(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 555 389**

(51) Int. CI.:

C12Q 1/68 (2006.01)
C12N 15/11 (2006.01)
C40B 50/06 (2006.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(96) Fecha de presentación y número de la solicitud europea: **23.03.2010 E 10762102 (1)**

(97) Fecha y número de publicación de la concesión europea: **21.10.2015 EP 2414548**

(54) Título: **Análisis de expresión génica en células individuales**

(30) Prioridad:

30.03.2009 US 164759 P

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
30.12.2015

(73) Titular/es:

ILLUMINA, INC. (100.0%)
5200 Illumina Way
San Diego, CA 92122, US

(72) Inventor/es:

LINNARSSON, STEN

(74) Agente/Representante:

ZEA CHECA, Bernabé

Observaciones :

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

DESCRIPCIÓN

Análisis de expresión génica en células individuales

Campo de la invención

La presente invención se refiere al análisis de la expresión génica en células individuales. En particular, la invención se refiere a un método para preparar una biblioteca de ADNc a partir de una pluralidad de células individuales, y a una biblioteca de ADNc producida por este método. Las bibliotecas de ADNc preparadas por el método de la invención son adecuadas para análisis de la expresión génica por secuenciación.

Antecedentes de la invención

La determinación del contenido de ARNm de una célula o tejido (es decir "perfil de expresión génica") proporciona un método para el análisis funcional de tejidos y órganos normales y enfermos. Por ejemplo, puede usarse el perfil de expresión génica en el estudio de embriogénesis; para la caracterización de muestras tumorales primarias; para el análisis de biopsias de tejido enfermo y normal, por ejemplo, en psoriasis; para el análisis comparativo de tipos celulares de especies diferentes para delinear la evolución del desarrollo; como un sistema de ensayo para diagnóstico; como un sistema de control de calidad en terapia de reemplazo celular (es decir para asegurar que un cultivo de células sea suficientemente puro, y las células se diferencien correctamente); y como una herramienta *in vitro* para medir el efecto de un gen transfectado o ARNip en dianas cadena abajo a pesar de una eficacia de transfección menor del 100 %.

El perfil de expresión génica se realiza habitualmente aislando ARNm de muestras tisulares y sometiendo este ARNm a hibridación de micromatriz. Sin embargo, dichos métodos permiten solamente analizar genes previamente conocidos, y no pueden usarse para analizar corte y empalme alternativo, promotores y señales de poliadenilación.

Por lo tanto, la secuenciación directa de todo, o partes, del contenido de ARNm de un tejido se está usando crecientemente (Cloonan *et al*., Nat Methods 5(7): 613-9 (2008)). Sin embargo, los métodos actuales de analizar el contenido de ARNm de células por secuenciación directa se basan en el análisis de ARNm en masa obtenido de muestras tisulares que típicamente contienen millones de células. Esto significa que mucha de la información funcional presente en células individuales se pierde o se distorsiona cuando la expresión génica se analiza en ARNm en masa. Además, no pueden observarse procesos dinámicos, tales como el ciclo celular, en promedios de población. De forma similar, solamente pueden estudiarse tipos celulares distintos en un tejido complejo (por ejemplo, el cerebro) si las células se analizan individualmente.

La expresión génica en células individuales se ha analizado previamente usando una diversidad de métodos (véase, por ejemplo, Brail *et al*., Mutat Res 406(2-4): 45-54 (1999); Levsky *et al*., Science 297(5582): 836-40 (2002); Bengtsson *et al*. Cenóme Res 15(10): 1388-92 (2005); Esumi *et al*., Nat Genet 37(2): 171-6 (2005)). En particular, la expresión génica de células individuales en células neurales se ha estudiado por análisis de micromatrices (véase Esumi *et al*. Neurosa Res 60(4): 439-51 (2008)). Sin embargo, estos métodos requieren que cada célula individual se analice individualmente y se trate por separado durante el procedimiento completo, que consume tiempo y es caro. Además, la preparación y amplificación de muestras de células individuales introduce potencialmente de forma independiente variación entre células. Además, ya que el ADNc de cada célula debe amplificarse hasta una cantidad que puede manipularse razonablemente para el análisis posterior, hay un desvío de amplificación potencial. Por ejemplo, una célula individual contiene aproximadamente 0,3 pg de ARNm, y se necesitan habitualmente al menos 300 ng para análisis posterior por secuenciación. Por lo tanto, se requiere una amplificación de al menos un millón de veces.

Adicionalmente, las micromatrices tienen dos defectos principales: están ligadas a genes conocidos, y tienen una sensibilidad y rango dinámico limitados. La secuenciación de ARN (ARN-Sec) supera estos problemas secuenciando el ARN directamente (Ozsolak *et al*., Nature 461: 814-818 (2009)) o después de transcripción inversa a ADNc (Cloonan *et al*., Nat. Methods 5: 613-619 (2008); Mortazavi *et al*., Nat. Methods 5: 621-628 (2008); Wang *et al*., Nature 456: 470-476 (2008)). Las lecturas de secuencia se mapean en el genoma para revelar sitios de transcripción y la cuantificación se basa simplemente en recuentos de aciertos, con gran sensibilidad y rango dinámico casi ilimitado.

Los tejidos son pocas veces homogéneos, sin embargo, y por lo tanto cualquier perfil de expresión basado en una muestra tisular, biopsia o cultivo celular confundirá los verdaderos perfiles de expresión de sus células constituyentes. Un modo de evitar este problema sería analizar células individuales en lugar de poblaciones celulares, y de hecho se han desarrollado métodos de células individuales para ambas micromatrices (Esumi *et al*., Neurosci. Res. 60: 439-451 (2008) y Kurimoto *et al*., Nucleic Acids Res. 34: 42 (2006)). Estos métodos son adecuados para el análisis de números pequeños de células individuales, y en particular pueden usarse para estudiar células que son difíciles de obtener en grandes números, tales como oocitos y las células del embrión temprano. Las células pueden aislarse por ejemplo por microdisección de captura de láser o por microcapilaridad, y

