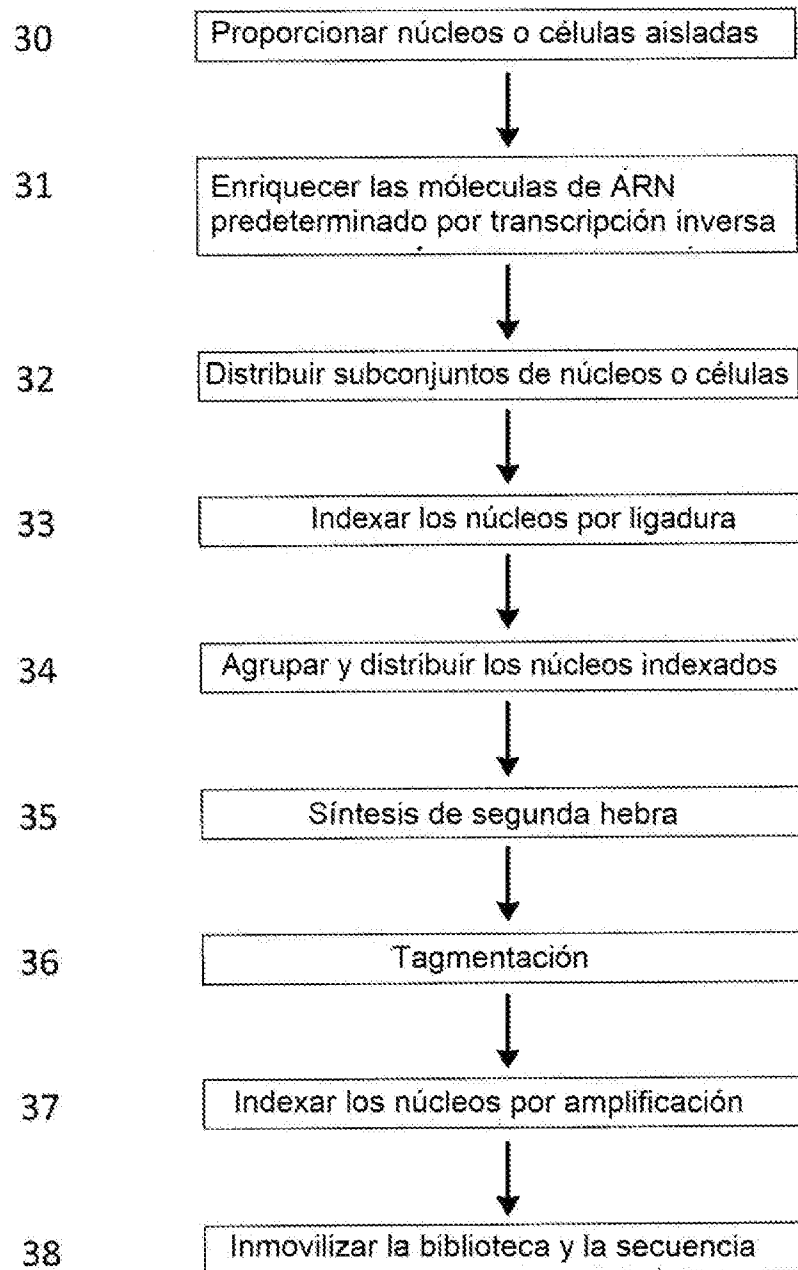


FIG. 3



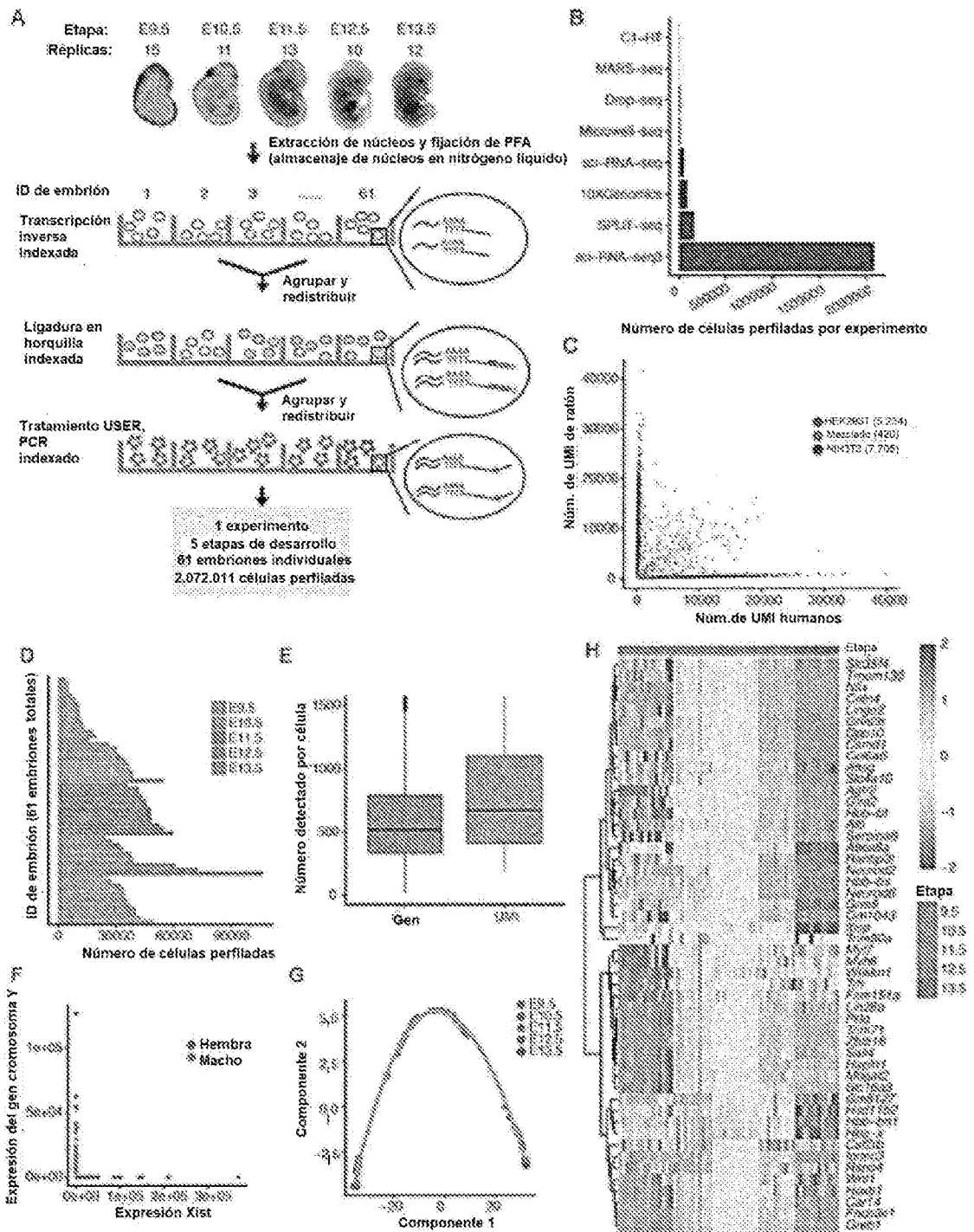


FIG. 4

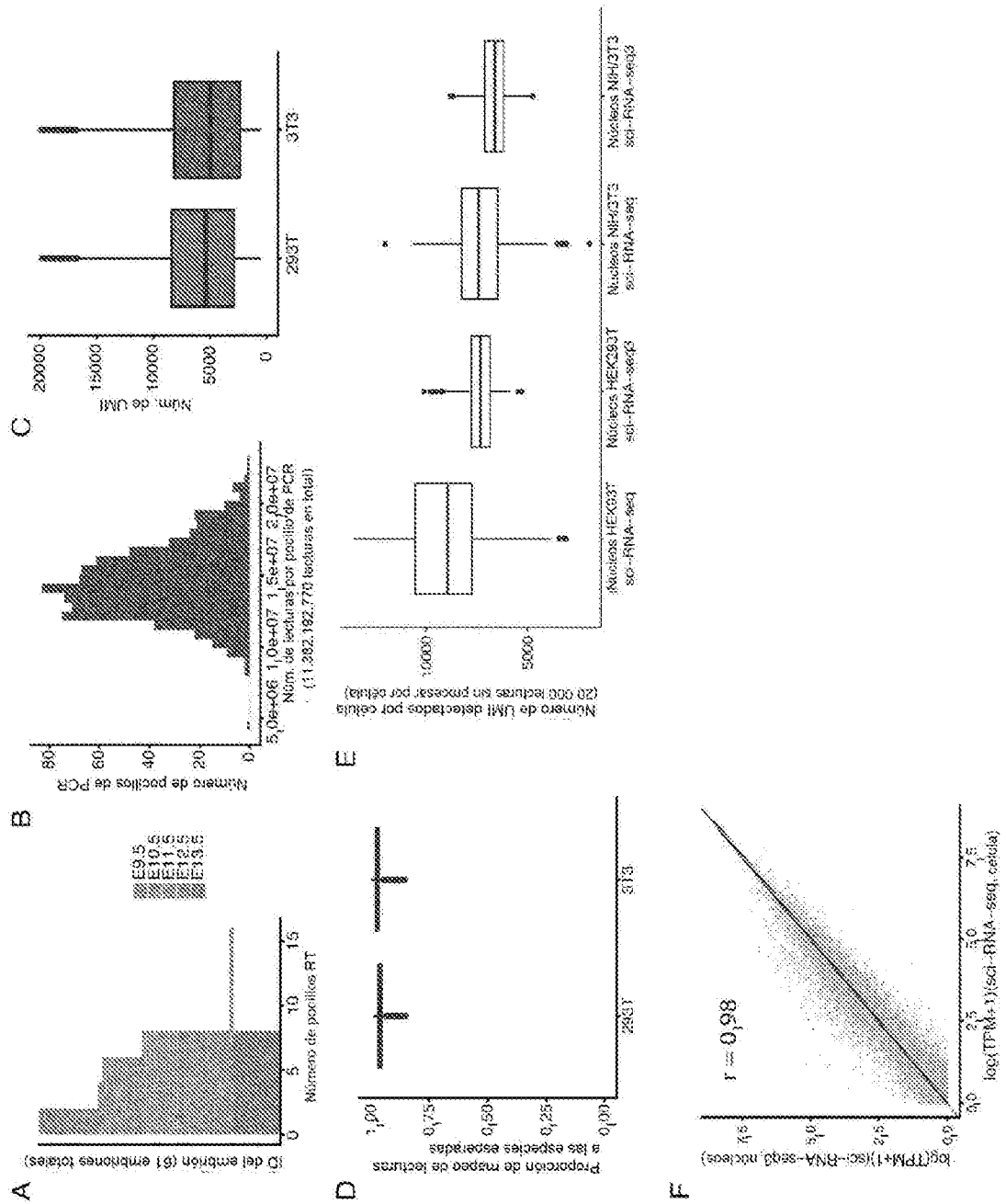


FIG. 5

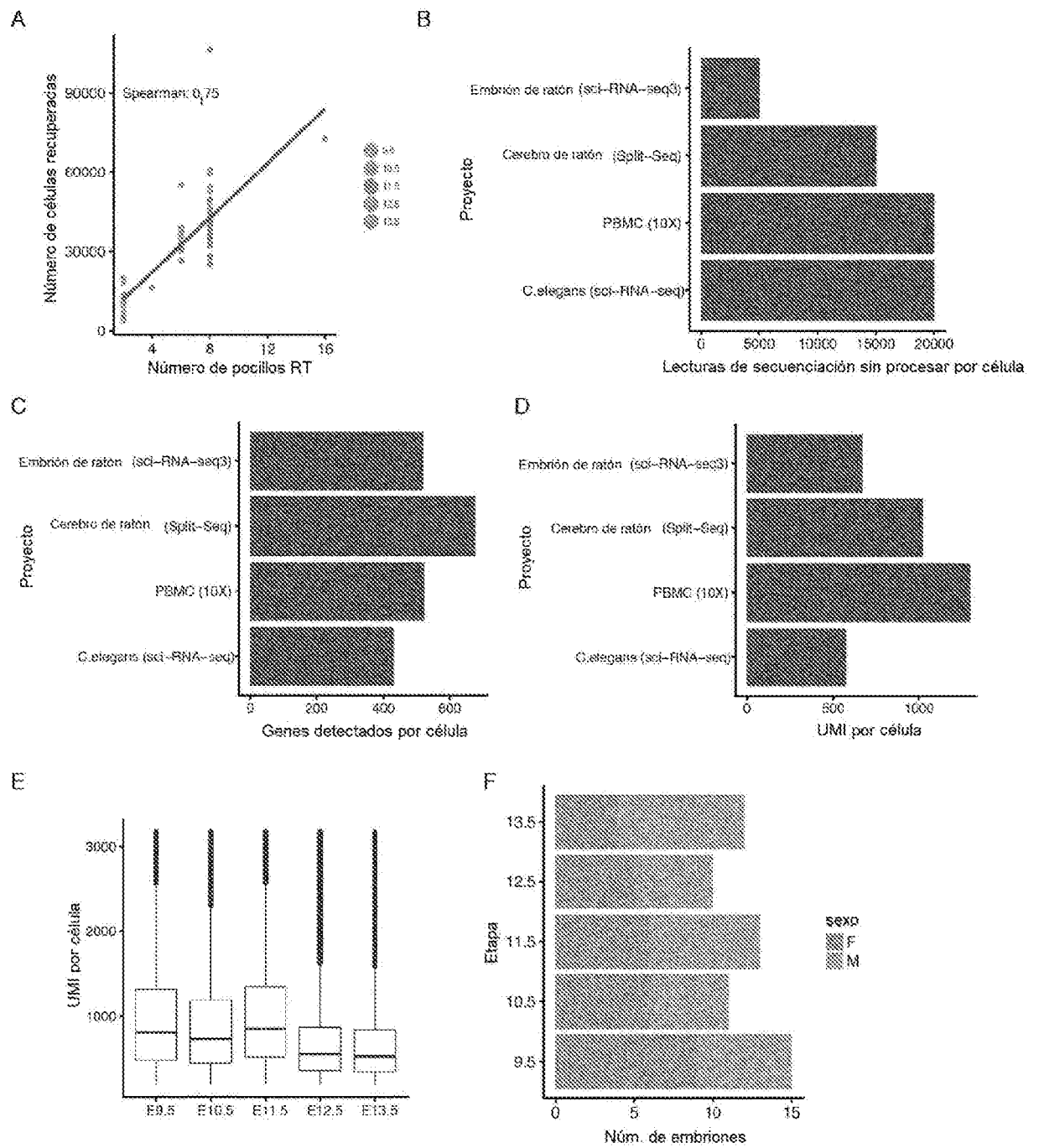