pueden usarse genes marcadores para localizar células de interés. Sin embargo, los transcriptomas de una única célula deben enfrentarse a dos grandes retos. En primer lugar, los marcadores adecuados para el aislamiento prospectivo de poblaciones celulares definidas no están disponibles para cada tipo celular, lo que refleja el hecho de que pocos tipos celulares están claramente definidos en términos moleculares. En segundo lugar, las abundancias de transcritos varían en gran medida de célula a célula. Por ejemplo, el contenido de ARNm de β-actina (Actb) varía más de tres órdenes de magnitud entre células de islotes pancreáticos (Bengtsson *et al.*, Genome Res. 15: 1388-1392 (2005)). Se han presentado resultados similares, usando una diversidad de métodos de detección, para ARN polimerasa II (Raj *et al*, PLoS Biol 4: 309 (2006)), GAPDH (Lagunavicius *et al.*, RNA 15: 765-771 (2009) y Warren *et al*, Proc. Nati. Acad. Sci. U.S.A. 103: 17807-17812 (2006)), PU.1 (Warren *et al*, mencionado anteriormente) y ARNm de TBP, B2M, SDHA y EE1FG (Taniguchi *et al*, Nat Methods 6: 503-506 (2009)) y en la actualidad parece ser una característica común del transcriptoma.

La mayoría de la variación puede ser intrínseca, provocada por activación estocástica de tipo estallido de transcripción, cuando breves episodios de síntesis de ARNm que duran varios minutos están separados por periodos de silencio transcripcional de duración similar (Chubb *et al*, Curr Biol 16: 1018-1025 (2006)). Cada estallido daría lugar a una población densa de ARNm en el núcleo, que se exporta después al citoplasma y rápidamente se degrada. Como consecuencia, una muestra aleatoria de células mostraría gran variación en su contenido de ARNm particulares, variando de las células que acaban de experimentar un estallido, a las que han degrado casi completamente su ARNm; esto se ha observado directamente para la transcripción de ARN polimerasa II *in situ* usando una sonda fluorescente que se dirigía a la repetición de 52 copias en ese gen (Raj *et al*, PLoS Biol 4: 309 (2006)).

En resumen, con frecuencia no hay marcadores de superficie celular adecuados para usar en el aislamiento de células individuales para su estudio, incluso cuando los hay, un número pequeño de células individuales no es suficiente para capturar el intervalo de variación natural en la expresión génica. La presente invención intenta superar, o reducir, estos problemas proporcionando un método para preparar bibliotecas de ADNc que pueden usarse para analizar la expresión génica en una pluralidad de células individuales.

Sumario de la invención

La invención proporciona un método para preparar una biblioteca de ADNc a partir de una pluralidad de células individuales. En un aspecto, el método incluye las etapas de liberar ARNm de cada célula individual para proporcionar una pluralidad de muestras de ARNm individuales, sintetizar una primera cadena de ADNc a partir del ARNm en cada muestra de ARNm individual e incorporar un marcador en el ADNc para proporcionar una pluralidad de muestras de ADNc marcadas, agrupar las muestras de ADNc marcadas y amplificar las muestras de ADNc agrupadas para generar una biblioteca de ADNc que tenga ADNc bicatenario. La invención también proporciona una biblioteca de ADNc producida por los métodos descritos en el presente documento. La invención proporciona además métodos para analizar la expresión génica en una pluralidad de células preparando una biblioteca de ADNc como se describe en el presente documento y secuenciando la biblioteca.

Breve descripción de los dibujos

Se pretende que las figuras ilustren conceptos amplios de la invención por referencia a ejemplos representativos para facilidad de análisis. No se pretende que limiten el alcance de la invención mostrando una de varias realizaciones alternativas o mostrando u omitiendo características opcionales de la invención.

La **Figura 1**, paneles A-F, muestra una visión de conjunto de un método para analizar la expresión génica en una pluralidad de células individuales. (A) El tejido de interés se disecciona; (B) se selecciona una pluralidad de células individuales; (C) se colocan células individuales en pocillos separados de una placa de 96 pocillos y se lisan, se realiza transcripción inversa marcada en cada muestra para producir ADNc; (D) se agrupan y amplifican muestras de ADNc; (E) se realiza secuenciación para obtener 100 millones de lecturas; y (F) identificación de genes expresados e identificación de células de las que se originaron.
La **Figura 2** muestra la síntesis de ADNc por cambio de molde.
La **Figura 3** muestra un ejemplo de oligonucleótido de cambio de molde que comprende una secuencia de cebador de amplificación 5' (APS), un marcador celular y una secuencia 3' para cambio de molde.
La **Figura 4** muestra un ejemplo de un cebador de síntesis de ADNc (CDS) que comprende una secuencia de cebador de amplificación 5' (APS), un marcador celular y una secuencia complementaria de ARN 3' (RCS).
La **Figura 5** muestra la visualización de muestras de ADNc L001 y L002 después de la amplificación de ADNc de longitud completa por PCR. Carril 1: escalera de marcadores de 100 pb; carriles 2-3: 25 ciclos; carriles 4-5: 30 ciclos; carriles 6-7: 35 ciclos. Los carriles pares contienen la muestra L001 y los carriles impares contienen la muestra L002.
La **Figura 6** muestra una serie de diluciones de una PCR de ensayo usando la muestra L001 (carriles 3-10). Los carriles 2 y 11 contienen una escalera de 100 pb como un marcador de tamaño.
La **Figura 7**, paneles A y B, muestra la separación por electroforesis en gel y el aislamiento de bibliotecas de ADNc. El panel A muestra una biblioteca de ADNc después de la amplificación final por PCR (16 ciclos) (carriles 5 y 6). El panel B muestra que la región de 125-200 pb se ha escindido. Los carriles 3 y 8 contienen una escalera de 100 pb

como un marcador de tamaño.

La **Figura 8** muestra un ejemplo de una molécula de ADNc secuenciada a partir de una biblioteca de ADNc marcada. Las secuencias de cebadores para secuenciación por SOLiD (P1 y P2) están subrayadas. El marcador específico de célula está encuadrado. Las 2-5 G del mecanismo de cambio de molde están sombreadas en una caja gris. La secuencia del vector de clonación TOPO se muestra en cursiva. El inserto en este caso es tubulina beta 2c.

La **Figura 9** muestra una representación gráfica de resultados de una PCR en tiempo real cuantitativa que compara el ADNc en masa (eje horizontal) frente a ADNc marcado de 96 células (eje vertical). Cada círculo representa un par de cebadores de PCR dirigido contra los genes indicados. Las unidades son arbitrarias y derivan del valor de ciclos hasta umbral, C_T.

La **Figura 10**, paneles A-E, muestra una visión en conjunto del método de transcripción inversa marcado con células individuales (STRT) y resultados ejemplares. (A) Visión de conjunto del método, que ilustra como se siguieron las células individuales. Se incorporaron códigos de barras específicos de pocillo (y por tanto específicos de célula) durante la síntesis de ADNc, dando como resultado una biblioteca en la que cada molécula portaba un código de barras que identificaba la célula de origen. (B) Ejemplo de lecturas mapeadas en ambas cadenas del locus de Pou5f1 de 5 kb, mostrado como una representación de cobertura. Las lecturas fueron específicas de cadena y se mapearon principalmente en exones. El fondo no de ARNm fue mínimo como se valoró por aciertos en intrones en la cadena directa (carril superior) o la cadena inversa (carril inferior). (C) Comparación con ARN-Sec convencional (Cloonan *et al.* Nat. Methods 5: 613-619 (2008)) para el locus Nanog de 6 kb. Ya que la síntesis de ADNc se cebó desde la cola de poli(A), las lecturas se agruparon típicamente en el extremo 3' (carril superior). La UTR 3' extendida de Nanog (azul claro) se detectó claramente por ambos métodos. Las lecturas de SQRL se extendieron a exones 5', pero mostraron más fondo intrónico. (D) El genoma mitocondrial. Como se esperaba, la transcripción se detectó casi exclusivamente en la cadena H (parte superior), originándose solamente algunos transcritos truncados (flecha) en la cadena L. Los genes codificantes de proteínas se indican en la parte inferior. (E) Expresión génica en el cromosoma 19. La cobertura en las cadenas directa (carril superior) e inversa (medio) estaba altamente correlacionada con la densidad génica local (inferior). Los genes se muestran como barras apiladas horizontales. Para mayor claridad, las escalas verticales en (B-E) están truncadas a la mitad de los valores máximos.

La **Figura 11**, paneles A-F, muestra representación gráfica de etapas usadas opcionalmente en el método de STRT. El panel A muestra una etapa de trascripción inversa. Un cebador de oligo dT con cola dirige la síntesis de una cadena de ADNc. Cuando se alcanza el final del ARN molde, la transcriptasa inversa añade 3-4 C en 3' de la cadena de ADNc (debido a su actividad transferasa terminal). El panel B muestra una etapa de cambio de molde. El oligo auxiliar con código de barras hibrida de forma transitoria, y la síntesis de ADNc continúa usando el oligo como molde. Como consecuencia, el extremo 3' del ADNc portará un código de barras (XXXX), secuencias de reconocimiento de BtsCI (Bts) y secuencias de hibridación de cebadores (Pr). El panel C muestra una única etapa de PCR de cebadores. Los extremos del ADNc tienen secuencias idénticas y se amplifican usando un único cebador de PCR, que ayuda a suprimir amplicones cortos. El panel D muestra una etapa de fragmentación. La biblioteca amplificada se fragmenta a 200 – 300 pb usando digestión con DNasa controlada. El panel E muestra una inmovilización y etapa de reparación de extremos. Los fragmentos tanto 5' (con código de barras (Br) y sitio BtsCI (Bts)) como 3' se unen con perlas, mientras que los fragmentos internos se retiran por lavado. El panel F muestra liberación de fragmentos y ligamiento con el adaptador. Se liberan fragmentos 5' por digestión con BtsCI, dejando solamente el código de barras (Bc) y el inserto (área blanca). Los fragmentos 3' permanecen unidos a las perlas. Se ligan los adaptadores compatibles con extremos apareados de Genome Analyzer (P1 y P2). La biblioteca se secuencia desde el cebador P1 (y podría opcionalmente secuenciarse también desde P2). Las lecturas de P1 comienzan con un código de barras de 5 pb, seguido de 3-4 G, seguido del inserto de ADNc. Las lecturas de P2 producirían solamente secuencia de ADNc.

La **Figura 12,** paneles A-D, muestra una representación gráfica de una ausencia de motivos que rodean el sitio de cambio de molde. Todas las lecturas se examinaron en la muestra L006 con respecto a la presencia de cualquier motivo alrededor del sitio de cambio de molde (es decir, alrededor del extremo 5' de cada lectura). Se muestran los logos de secuencia para las 20 bases de secuencia genómica cadena arriba y cadena abajo del primer nucleótido (flecha) de cada lectura mapeada. Como se ejemplifican en A y B, en casos típicos (92 de 96), no se detectó ningún motivo fuerte, lo que indica una ausencia de acontecimientos de cebado erróneo significativos, que habrían generado un motivo cadena arriba complementario al cebador. En cuatro casos (C y D), hubo una preferencia general de secuencias ricas en T particularmente en las primeras 20 bases de la lectura. Esto se produjo en pocillos con números muy pequeños de lecturas, lo que indica una reacción fallida. Sin embargo, en un caso individual el motivo rico en T se observó a pesar de un gran número de lecturas.

La **Figura 13**, paneles A y B, muestra una representación gráfica de puntos calientes para cambio de molde. (A) El locus de Actb se expresa a partir de la cadena inferior (inversa), de derecha a izquierda en la figura. Las dos pistas superiores muestran aciertos de agregados en las cadenas directa e inversa, respectivamente, lo que demuestra especificidad de cadena y falta de fondo en intrones. La pista media (azul) muestra la estructura de exones/intrones del gen. La pista inferior muestra aciertos individuales de células individuales. Cada fila de píxeles representa los aciertos de una célula individual como puntos negros. Hay 96 filas de píxeles en total. (B) Se realizó el mismo análisis para Sox2, un gen de un único exón transcrito en la cadena directa (superior), que muestra el desvío 3' habitual. Tanto en (A) como en (B), los puntos calientes eran claramente visibles, y se compartían entre células, lo que sugiere que representan sitios estructurales en ARNm que favorecen la terminación de la síntesis de ADNc, hidrólisis de ARN y/o cambio de molde.