FIG. 6

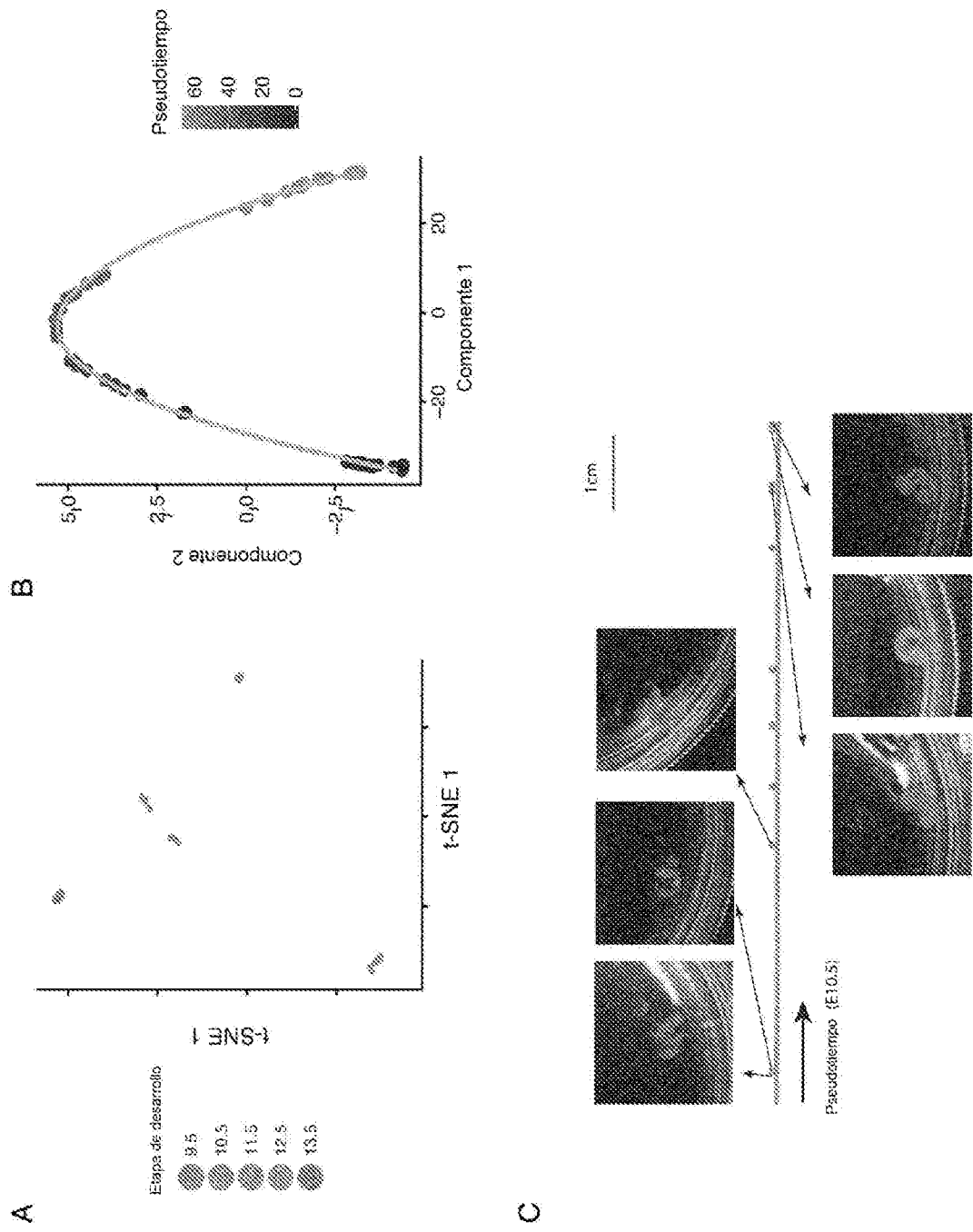


FIG. 7

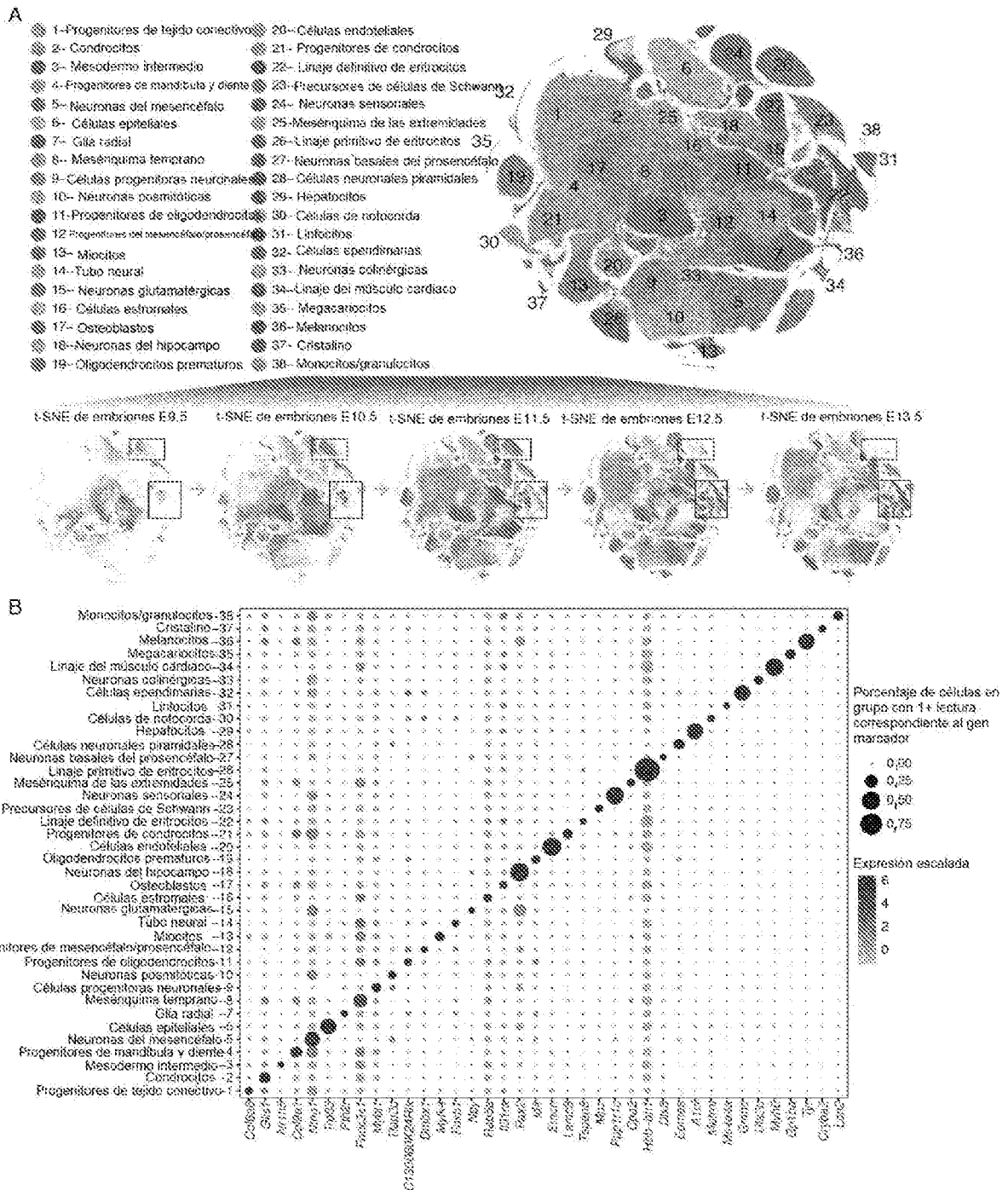
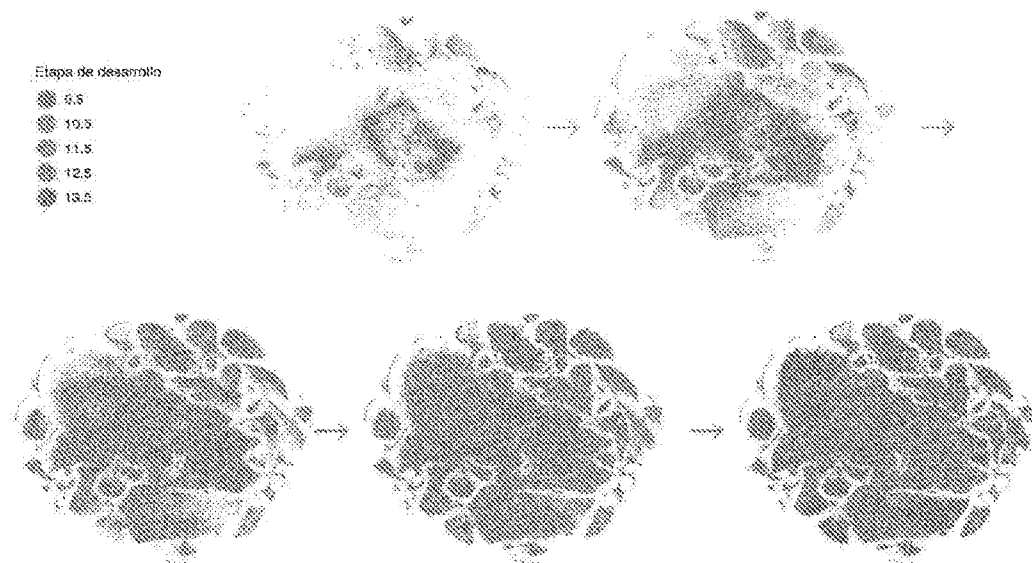
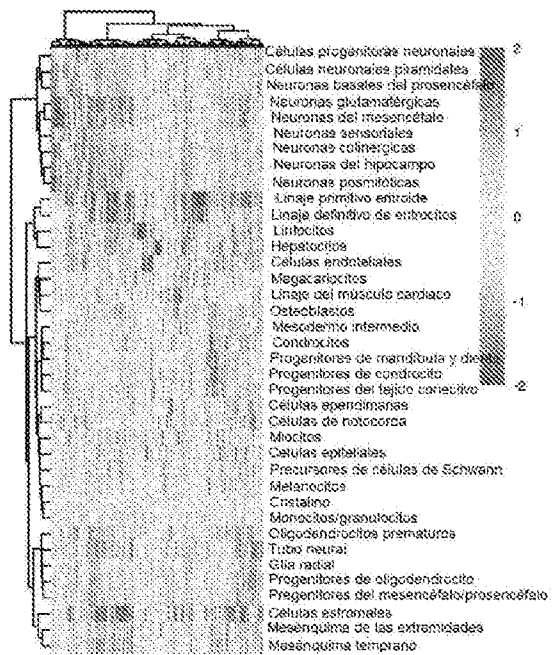