La **Figura 14**, paneles A y B, muestra una representación gráfica de la nueva tasa de descubrimiento. (A) Muestra la

tasa de descubrimiento de lecturas de mapeo distintas en función del número total de lecturas. Ninguna de las muestras se secuenció hasta saturación, y la mayoría de las bibliotecas contendrían habitualmente al menos 3 millones de moléculas distintas, lo que indica que en promedio al menos 30.000 moléculas distintas por célula se convirtieron con éxito en ADNc amplificable. Las curvas son onduladas debido a heterogeneidades en los datos, supuestamente causadas por imperfecciones en el proceso de PCR de grupos que pueden generar duplicados locales y por lo tanto muestreo no aleatorio. En (B) la tasa de descubrimiento de características anotadas distintas se muestra en función del número de lecturas mapeadas (para la muestra L006). La saturación se alcanzó rápidamente, lo que muestra que la mayoría de las características presentes en la muestra podrían descubrirse a profundidad de toma de muestras modesta.

La **Figura 15** muestra una representación gráfica para distinguir la expresión de genes solapantes. Debido a la especificidad de cadena del mecanismo de cambio de molde, el número de cadenas podría mantenerse durante todo el protocolo. Esto fue especialmente importante para genes con exones solapantes. Se representa en la figura un ejemplo de dicho par de genes, Catepsina A (Ctsa) y proteína de transferencia de Fosfolípidos (Pltp), cuyos últimos exones se solapan. Sin información de cadena, las lecturas en los últimos cuatro exones de Pltp no podrían distinguirse de las lecturas que se originan en el último exón de Ctsa. Hay aproximadamente 3.000 genes con exones 3' solapantes similares.

La **Figura 16** muestra una representación gráfica de desvío de longitud para transcritos. Para detectar cualquier desvío contra transcritos cortos o largos, se calculó el nivel de expresión promedio en función de la longitud de ARNm (en grupos de 200 pb) para la muestra L019. Cada barra muestra el nivel de expresión de genes con transcritos más cortos que la longitud indicada (por lo tanto la primera barra contiene transcritos de 0 - 200 pb de longitud). Sobre una amplia serie de longitudes de ARNm, no hubo ninguna diferencia evidente en los niveles de expresión medidos. Los transcritos más cortos (<200 pb) se suprimían supuestamente por la etapa de purificación en gel en la que se seleccionaron insertos de > 100 pb. La sobreexpresión aparente de genes en el intervalo de 400 - 800 pb puede explicarse posiblemente por un cambio de molde más eficaz en este intervalo, en el que la síntesis de ADNc con frecuencia alcanzaría el extremo 5' de ARNm. Como alternativa, puede deberse simplemente a la presencia de algunos genes muy altamente expresados en este intervalo, incluyendo Dppa5 y Rps14.

La **Figura 17**, paneles A-E, muestra una representación gráfica de la precisión cuantitativa del método de STRT. (A) La distribución de los niveles de expresión génica en transcritos por millón (t.p.m.) mostraron predominantemente expresión baja, en el intervalo de 10 - 100 t.p.m. (B) La comparación de las secuencias de dos células hasta una profundidad de aproximadamente 500.000 lecturas/célula. En este caso, los genes por debajo de 100 t.p.m. podrían cuantificarse con precisión. (C) Comparación de dos células secuenciadas hasta aproximadamente 100.000 lecturas/célula. En este caso, la sensibilidad descendió hasta aproximadamente 1000 t.p.m. (D) Probabilidad de detección en función del nivel de expresión. Cada punto muestra un gen, con un nivel de expresión promedio dado (entre todas las células) y fracción de células que tienen expresión no cero de este gen. La distribución aborda el límite teórico de la toma de muestras aleatoria dada la profundidad real de secuenciación usada aquí (línea discontinua). (E) Comparación con PCR en tiempo real cuantitativa para las cuatro células ES mostradas en (B) y (C), usando marcadores seleccionados de pluripotencialidad y diferenciación. En general, la precisión cuantitativa fue buena, con un único falso positivo potencial (Eomes en Célula ES nº 4). Sin embargo este fue un acontecimiento poco común, y no se observó para este gen en ninguna de las otras 160 células ES examinadas. Se convirtieron los niveles de Q-PCR a t.p.m. por normalización del par de cebadores Actb/1081 con las células ES nº 1 y nº 2. Después se midió independientemente la actina usando un par de cebadores diferente (Actb/1832) para confirmar la precisión de la normalización.

La **Figura 18** muestra una comparación gráfica entre STRT, Q-PCR y análisis de micromatrices. Los genes que se esperaba que se expresaran (Actb, Pou5f1, Zfp42, Sox2, Klf4, Nanog, Plk1, Zic3) o que no se expresaran (Gata4, Brachyury, Eomes, Otx1, Cdx2, Gata5, Calb1, Gfap, Dppa3 y NeuroD1) en células ES indiferenciadas se analizaron por STRT, PCR en tiempo real cuantitativa (Q-PCR) y micromatriz Illumina. Hubo una buena correlación entre STRT y Q-PCR, y en menor grado con datos de micromatrices. En particular, Sox2 apareció poco en la micromatriz, mientras que Otx1 y Dppa3 fueron falsos positivos aparentes. Los datos de micromatrices son la medida de dos reacciones de hibridación, se realización Q-PCR por duplicado y se repitió una vez para confirmación, y los datos de STRT son la media de 160 células ES individuales.

La **Figura 19**, paneles A y B, muestra una representación gráfica de la distribución de expresión génica entre células. (A) muestra la distribución de los niveles de expresión de Actb entre todas las células usando STRT. (B) muestra la expresión de Actb en células de islotes pancreáticos de Bengtsson *et al*. Genome Res 15(10): 1388-92 (2005) para comparación.

La **Figura 20** muestra un análisis de componentes principales. Para descubrir y agrupar tipos celulares basándose en datos de expresión, las cinco muestras de 96 células preparadas independientemente se sometieron a análisis de componentes principales. Los tres tipos de células (ES, Neuro-2A y MEF) claramente se agruparon por separado, aunque los MEF no formaron un grupo muy definido. Además, las células ES preparadas de forma independiente se agruparon entre sí, lo que muestra que la PCA no captó simplemente diferencias en la preparación de muestras. Esto demuestra que los datos de expresión de células individuales pueden usarse para clasificar con precisión tipos celulares.

La **Figura 21**, paneles A-C, muestra una visualización basada en gráficos ("mapeo celular") del patrón de expresión. (A) Las células, representadas por nodos gráficos (círculos) se extendieron aleatoriamente, y se dibujaron los bordes a partir de cada célula a las otras cinco células con las que estaba más altamente correlacionada. (B) Se usó la distribución dirigida por fuerza para trazar la gráfica en un plano. En esta distribución, las células se repelían entre sí

uniformemente, pero se mantuvieron juntas por los bordes que actuaban como resortes elásticos. El mapa visual resultante fue coherente con las identidades celulares conocidas (células ES frente a células Neuro-2A), mostrando algunas células menos profundamente secuenciadas escasa separación. (C) La adición de más células ES así como fibroblastos (MEF) expandió el mapa y demostró que las células ES preparadas de forma independiente se agrupaban juntas con precisión.

La **Figura 22** muestra la visualización de la expresión génica en un mapa celular de la **Figura 21**. Cada mapa conserva su distribución de la **Figura 21C**, pero las células se sombrean de acuerdo con la expresión del gen indicado. Se usó una escala logarítmica (superior derecha). El ARN 2 de ribosoma mitocondrial (mt_Rnr2) fue el gen más expresado de todos. Se detectaron genes constitutivos tales como Actina (Actb) y la proteína ribosómica L4 (Rp14) en todos los tipos celulares, pero no en cada célula individual. La potencia del perfil de expresión de células individuales al azar se reveló para genes poco expresados como K-ras (Kras), que se detectó solamente en aproximadamente la mitad de las células, pero aún se expresaba claramente en todos los tipos celulares. La calbindina (Calb1) estaba ausente, como se esperaba y se confirmó por Q-PCR. Un conjunto de marcadores celulares de ES bien conocidos (Dppa5, Sox2, Sa114, Pou5f1, Nanog, Zfp42, Zic3 y Esrrb) estaban claramente expresados específicamente en el grupo de células ES, mientras que Klf4, Myc y Klf2 estaban más ampliamente distribuidos. Dppa3 no se detectó, como se confirmó por Q-PCR (**Figura 18**).

Descripción detallada de la invención

La presente invención proporciona métodos y composiciones para el análisis de la expresión génica en células individuales o en una pluralidad de células individuales. En particular, la invención proporciona métodos para preparar una biblioteca de ADNc a partir de una pluralidad de células individuales. Los métodos se basan en la determinación de los niveles de expresión génica a partir de una población de células individuales, que pueden usarse para identificar variaciones naturales en la expresión génica en un nivel célula a célula. Los métodos también pueden usarse para identificar y caracterizar la composición celular de una población de células en ausencia de marcadores de superficie celular adecuados. Los métodos descritos en el presente documento también proporcionan la ventaja de generar una biblioteca de ADNc representativa de contenido de ARN en una población celular usando células individuales, mientras que bibliotecas de ADNc preparadas por métodos clásicos típicamente requieren ARN total aislado de una población grande (véase Ejemplo 1). Por lo tanto, una población de ADNc producida usando los métodos de la invención proporciona al menos representación equivalente del contenido de ARN en una población de células utilizando una subpoblación menor de células individuales junto con ventajas adicionales como se describe en el presente documento.

Las realizaciones también proporcionan toma de muestras de un gran número de células individuales. Usando similitud de patrones de expresión, puede construirse un mapa de células que muestre cómo las células se relacionan entre sí. Este mapa puede usarse para distinguir tipos celulares por ordenador, detectando grupos de células estrechamente relacionadas (véase Ejemplo II). Por la toma de muestras pueden usarse similitud de patentes de expresión de no solamente algunas, sino de grandes números de células individuales, para construir un mapa de células y cómo se relacionan entre sí. Este método permite acceso a datos de expresión puros de cada tipo de célula distinto presente en una población, sin la necesidad de purificación previa de esos tipos celulares. Además, cuando están disponibles marcadores conocidos, estos pueden usarse por ordenador para delinear células de interés. La validez de este enfoque se muestra en el Ejemplo II, que analiza una colección de células mostradas de tres tipos celulares distintos (células madre embrionarias de ratón, fibroblastos embrionarios y células de neuroblastoma) de distintos orígenes embrionarios (células madre pluripotenciales frente a capas germinales mesodérmicas y ectodérmicas) y patología (normal frente a transformada).

Las realizaciones de la invención proporcionan un método para preparar una biblioteca de ADNc a partir de una pluralidad de células individuales liberando ARNm de cada célula individual para proporcionar una pluralidad de muestras individuales, en el que ARNm en cada muestra de ARNm individual es de una célula individual, sintetizando una primera cadena de ADNc a partir del ARNm en cada muestra de ARNm individual e incorporando un marcador en el ADNc para proporcionar una pluralidad de muestras de ADNc marcadas, en el que el ADNc en cada muestra de ADNc marcada es complementario de ARNm de una célula individual agrupando las muestras de ADNc marcadas y amplificando las muestras de ADNc agrupadas para generar una biblioteca de ADNc que comprende ADNc bicatenario. Utilizando el método anterior, es factible preparar muestras para secuenciar a partir de varios cientos de células individuales en un tiempo corto y con una cantidad mínima de trabajo. Los métodos tradicionales para preparar una biblioteca de fragmentos de ARN para secuenciar incluyen etapas de escisión en gel que son trabajosas. En ausencia de equipamiento especial, no es conveniente preparar más de unas cuantas muestras en paralelo. En algunos aspectos de los métodos descritos en el presente documento, se prepara un conjunto de 96 células como una única muestra (después de síntesis de ADNc), lo que hace factible preparar varios cientos de células para secuenciación. Adicionalmente, la variación técnica se minimiza porque cada conjunto de 96 células se prepara junto (en único tubo).

En algunos aspectos de la invención, cada muestra de ADNc obtenida de una única célula se marca, lo que permite analizar la expresión génica al nivel de una única célula. Esto permite estudiar los procesos dinámicos, tales como el ciclo celular, y analizar distintos tipos celulares en un tejido complejo (por ejemplo el cerebro). En algunos aspectos,

las muestras de ADNc pueden agruparse antes de su análisis. La agrupación de las muestras simplifica la manipulación de las muestras de cada célula individual y reduce el tiempo requerido para analizar la expresión génica en las células individuales, lo que permite un análisis de alto rendimiento de la expresión génica. El agrupamiento de las muestras de ADNc antes de su amplificación también proporciona la ventaja de que prácticamente se elimina la variación técnica entre muestras. Además, como las muestras de ADNc se agrupan antes de la amplificación, se requiere menos amplificación para generar suficientes cantidades de ADNc para un análisis posterior en comparación con la amplificación y el tratamiento de muestras de ADNc de cada célula individual por separado. Esto reduce el desvío de amplificación, y también significa que cualquier desvío será similar entre todas las células usadas para proporcionar muestras de ADNc agrupadas. Tampoco se requiere purificación de ARN, almacenamiento y manipulación, lo que ayuda a eliminar problemas provocados por la naturaleza inestable del ARN.