FIG. 8

A



B



C

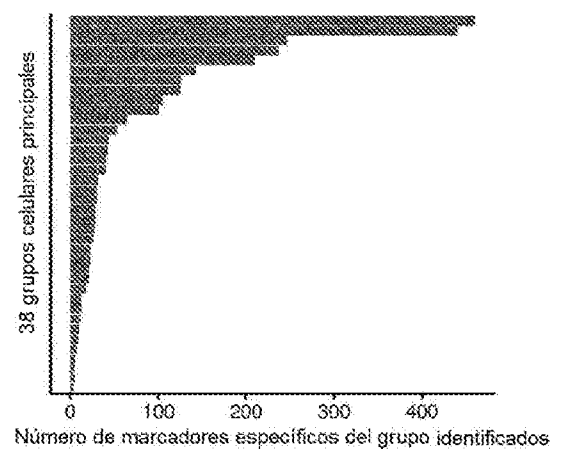


FIG. 9

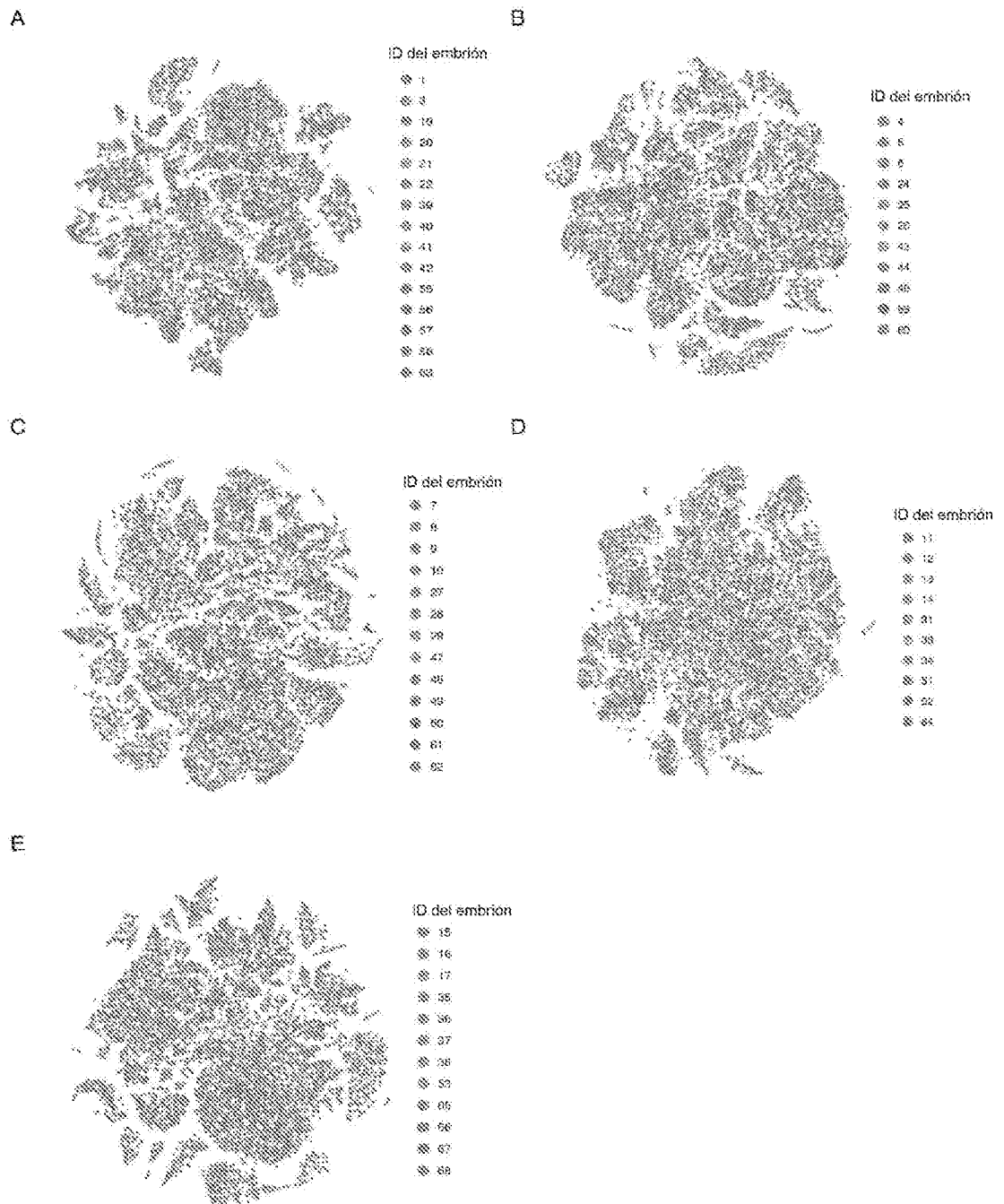


FIG. 10

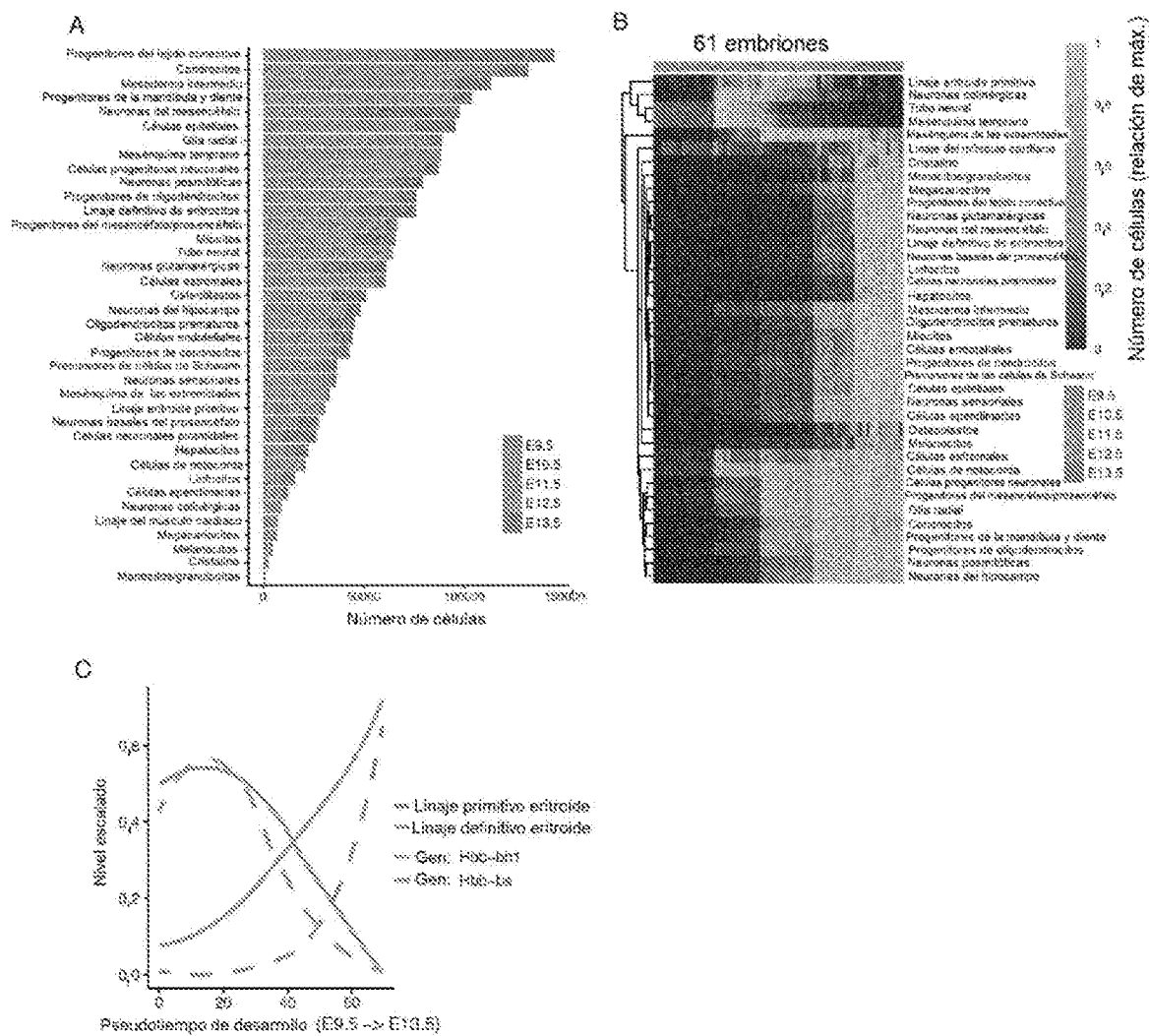


FIG. 11

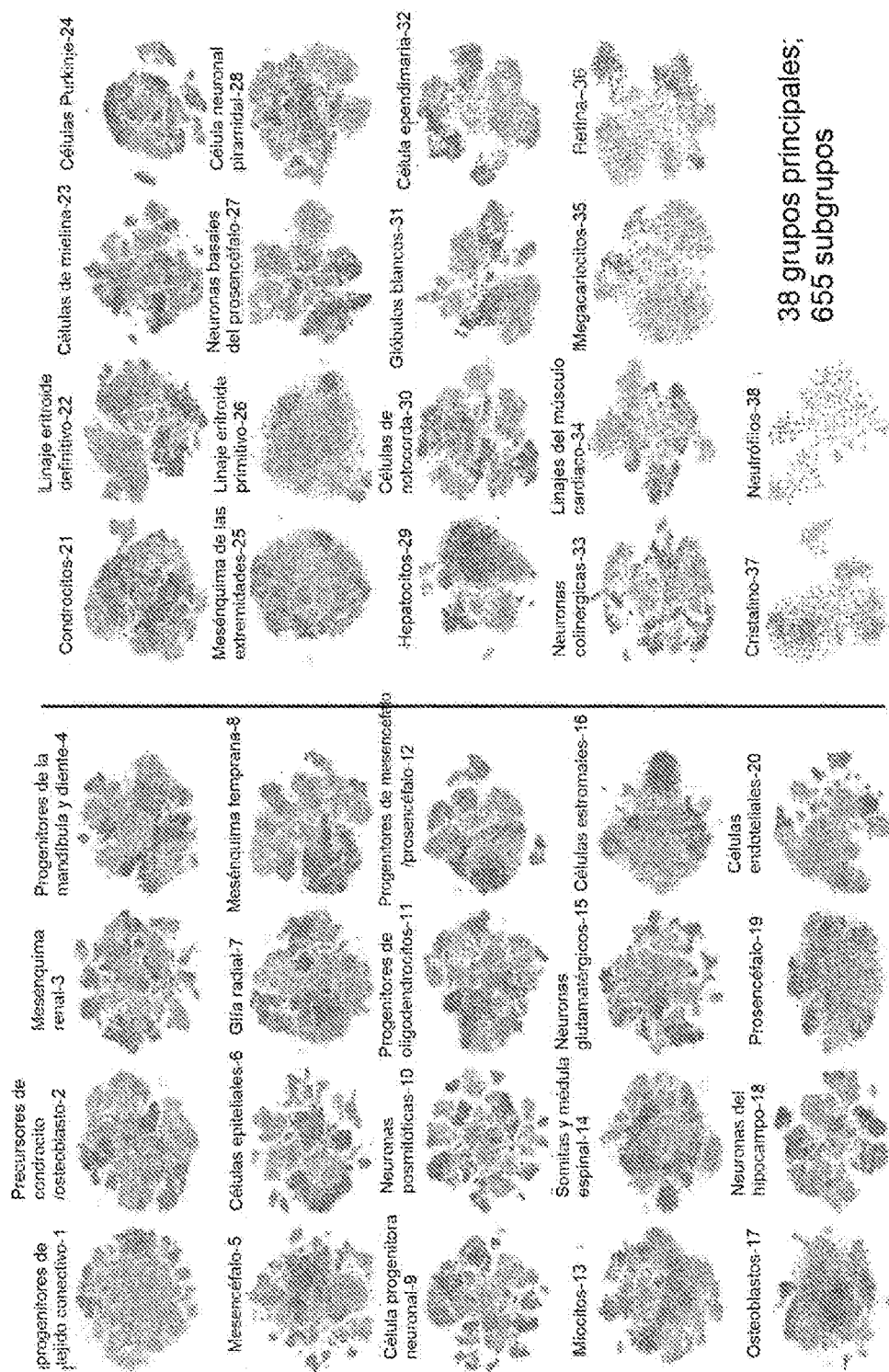


FIG. 12

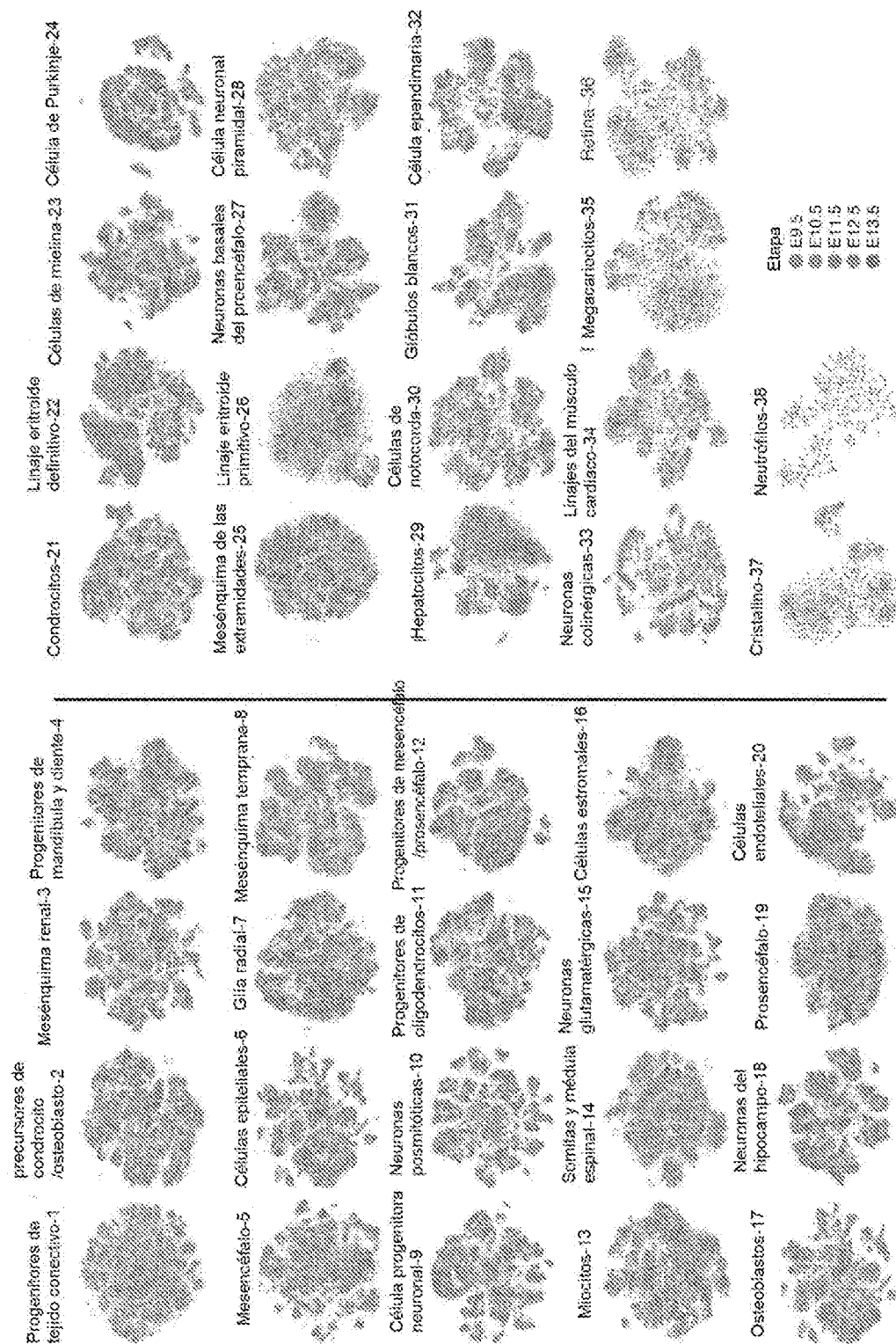
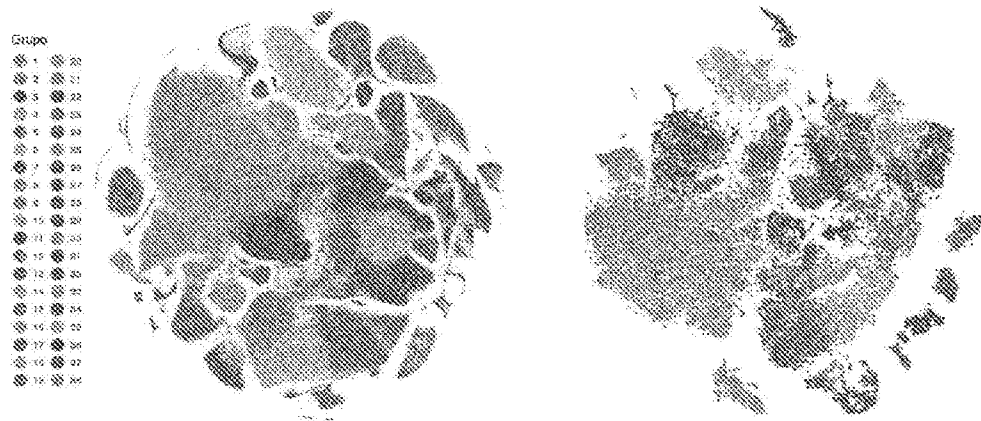
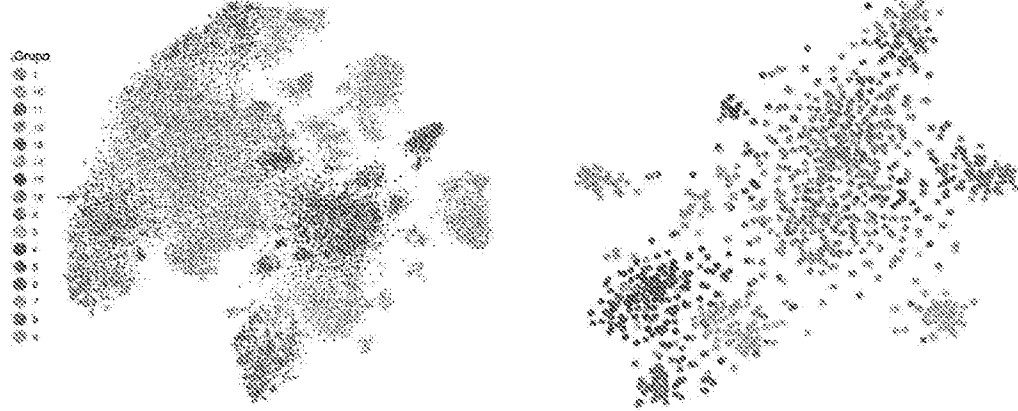