Como las bibliotecas de ADNc producidas por el método de la invención son adecuadas para análisis de los perfiles de expresión génica de células individuales por secuenciación directa, es posible usar estas bibliotecas para estudiar la expresión de genes que no se conocían previamente, y también analizar el corte y empalme alternativo, promotores y señales de poliadenilación. La preparación de las bibliotecas de ADNc como se describe en el presente documento, proporciona un método sensible para detectar un transcrito de ARN individual o de bajo número de copias. La sensibilidad del método se muestra en la **Figura 17D** y se describe en el Ejemplo II. Por ejemplo, se detectan genes expresados a 100 transcritos por millón (t.p.m.) aproximadamente el 50 % de las veces. Sin embargo, como se muestra en la **Figura 14A**, las muestras no se saturaron, de modo que hay sensibilidad adicional que puede conseguirse con secuenciación más profunda de las muestras. En consecuencia, el método para preparar las bibliotecas de ADNc como se describe en el presente documento detecta un transcrito de ARN individual o de bajo número de copias al menos el 30 % de las veces, como alternativa al menos el 40 % de las veces, al menos el 50 % de las veces, como alternativa al menos el 60 % de las veces, como alternativa al menos el 70 % de las veces, como alternativa al menos el 80 % de las veces, como alternativa al menos el 90 % de las veces o como alternativa al menos el 95 % de las veces.

Las realizaciones también proporcionan un método para identificar un tipo celular individual de una muestra y/o determinar el transcriptoma de una célula individual preparando una biblioteca de ADNc como se describe en el presente documento, determinar los niveles de expresión de células individuales en una población, y mapear las células individuales basándose en la similitud de los patrones de expresión. El mapeo de células individuales puede realizarse por ordenador por un experto en la materia y en particular utilizando los métodos descritos en el presente documento, tal como se muestra en el Ejemplo II. El número de células necesario para determinar la frecuencia de un tipo celular dado en la pluralidad de células seguirá una distribución binomial. Por ejemplo, pueden tomarse muestras de un número predeterminado de células individuales de modo que se espere detectar al menos diez del tipo deseado. En consecuencia, si la frecuencia del tipo celular en la muestra es del 10 %, será necesario preparar y analizar como se describe en el presente documento una biblioteca de ADNc de aproximadamente 100 células.

La expresión “biblioteca de ADNc” se refiere a una colección de fragmentos de ADN complementario (ADNc) clonado, que constituyen juntos alguna parte del transcriptoma de una célula individual o una pluralidad de células individuales. Se produce ADNc a partir de ARNm completamente transcrito hallado en una célula y por lo tanto contiene solamente los genes expresados de una única célula o cuando se agrupan entre sí los genes expresados de una pluralidad de células individuales.

Como se usa en el presente documento, una “pluralidad” se refiere a una población de células y puede incluir cualquier número de células que se desea analizar. En algunos aspectos, una pluralidad de células incluye al menos 10 células, como alternativa al menos 25 células, como alternativa al menos 50 células, como alternativa al menos 100 células, como alternativa al menos 200 células, como alternativa al menos 500 células, como alternativa al menos 1000 células, como alternativa 5.000 células o como alternativa 10.000 células. En otro aspecto, una pluralidad de células incluye de 10 a 100 células, como alternativa de 50 a 200 células, como alternativa de 100 a 500 células, como alternativa de 100 a 1000, o como alternativa de 1.000 a 5.000 células.

La expresión “amplificación” o “amplificar” se refiere a un proceso por el que se forman copias extra o múltiples de un polinucleótido particular. La amplificación incluye métodos tales como PCR, amplificación por ligamiento (o reacción en cadena de la ligasa, LCR) y métodos de amplificación. Estos métodos se conocen y se practican ampliamente en la técnica. Véase, por ejemplo, Patente de Estados Unidos nº 4.683.195 y 4.683.202 e Innis *et al.*, "PCR protocols: a guide to method and applications" Academic Press, Incorporated (1990) (para PCR); y Wu *et al.* (1989) Genomics 4: 560-569 (para LCR). En general, el procedimiento de PCR describe un método de amplificación génica que está comprendido por (i) hibridación específica de secuencia de cebadores con genes específicos dentro de una muestra de ADN (o biblioteca), (ii) amplificación posterior que implica múltiples ciclos de hibridación, elongación y desnaturalización usando una ADN polimerasa, y (iii) exploración de los productos de PCR con respecto a una banda del tamaño correcto. Los cebadores usados son oligonucleótidos de longitud suficiente y secuencia apropiada para proporcionar inicio de polimerización, es decir cada cebador se diseña específicamente para que sea complementario de cada cadena del locus genómico para amplificar.

En el comercio se dispone de reactivos y hardware para realizar una reacción de amplificación. Los cebadores útiles para amplificar secuencias de una región génica particular son preferentemente complementarios de, e hibridan específicamente con, secuencias en la región diana o en sus regiones flanqueantes, y pueden prepararse usando las secuencias polinucleotídicas proporcionadas en el presente documento. Las secuencias de ácido nucleico generadas por amplificación pueden secuenciarse directamente.

Cuando se produce hibridación en una configuración antiparalela entre dos polinucleótidos monocatenarios, la reacción se denomina "hibridación" y los polinucleótidos se describen como "complementarios". Un polinucleótido bicatenario puede ser complementario u homólogo de otro polinucleótido, si puede producirse hibridación entre una de las cadenas del primer polinucleótido y el segundo. La complementariedad u homología (el grado en que un polinucleótido es complementario de otro) es cuantificable en términos de la proporción de bases en cadenas opuestas que se espera que formen enlaces de hidrógeno entre sí, de acuerdo con las normas de formación de pares de bases aceptadas en general.

Como se usa en el presente documento, una "célula individual" se refiere a una célula. Pueden obtenerse células individuales útiles en los métodos descritos en el presente documento de un tejido de interés, o de una biopsia, muestra de sangre o cultivo celular. Adicionalmente, pueden obtenerse células de órganos específicos, tejidos, tumores, neoplasias o similares y usarse los métodos descritos en el presente documento. Además, en general, pueden usarse en los métodos células de cualquier población, tal como una población de organismos unicelulares procariotas o eucariotas incluyendo bacterias o levaduras. En algunos aspectos, el método para preparar la biblioteca de ADNc puede incluir la etapa de obtener células individuales. Puede obtenerse una suspensión de células individuales usando métodos convencionales conocidos en la técnica incluyendo, por ejemplo, enzimáticamente usando tripsina o papaína para digerir proteínas que conectan células en muestras tisulares o liberar células adherentes en cultivo, o mecánicamente separando células en una muestra. Pueden colocarse células individuales en cualquier recipiente de reacción adecuado en el que puedan tratarse individualmente células individuales, por ejemplo una placa de 96 pocillos, de modo que cada célula individual se coloca en un único pocillo.

Se conocen en la técnica métodos para manipular células individuales e incluyen separación de células activadas por fluorescencia (FACS), micromanipulación y el uso de seleccionadores de células semiautomáticos (por ejemplo, el sistema de transferencia de células Quixell™ de Stoelting Co.). Las células individuales pueden seleccionarse, por ejemplo, individualmente basándose en características detectables por observación microscópica, tales como localización, morfología o expresión de gen indicador.

En algunos aspectos, puede liberarse ARNm de las células lisando las células. El análisis puede conseguirse, por ejemplo, calentando las células, o mediante el uso de detergentes u otros métodos químicos, o por una combinación de estos. Sin embargo, puede usarse cualquier método de lisis adecuado conocido en la técnica. Puede usarse provechosamente un procedimiento de lisis suave para prevenir la liberación de cromatina nuclear, evitando de este modo la contaminación genómica de la biblioteca de ADNc y para minimizar la degradación de ARN. Por ejemplo, calentar las células a 72 ºC durante 2 minutos en presencia de Tween-20 es suficiente para lisar las células sin dar como resultado contaminación genómica detectable de la cromatina nuclear. Como alternativa, las células pueden calentarse a 65 ºC durante 10 minutos en agua (Esumi *et al.*, Neurosci Res 60(4): 439-51 (2008)); o 70 ºC durante 90 segundos en tampón de PCR II (Applied Biosystems) complementado con NP-40 0,5 % (Kurimoto *et al.*, Nucleic Acids Res 34(5): e42 (2006)); o puede conseguirse lisis con una proteasa tal como Proteinasa K o mediante el uso de sales caotrópicas tales como guanidina isotiocianato (Publicación de Estados Unidos nº 2007/0281313).

Puede realizarse síntesis de ADNc a partir de ARNm en los métodos descritos en el presente documento directamente en lisados celulares, de modo que se añada una mezcla de reacción para transcripción inversa directamente a lisados celulares. Como alternativa, puede purificarse ARNm después de su liberación de las células. Esto puede ayudar a reducir la contaminación mitocondrial y ribosómica. Puede conseguirse purificación de ARNm por cualquier método conocido en la técnica, por ejemplo, uniendo el ARNm con una fase sólida. Los métodos de purificación usados habitualmente incluyen perlas paramagnéticas (por ejemplo, Dynabeads). Como alternativa, pueden retirarse selectivamente contaminantes específicos, tales como ARN ribosómico usando purificación de afinidad.

Se sintetiza típicamente ADNc a partir de ARNm por transcripción inversa. Se han descrito previamente métodos para sintetizar ADNc a partir de cantidades pequeñas de ARNm, incluyendo de células individuales (Kurimoto *et al.*, Nucleic Acids Res 34(5): e42 (2006): Kurimoto *et al.*, Nat Protoc 2(3): 739-52 (2007); y Esumi *et al.*, Neurosci Res 60(4): 439-51 (2008)). Para generar un ADNc amplificable, estos métodos introducen una secuencia de hibridación de cebadores en ambos extremos de cada molécula de ADNc de tal modo que la biblioteca de ADNc puede amplificarse usando un único cebador. El método de Kurimoto usa una polimerasa para añadir una cola de poli A 3' a la cadena de ADNc, que puede después amplificarse usando un cebador de oligo T universal. Por el contrario, el método de Esumi usa un método de cambio de molde para introducir una secuencia arbitraria en el extremo 3' del ADNc, que se diseña para ser complementaria inversa de la cola 3' del cebador de síntesis de ADNc. De nuevo, la biblioteca de ADNc puede amplificarse por un único cebador de PCR. La PCR de un único cebador aprovecha el efecto de supresión de PCR para reducir la amplificación de amplicones contaminantes cortos y dímeros de

cebadores (Dai *et al*., J Biotechnol 128(3): 435-43 (2007)). Como los dos extremos de cada amplicón son complementarios, los amplicones cortos formarán horquillas estables, que son malos moldes para PCR. Esto reduce la cantidad de ADNc truncado y mejora el rendimiento de moléculas de ADNc más largas.

En algunos aspectos de la invención, la síntesis de la primera cadena del ADNc puede dirigirse por un cebador de síntesis de ADNc (CDS) que incluye una secuencia complementaria de ARN (RCS). En algunos aspectos de la invención, la RCS es al menos parcialmente complementaria de uno o más ARNm en una muestra de ARNm individual. Esto permite que el cebador, que es típicamente un oligonucleótido, hibride con al menos algo de ARNm en una muestra de ARNm individual para dirigir la síntesis de ADNc usando el ARNm como molde. La RCS puede comprender oligo (dT), o ser específica de familia génica, tal como una secuencia de ácidos nucleicos presente en todos o una mayoría de los genes relacionados, o puede estar compuesta de una secuencia aleatoria, tal como hexámeros aleatorios. Para evitar que el CDS sea cebador de sí mismo y por lo tanto genere productos secundarios indeseados, puede usarse una secuencia semialeatoria no autocomplementaria. Por ejemplo, puede excluirse una letra del código genético, o puede usarse un diseño más complejo restringiendo al mismo tiempo el CDS para que sea no autocomplementario.