FIG. 13

A



B



C

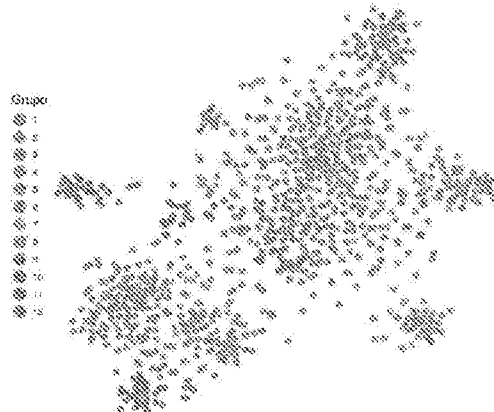


FIG. 14

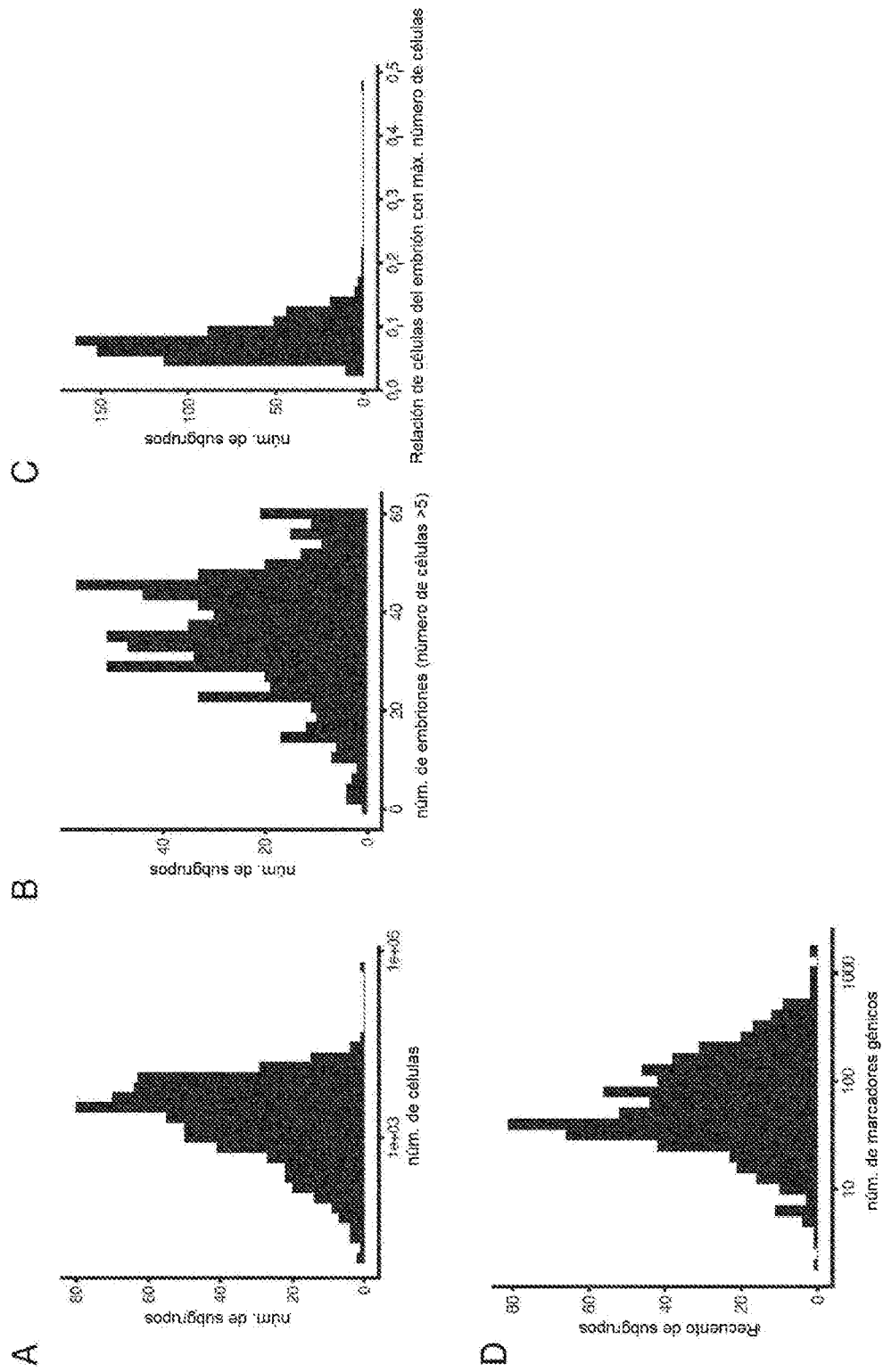


FIG. 15

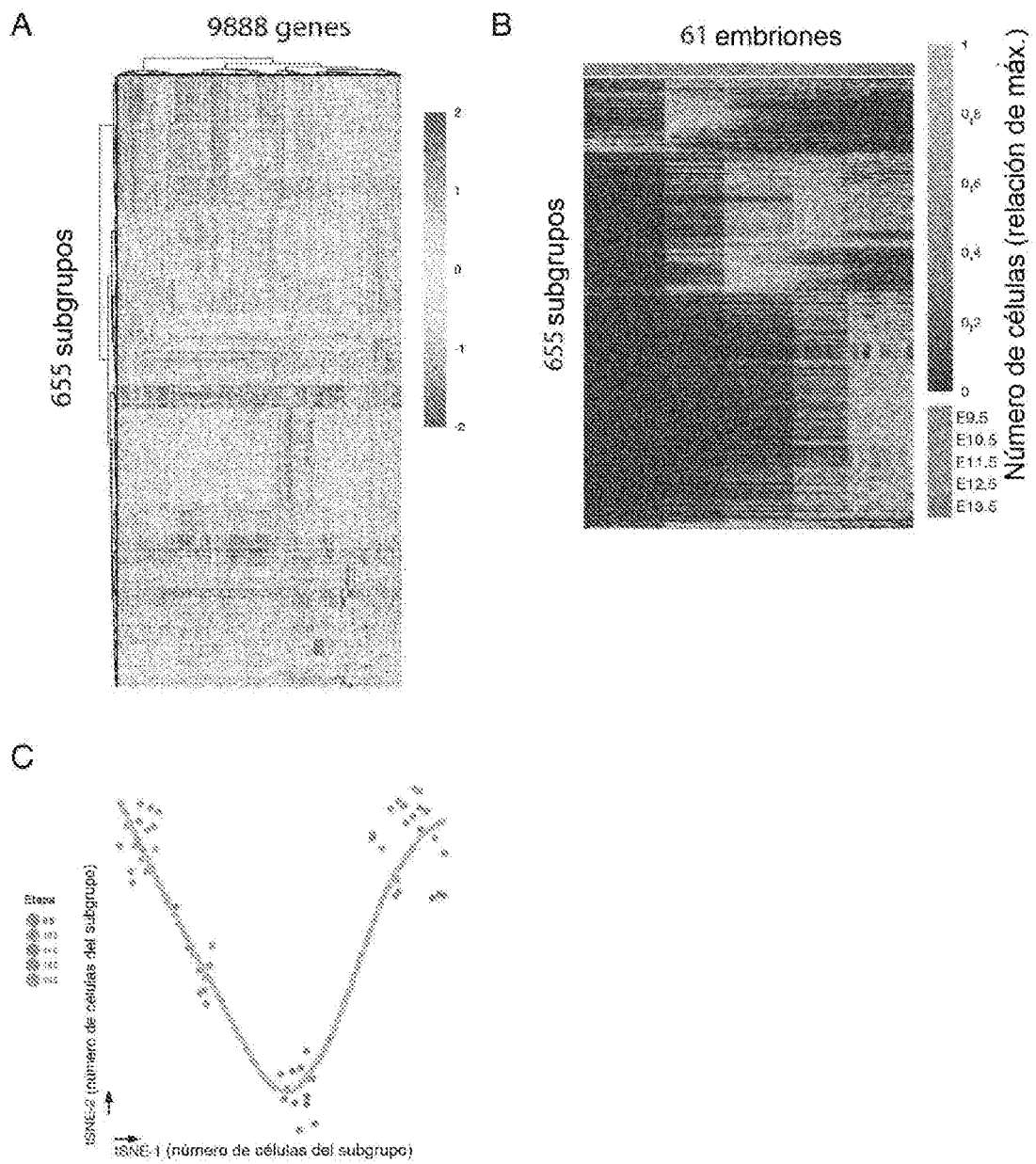


FIG. 16

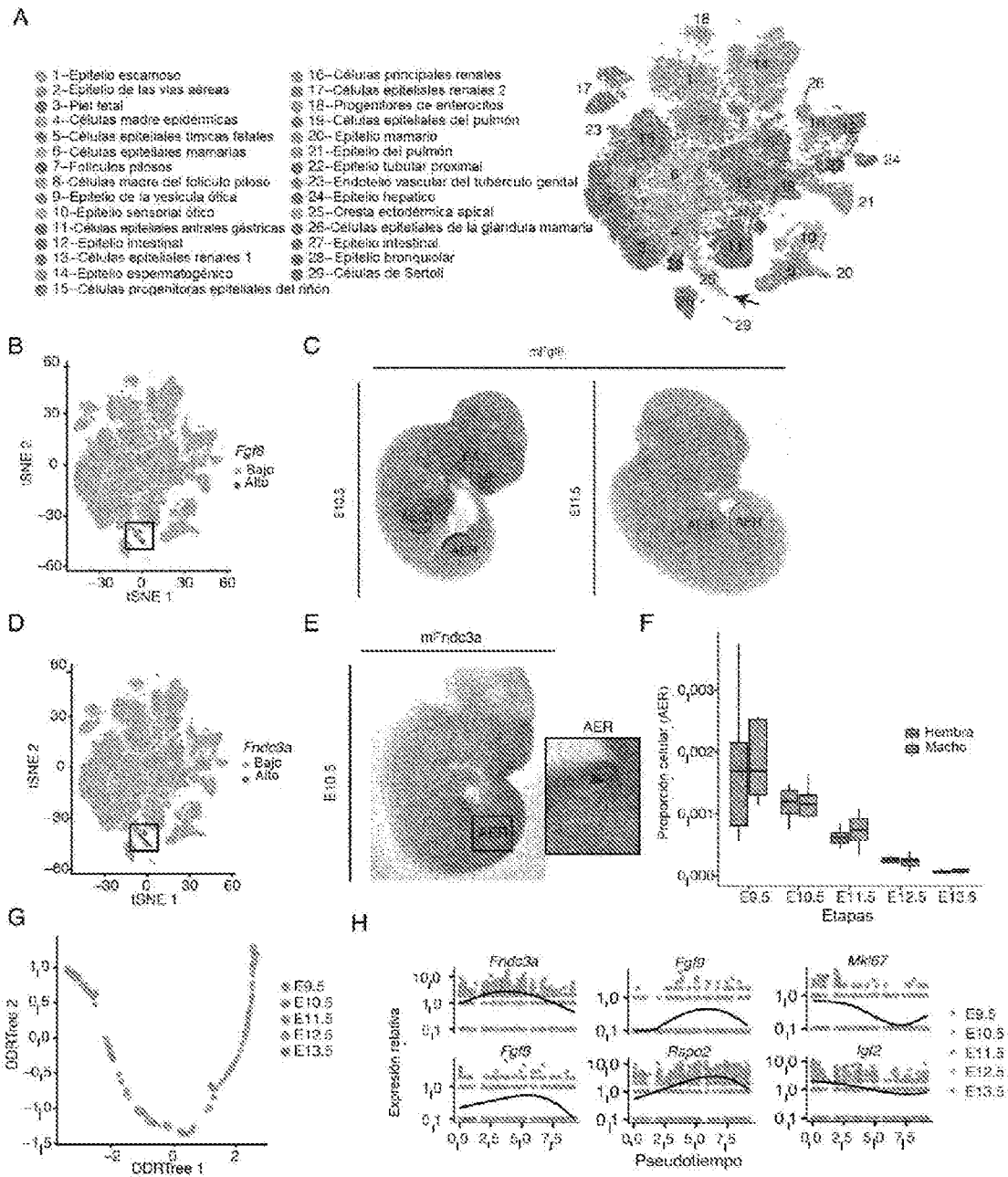
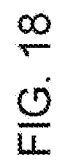


FIG. 17



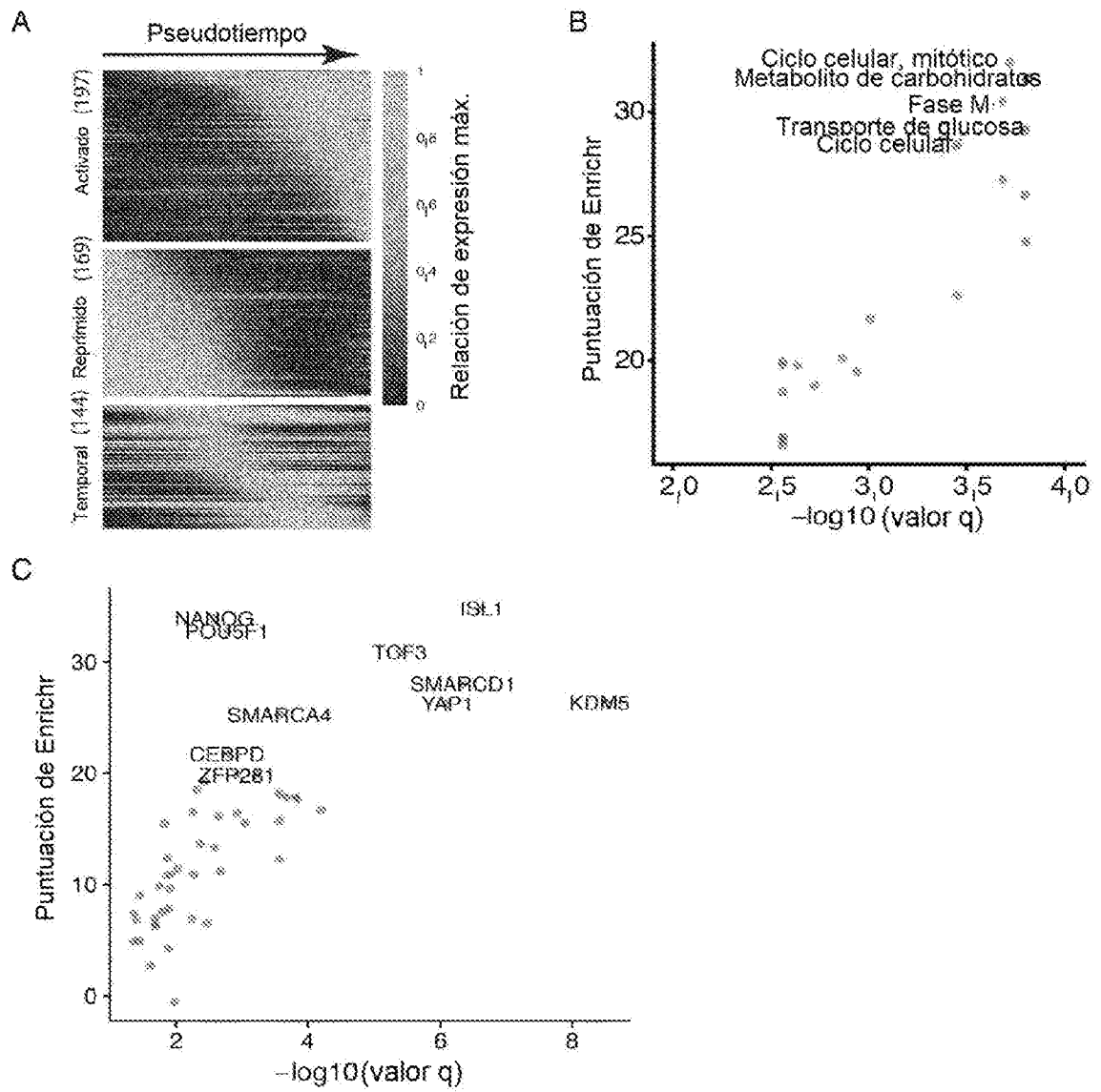


FIG. 19

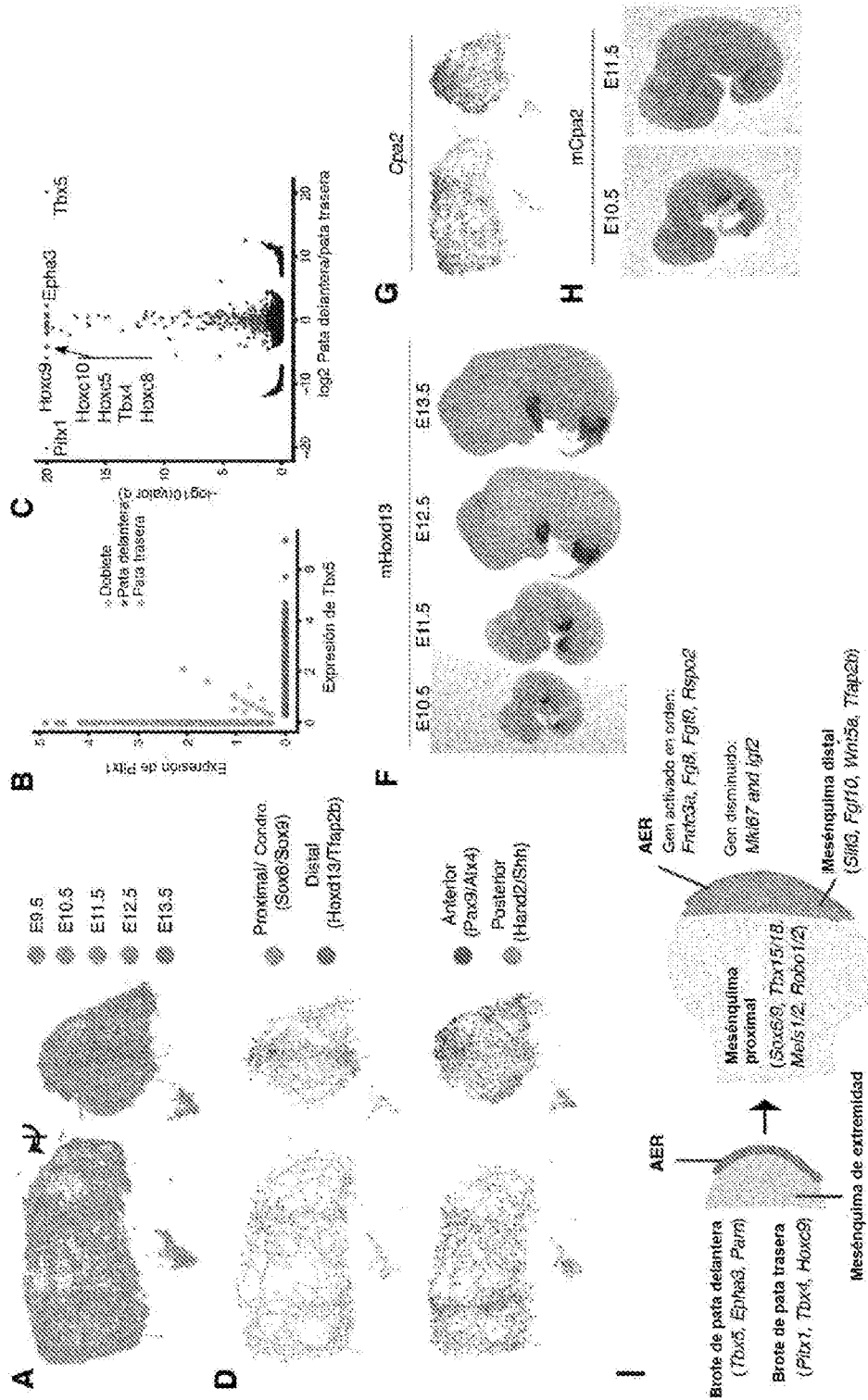


FIG. 20

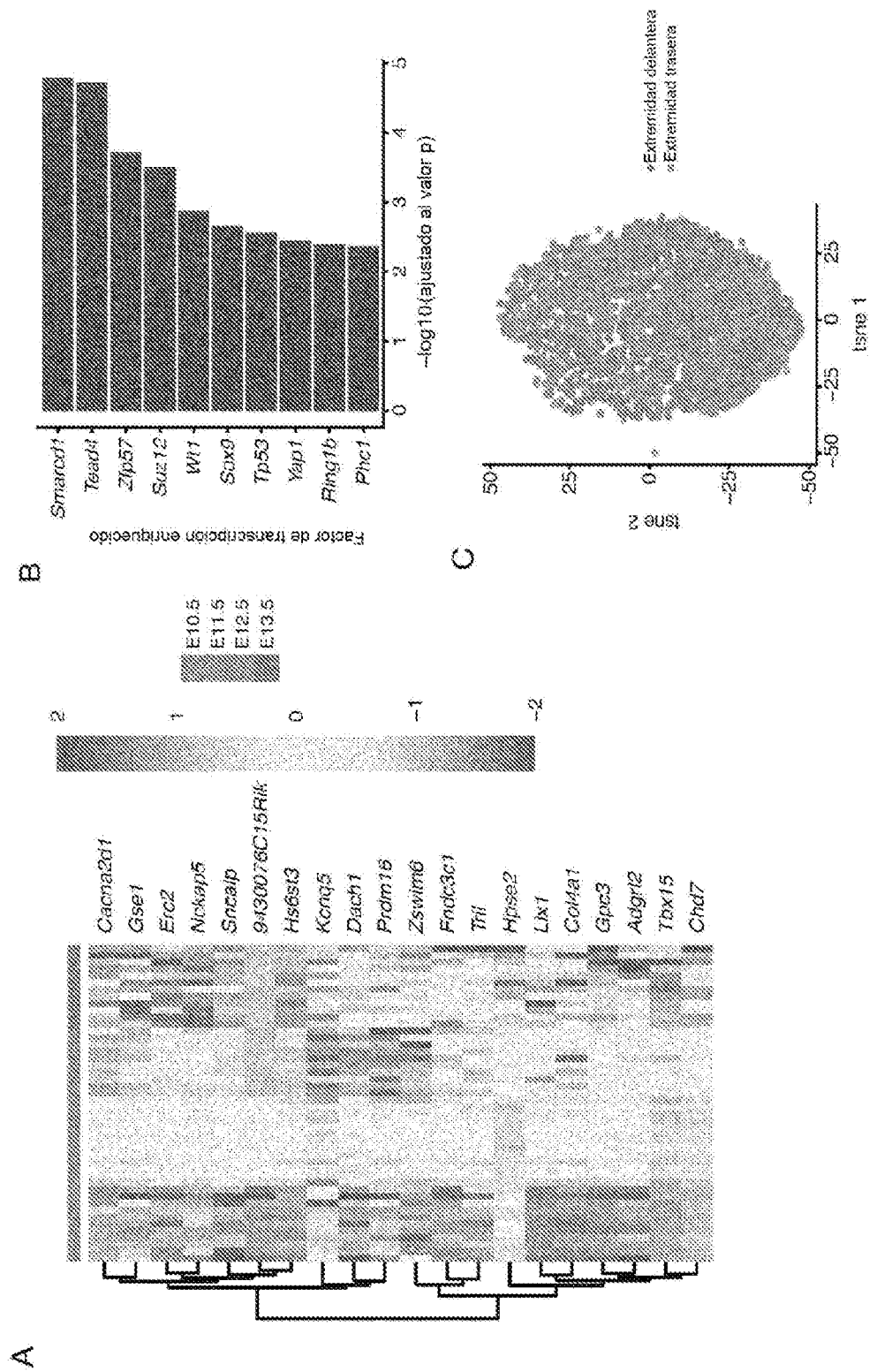


FIG. 21

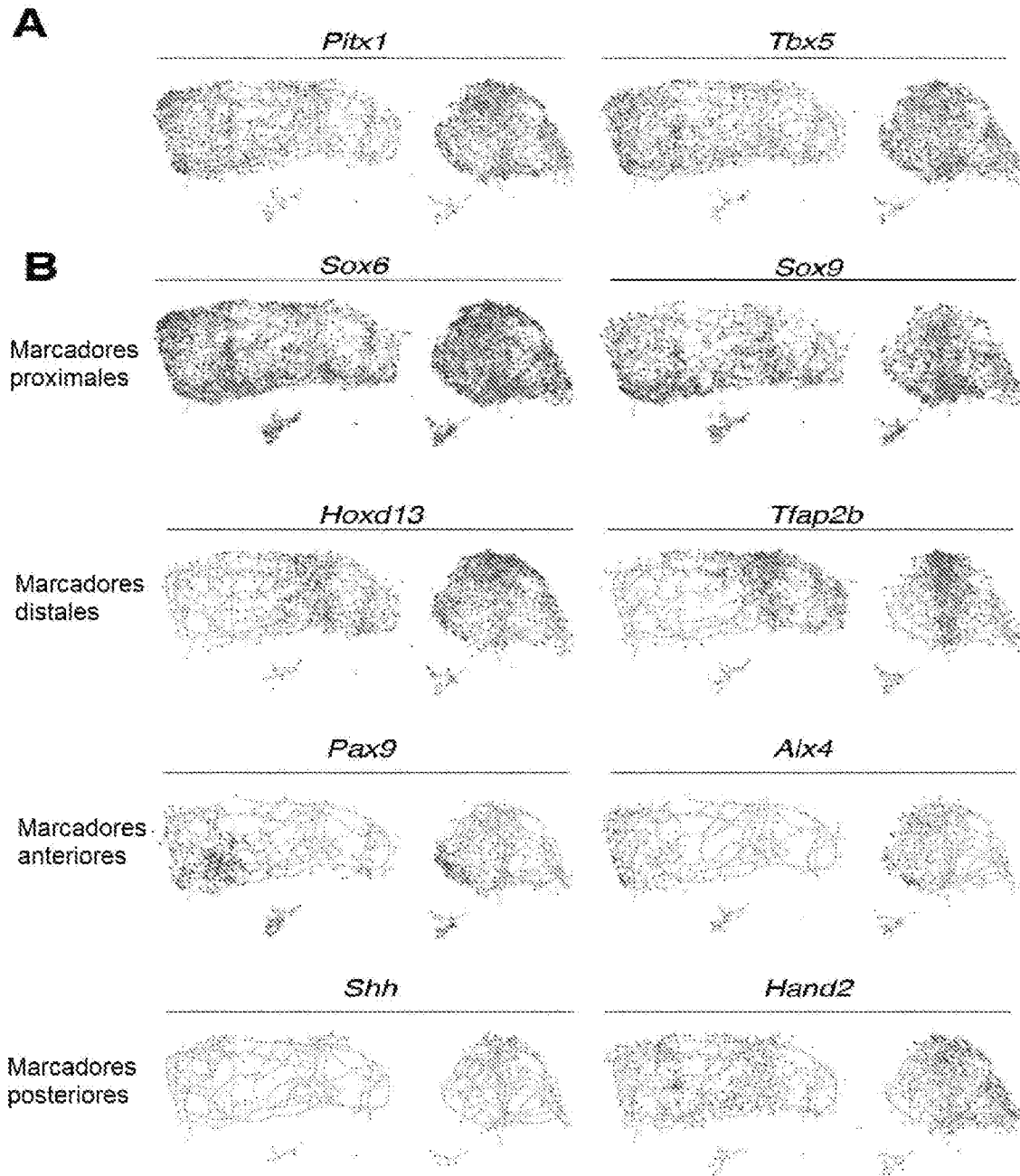


FIG. 22

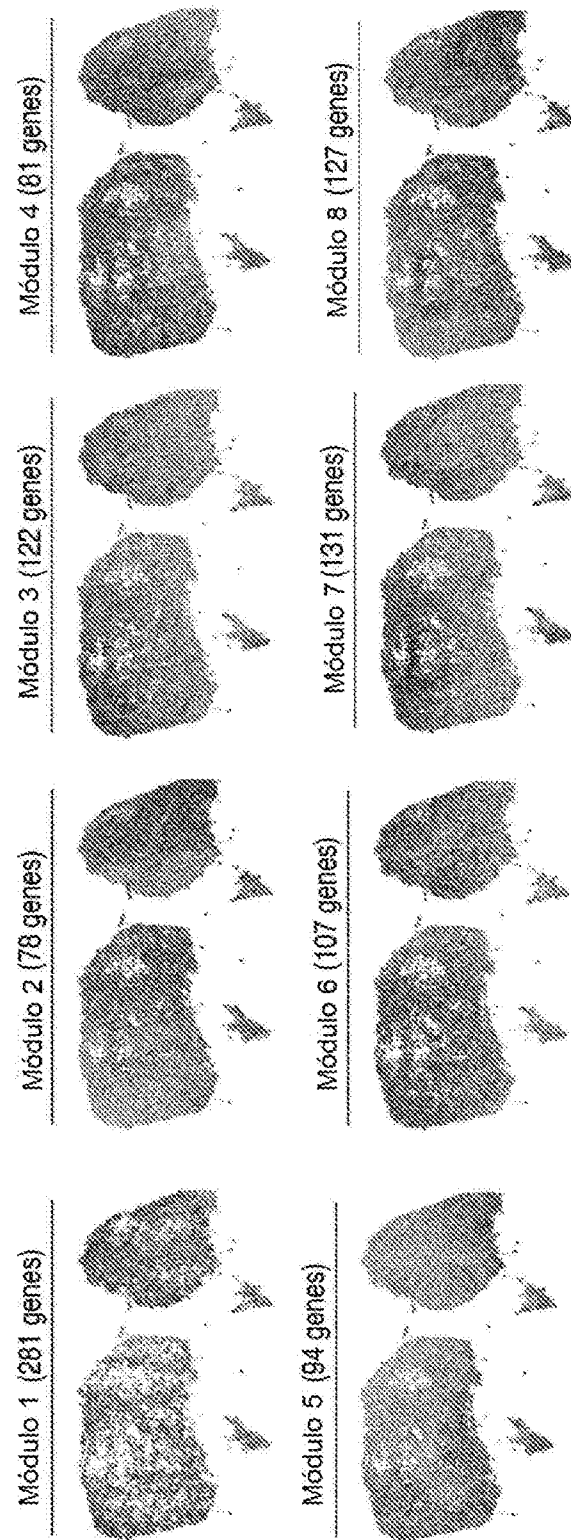


FIG. 23

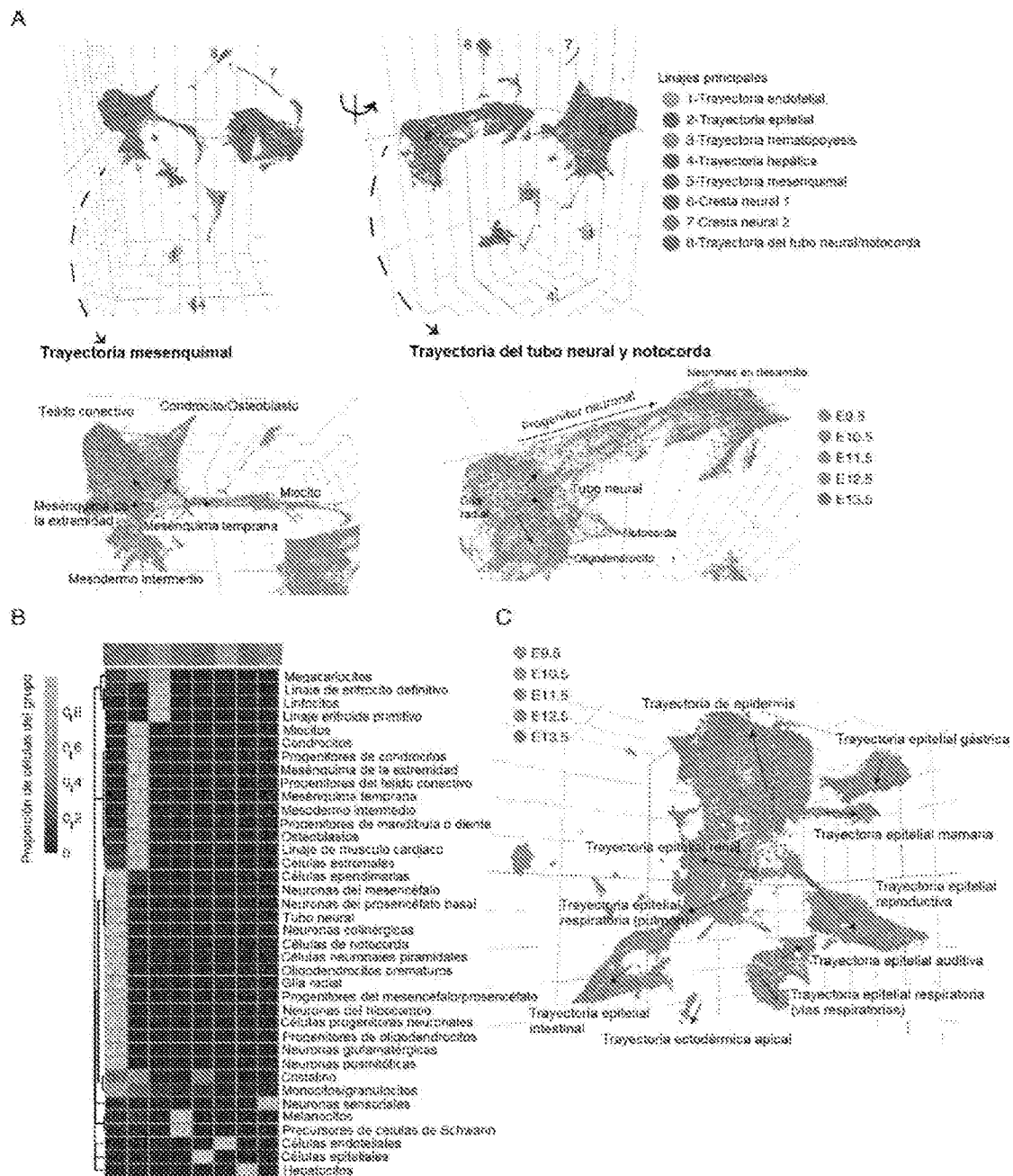


FIG. 24

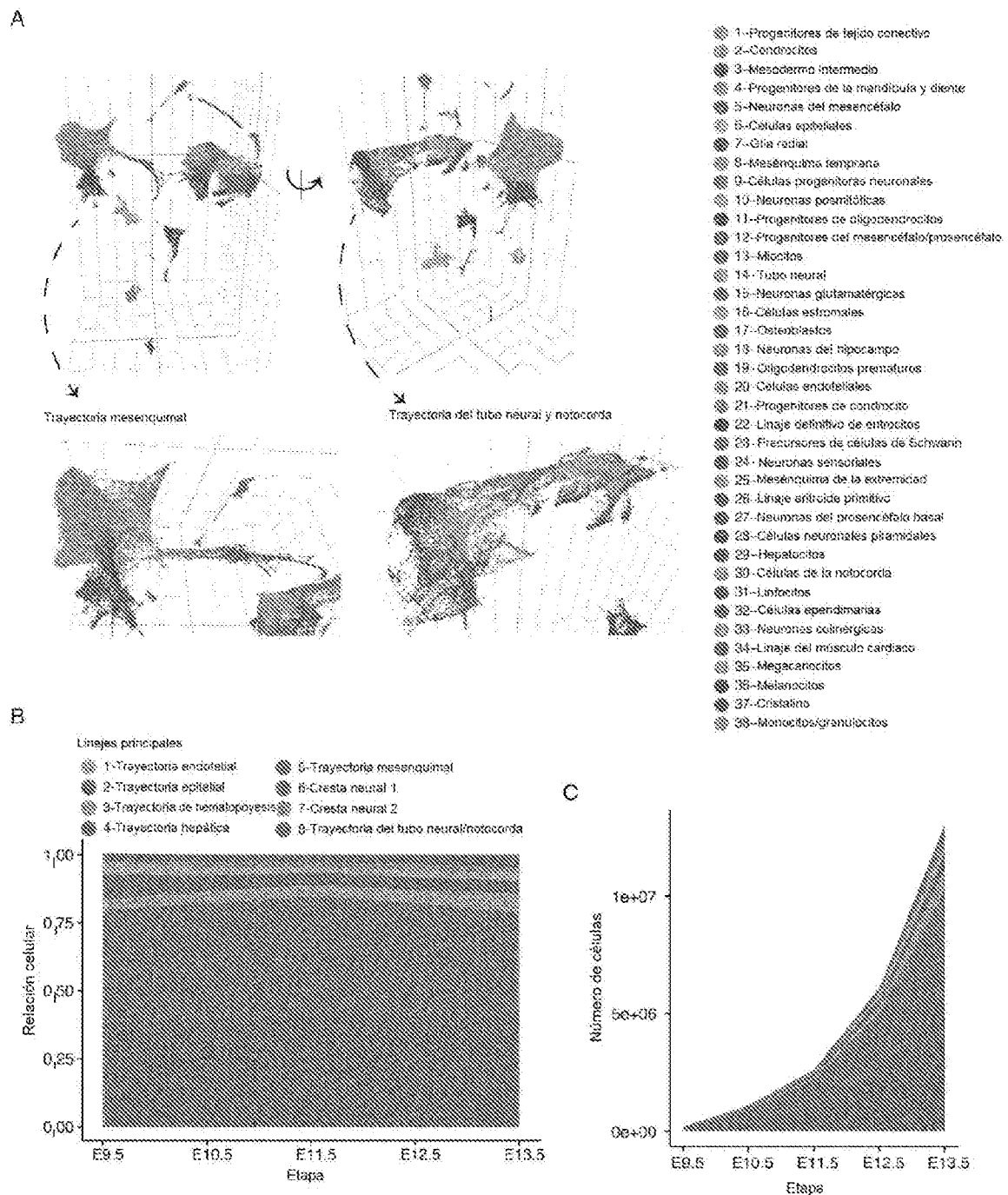


FIG. 25

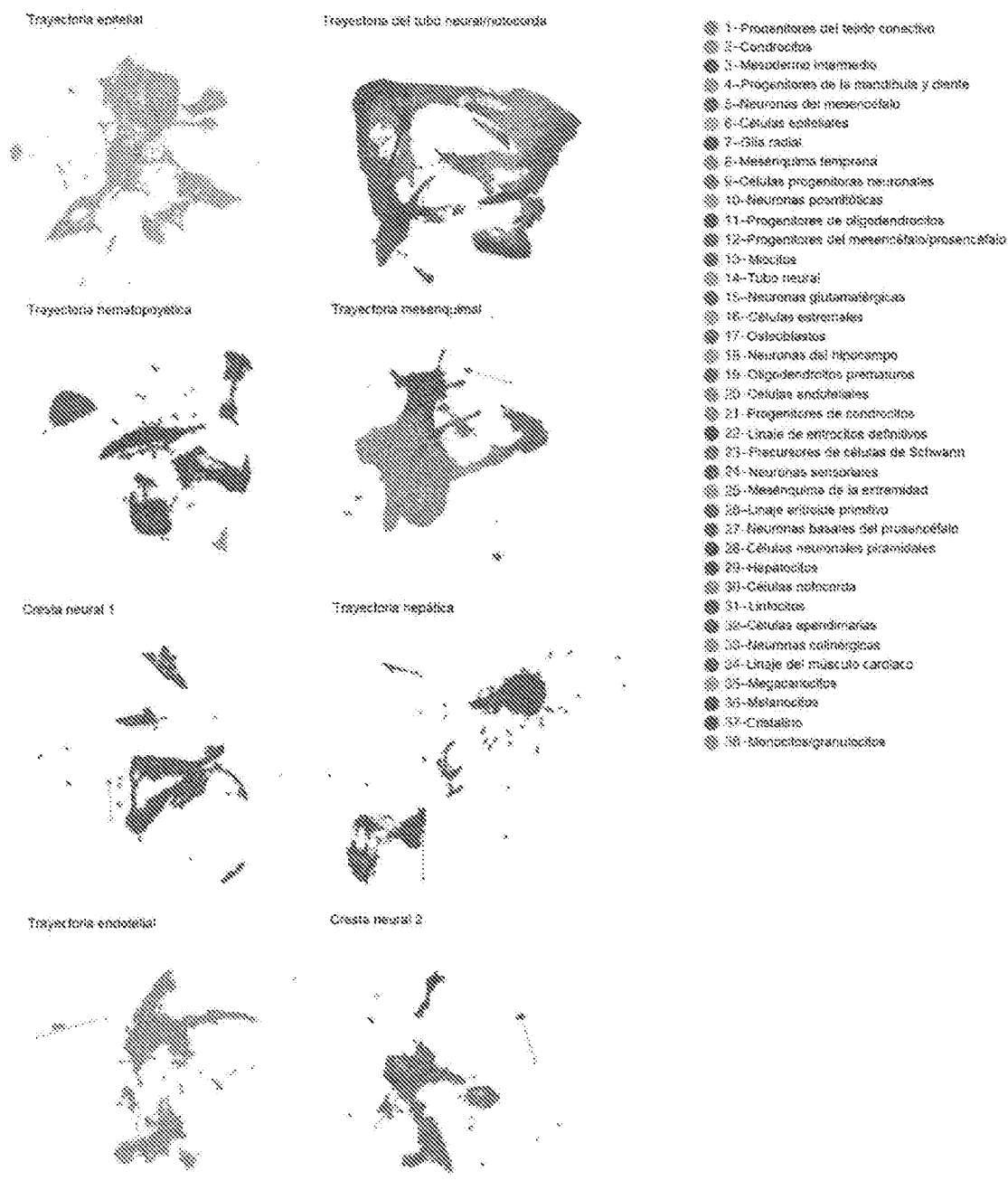
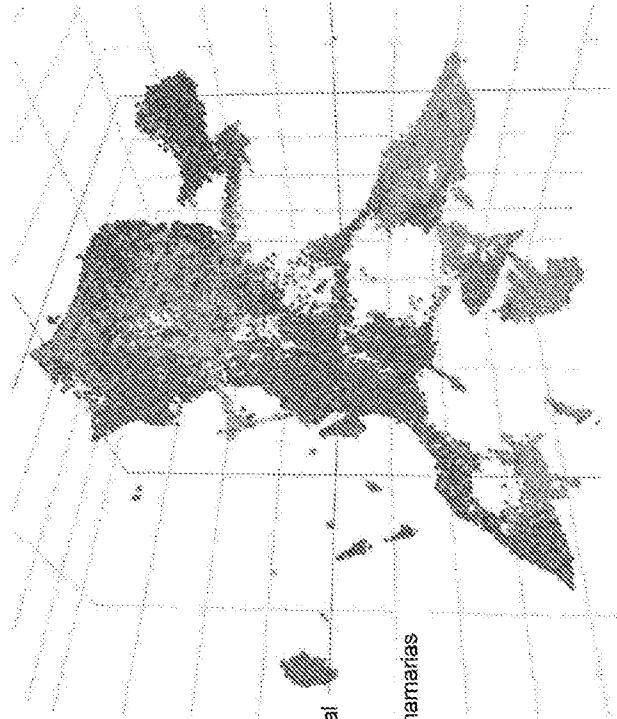


FIG. 26



FIG. 27



- 1-Epitelio escamoso
- 2-Epitelio de las vías aéreas
- 3-Piel fetal
- 4-Células madre epidérmicas
- 5-Células epiteliales tímica fetales
- 6-Células epiteliales mamarias
- 7-Folículos pilosos
- 8-Células madre del folículo piloso
- 9-Epitelio de vesícula ótica
- 10-Epitelio sensorial ótico
- 11-Células epiteliales antrales gástricas
- 12-Epitelio intestinal
- 13-Células epiteliales renales 1
- 14-Epitelio espermatogénico
- 15-Células progenitoras epiteliales renales
- 16-Células renales principales
- 17-Células epiteliales renales 2
- 18-Progenitores de enterocito
- 19-Células epiteliales pulmonares
- 20-Epitelio mamario
- 21-Epitelio pulmonar
- 22-Epitelio tubular proximal
- 23-Endotelio vascular del tubérculo genital
- 24-Epitelio hepático
- 25-Cresta ectodérmica apical
- 26-Células epiteliales de las glándulas mamarias
- 27-Epitelio intestinal
- 28-Epitelio bronquial
- 29-Células de Sertoli

FIG. 28

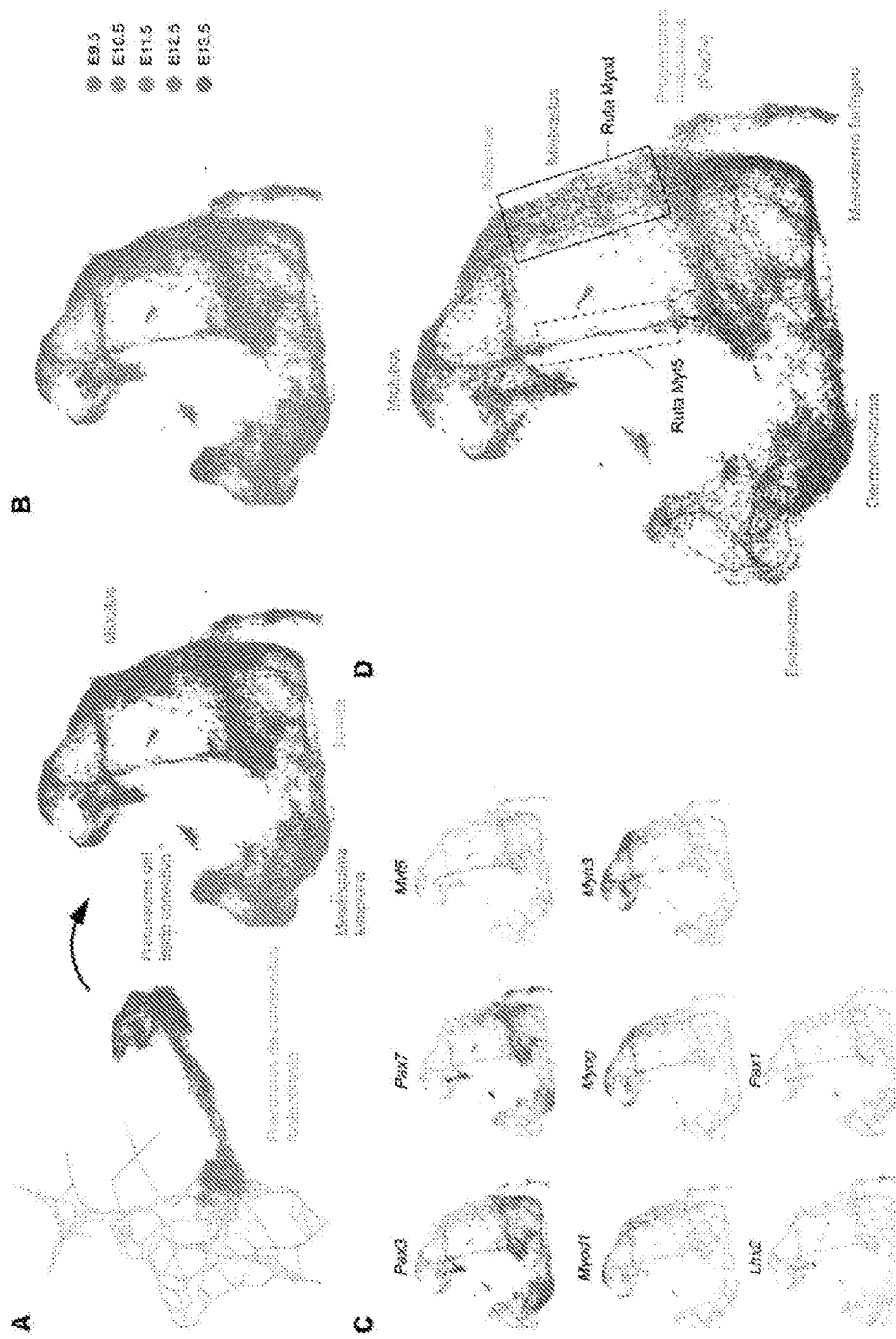


FIG. 29

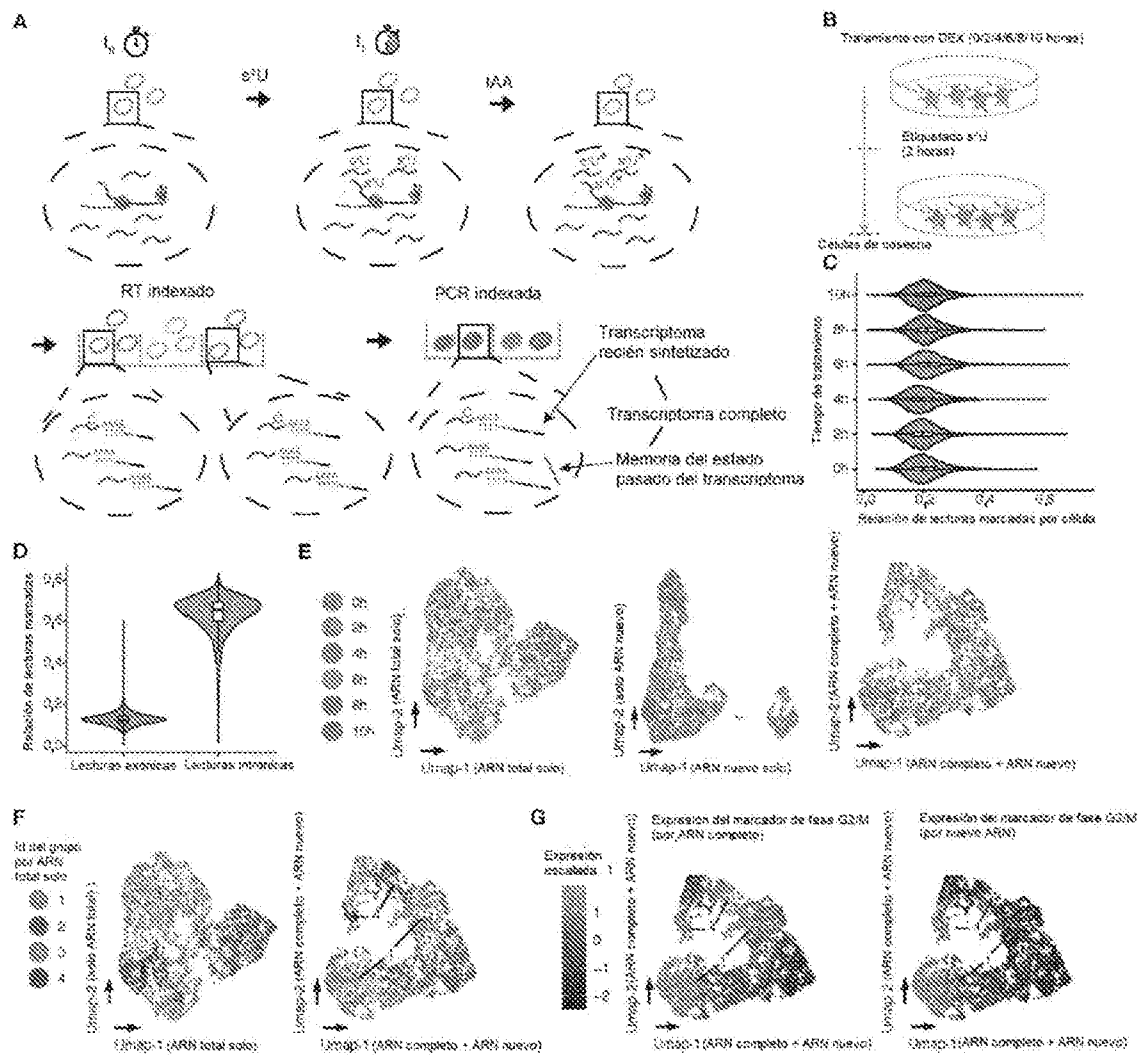


FIG. 30

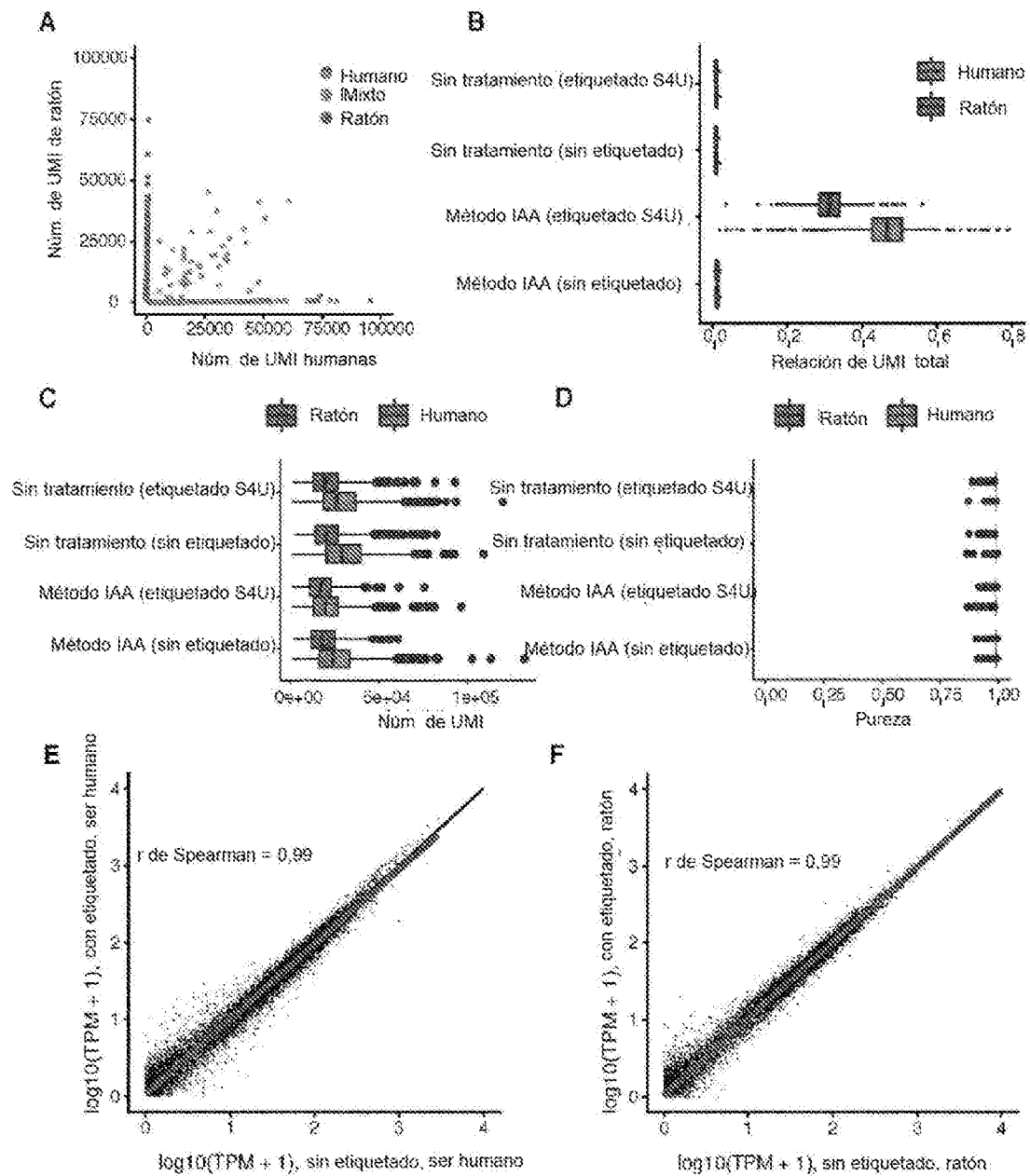


FIG. 31

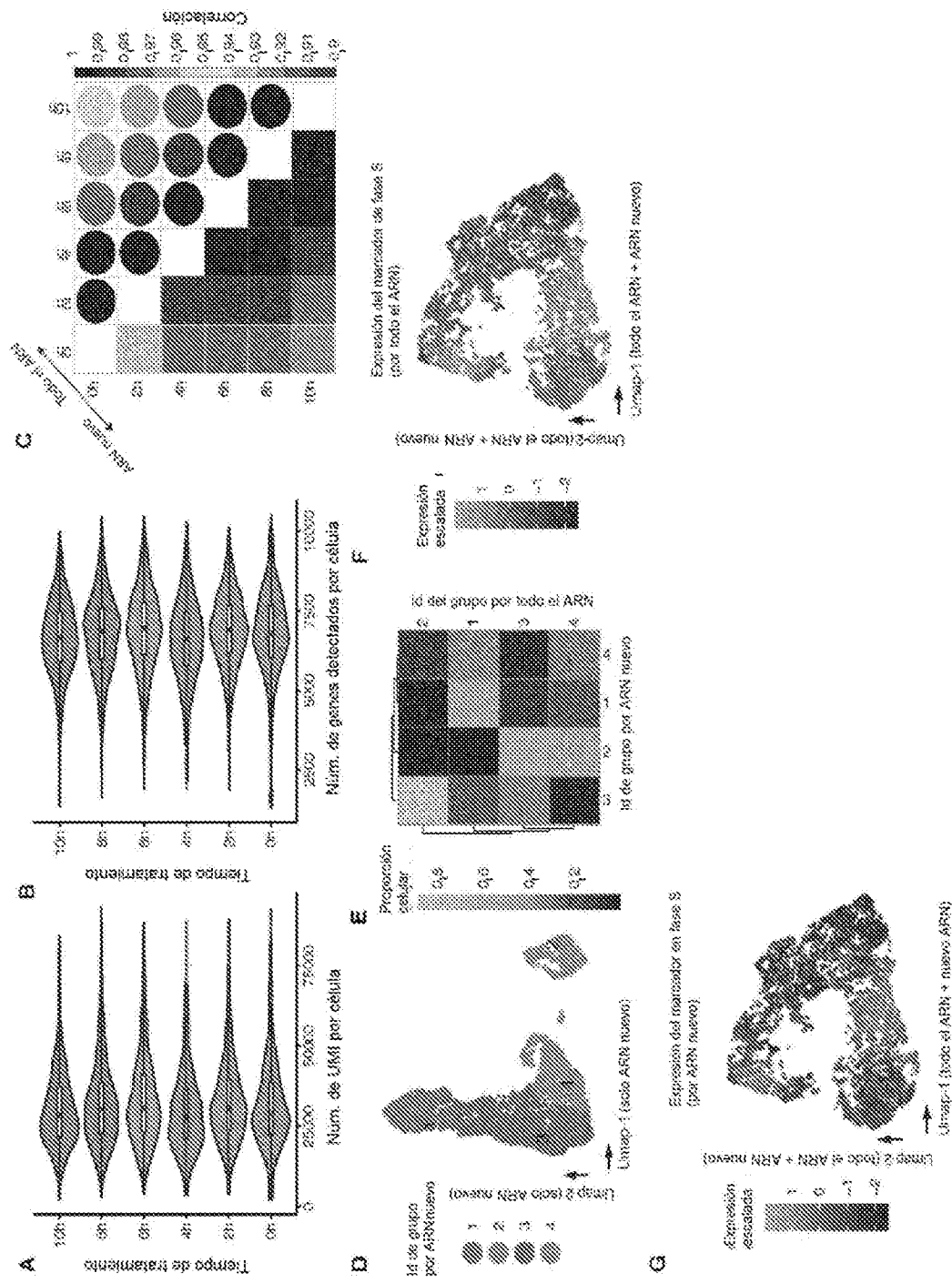


FIG. 32

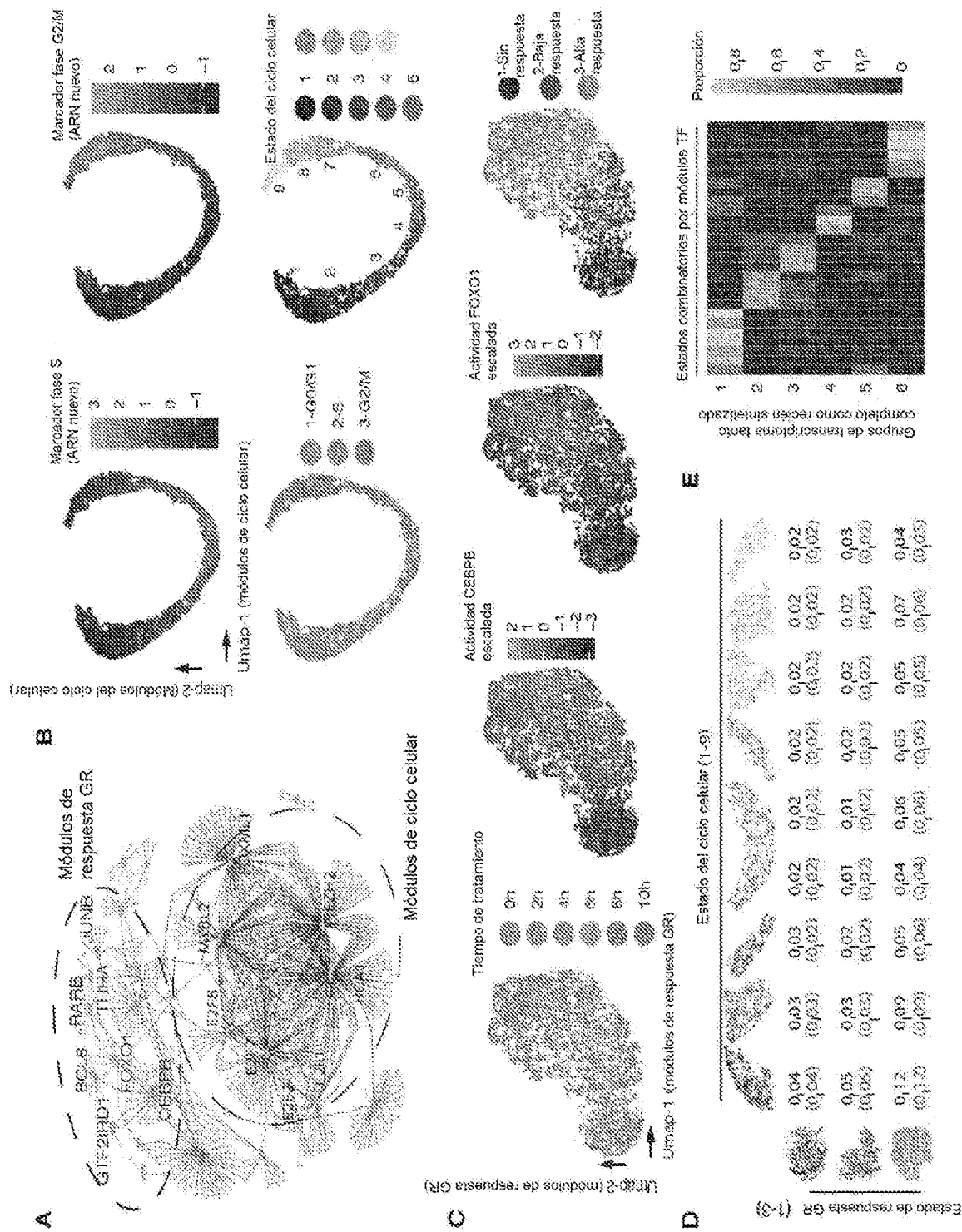


FIG. 33

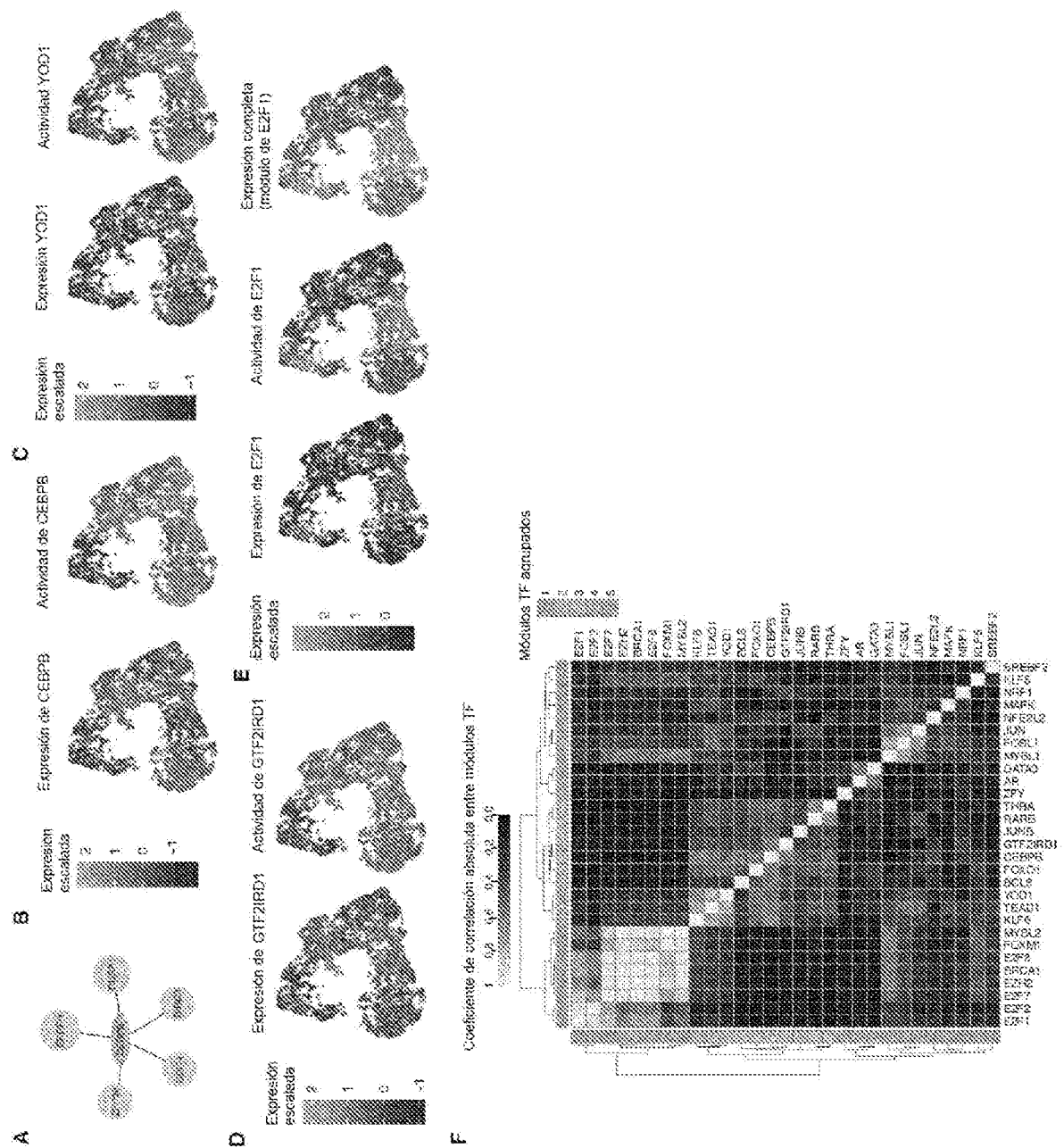
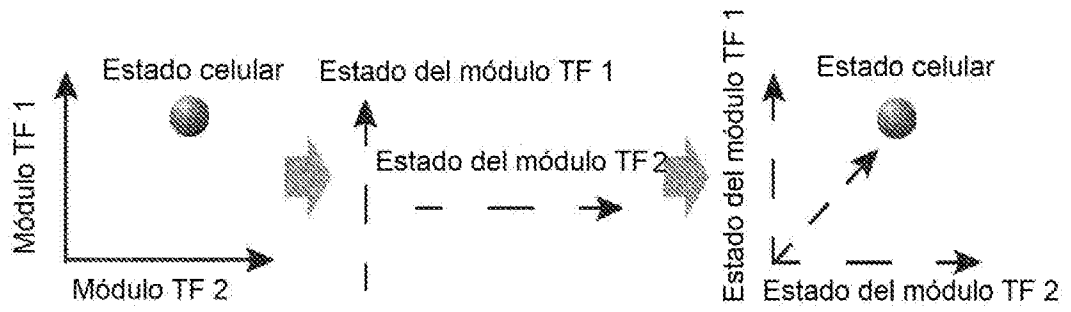


FIG. 34

A



B

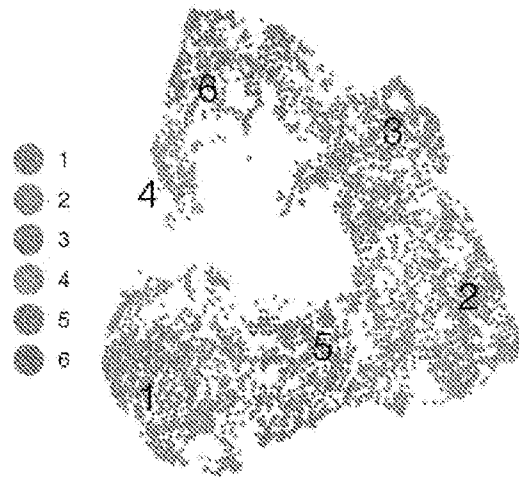


FIG. 35

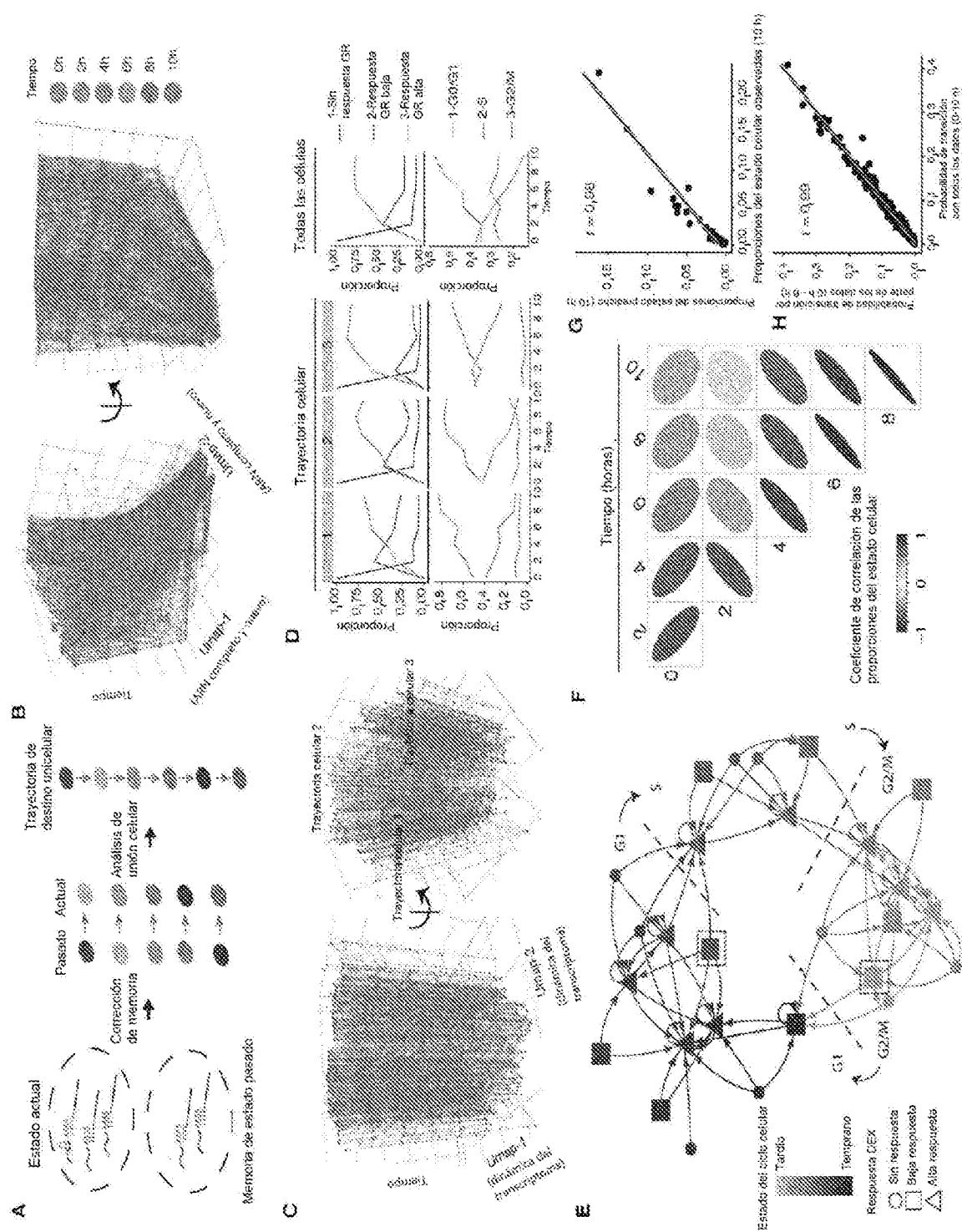


FIG. 36

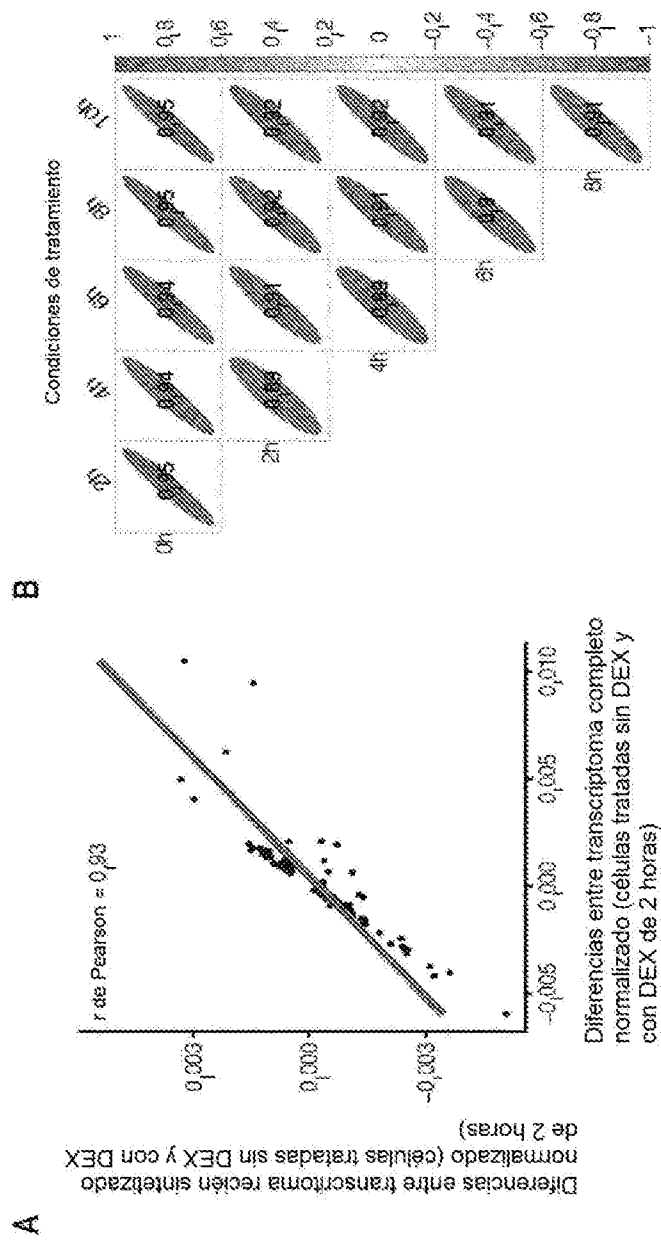


FIG. 37

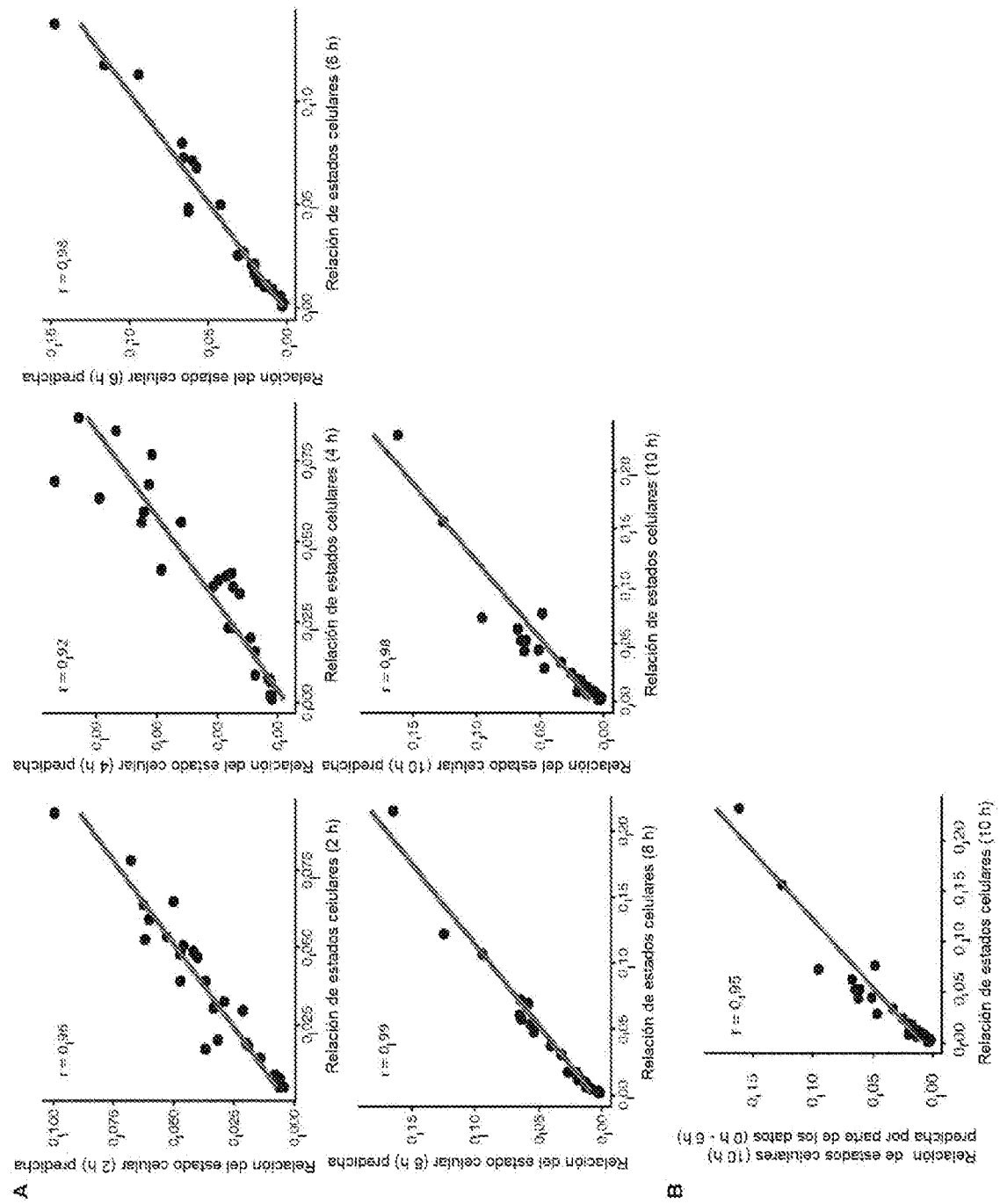


FIG. 38

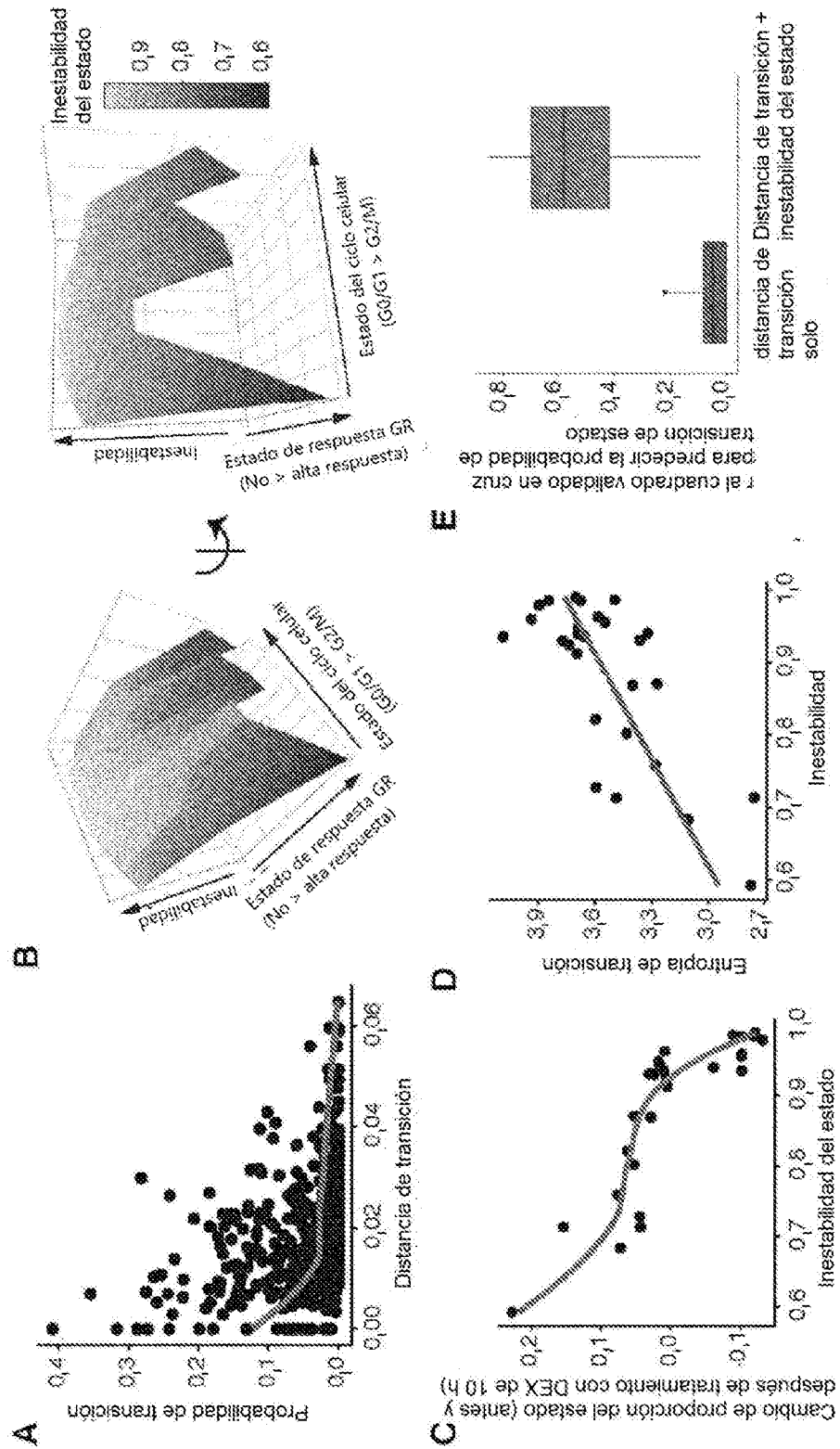


FIG. 39

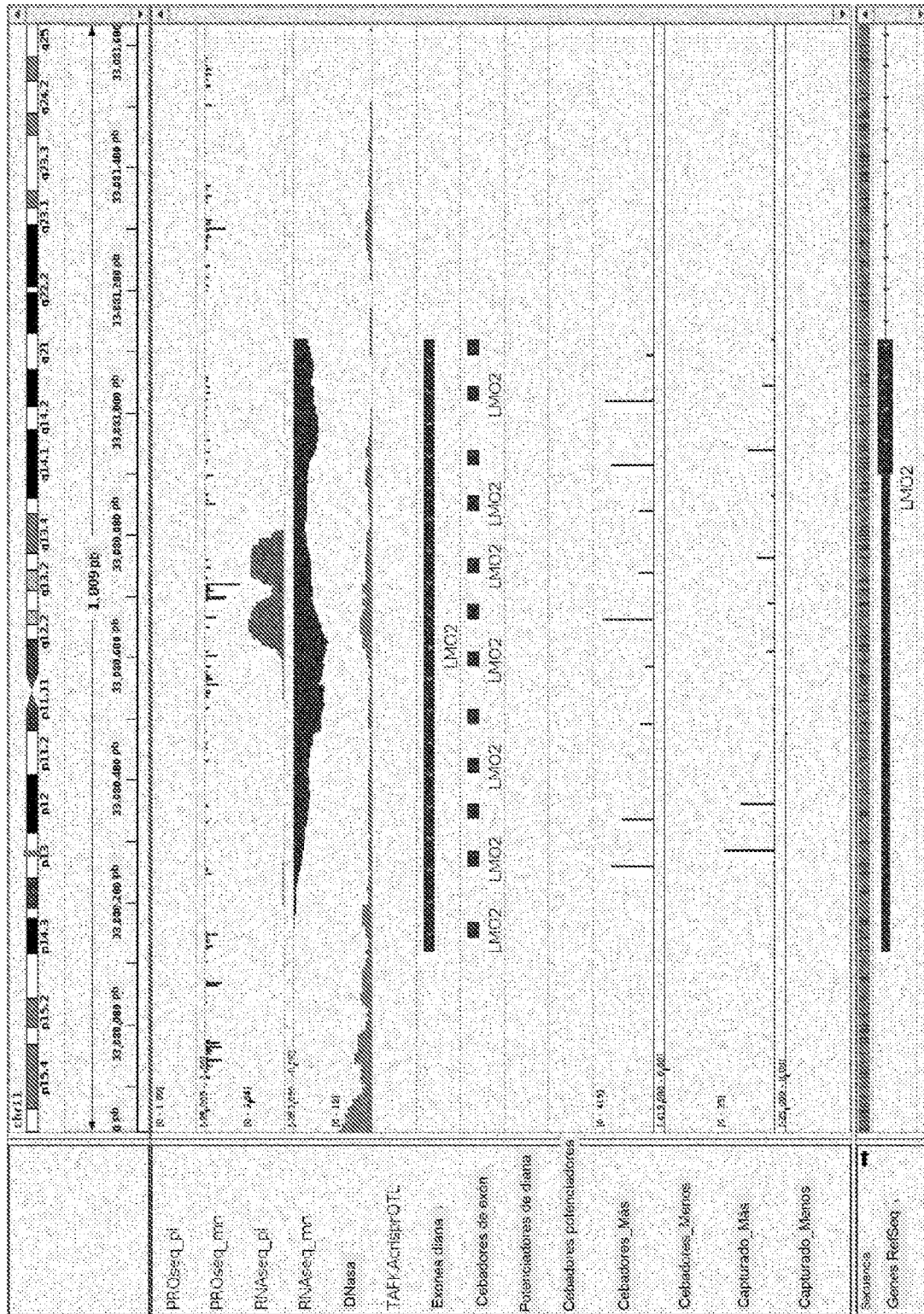


FIG. 40

Cromosoma	Comienzo	Final	Gen	Recuentos totales	Codificación	Nivel	Codificación
chrM	1672	3230	RNR2	4566	653212	1	1
chr11	34073229	34124157	CAPRIN1	2061	2344	2	451
chr11	33880122	33913836	LMO2	1983	376	3	4627
chrM	649	1603	RNR1	1017	156276	4	4
chr11	34127110	34168458	NAT10	641	1075	5	1489
chr11	33724555	33758025	CD59	420	2503	6	396
chr11	34460471	34493607	CAT	395	263	7	5919
chr11	65265232	65273983	MALAT1	291	443474	8	2
chr11	33278217	33378568	HIPK3	238	505	9	3550
chr22	21061978	21213100	P4KA	92	4667	10	120
chr11	34937676	35017675	PDHX	90	146	11	7928
chr9	37779710	37785089	EXOSC3	72	356	12	4827

FIG. 41

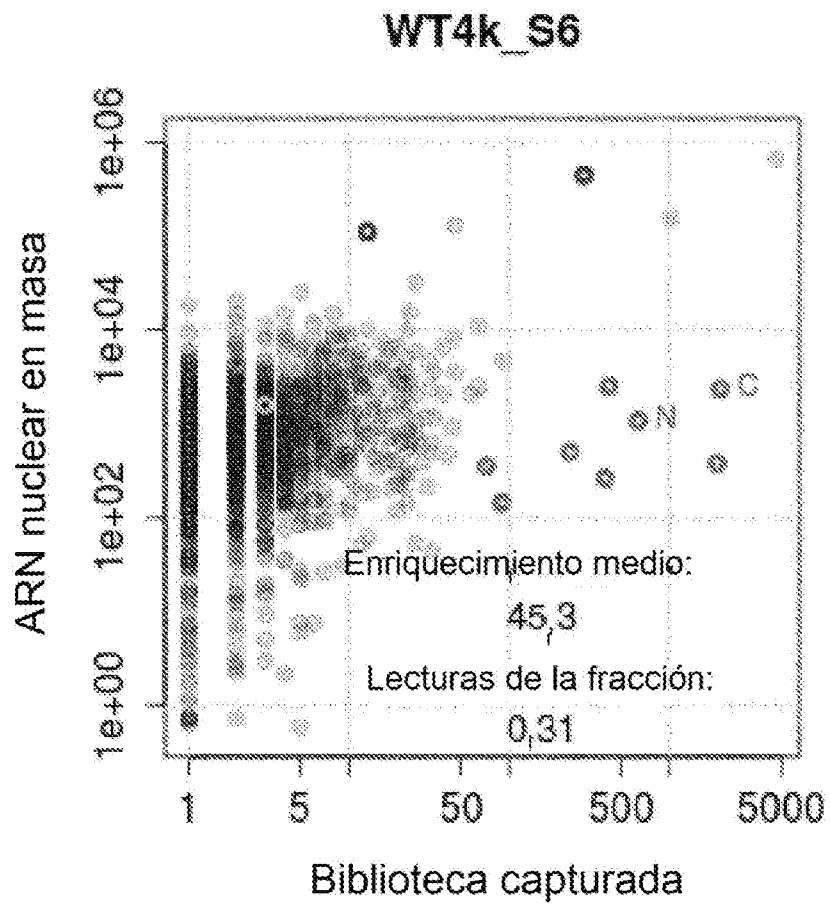


FIG. 42