Los términos “oligonucleótido” y “polinucleótido” se usan indistintamente y se refieren a una forma polimérica de nucleótidos de cualquier longitud, bien desoxirribonucleótidos o bien ribonucleótidos o análogos de los mismos. Los polinucleótidos pueden tener cualquier estructura tridimensional y pueden realizar cualquier función, conocida o desconocida. Los siguientes son ejemplos no limitantes de polinucleótidos: un gen o fragmento génico (por ejemplo, una sonda, un cebador, EST o marcador SAGE), exones, intrones, ARN mensajero (ARNm), ARN de transferencia, ARN ribosómico, ribozimas, ADNc polinucleótidos recombinantes, polinucleótidos ramificados, plásmidos, vectores, ADN aislado de cualquier secuencia, ARN aislado de cualquier secuencia, sondas y cebadores de ácido nucleico. Un polinucleótido puede comprender nucleótidos modificados, tales como nucleótidos metilados y análogos de nucleótidos. El término también se refiere a moléculas tanto bicatenarias como monocatenarias. A no ser que se especifique o requiera de otro modo, cualquier realización que comprenda un polinucleótido abarca tanto la forma bicatenaria como cada una de las dos formas monocatenarias complementarias que se sabe o se predice que componen la forma bicatenaria.

Un polinucleótido está compuesto de una secuencia específica de cuatro bases nucleotídicas: adenina (A); citosina (C); guanina (G); timina (T); y uracilo (U) en lugar de timina cuando el polinucleótido es de ARN. Por lo tanto, la expresión secuencia polinucleotídica es la representación alfabética de una molécula polinucleotídica. Esta representación alfabética puede introducirse en bases de datos en un ordenador que tenga una unidad de procesamiento central y usarse para aplicaciones bioinformáticas tales como genómica funcional y búsqueda de homología.

Un “cebador” es un polinucleótido corto, generalmente con un grupo OH 3’ libre que se une con una diana o un molde potencialmente presente en una muestra de interés hibridando con la diana, y a continuación promoviendo la polimerización de un polinucleótido complementario de la diana. Los cebadores de la presente invención están comprendidos por nucleótidos que varían de 17 a 30 nucleótidos. En un aspecto, el cebador es de al menos 17 nucleótidos, como alternativa, al menos 18 nucleótidos, como alternativa, al menos 19 nucleótidos. como alternativa, al menos 20 nucleótidos, como alternativa, al menos 21 nucleótidos, como alternativa, al menos 22 nucleótidos, como alternativa, al menos 23 nucleótidos, como alternativa, al menos 24 nucleótidos, como alternativa, al menos 25 nucleótidos, como alternativa, al menos 26 nucleótidos, como alternativa, al menos 27 nucleótidos, como alternativa, al menos 28 nucleótidos, como alternativa, al menos 29 nucleótidos, como alternativa, al menos 30 nucleótidos, como alternativa al menos 50 nucleótidos, como alternativa al menos 75 nucleótidos o como alternativa al menos 100 nucleótidos.

La RCS también puede ser al menos parcialmente complementaria de una parte de la primera cadena de ADNc, de modo que sea capaz de dirigir la síntesis de una segunda cadena de ADNc usando la primera cadena del ADNc como molde. Por lo tanto, después de la síntesis de primera cadena, puede añadirse una enzima RNasa (por ejemplo una enzima que tenga actividad RNasaH) después de la síntesis de la primera cadena de ADNc para degradar la cadena de ARN y para permitir que el CDS hibride de nuevo en la primera cadena para dirigir la síntesis de una segunda cadena de ADNc. Por ejemplo, la RCS podría comprender hexámeros aleatorios, o una secuencia semialeatoria no autocomplementaria (que minimiza la autohibridación del CDS).

Puede añadirse un oligonucleótido de cambio de molde (TSO) que incluye una parte que es al menos parcialmente complementaria de una parte del extremo 3’ de la primera cadena de ADNc a cada muestra de ARNm individual en los métodos descritos en el presente documento. Dicho método de cambio de molde se describe en (Esumi *et al*., Neurosci Res 60(4): 439-51 (2008)) y permite sintetizar ADNc de longitud completa que comprende el extremo 5’ completo del ARNm. Como la actividad transferasa terminal de la transcriptasa inversa típicamente provoca que se incorporen 2-5 citosinas en el extremo 3’ de la primera cadena de ADNc sintetizado a partir de ARNm, la primera cadena de ADNc puede incluir una pluralidad de citosinas, o análogos de citosina que forman pares de bases con guanosina, en su extremo 3’ (véase documento US 5.962.272). En un aspecto de la invención, la primera cadena de ADNc puede incluir una parte 3’ que comprende al menos 2, al menos 3, al menos 4, al menos 5 o 2, 3, 4 o 5

citosinas o análogos de citosina que forman pares de bases con guanosina. Un ejemplo no limitante de un análogo de citosina que forma pares de bases con guanosina es 5-aminoalil-2'-desoxicitidina.

En un aspecto de la invención, el TSO puede incluir una parte 3' que comprende una pluralidad de guanosinas o análogos de guanosina que forman pares de bases con citosina. Los ejemplos no limitantes de guanosinas o análogos de guanosina útiles en los métodos descritos en el presente documento incluyen, pero sin limitación, desoxirriboguanosina, riboguanosina, guanosina de ácido nucleico bloqueado y guanosina de ácido nucleico peptídico. Las guanosinas pueden ser ribonucleósidos o monómeros de ácido nucleico bloqueado.

Un ácido nucleico bloqueado (LNA) es un nucleótido de ARN modificado. El resto de ribosa de un nucleótido LNA está modificado con un enlace extra que conecta el oxígeno 2' y el carbono 4'. El enlace "bloquea" la ribosa en la conformación 3'-endo (Norte). Algunas de las ventajas de usar LNA en los métodos de la invención incluyen aumentar la estabilidad térmica de dobles cadenas, aumento de la especificidad de diana y resistencia de exo y endonucleasas.

Un ácido nucleico peptídico (PNA) es un polímero sintetizado de forma artificial similar a ADN o ARN, en el que la cadena principal está compuesta de unidades de N-(2-aminoetil)-glicina repetidas unidas por enlaces peptídicos. La cadena principal de un PNA es sustancialmente no iónica en condiciones neutras, a diferencia de la cadena principal de fosfodiéster altamente cargada de ácidos nucleicos de origen natural. Esto proporciona dos ventajas no limitantes. En primer lugar, la cadena principal de PNA muestra cinética de hibridación mejorada. En segundo lugar, los PNA tienen mayores cambios en la temperatura de fusión (Tf) para pares de bases perfectamente coincidentes frente a desapareados. El ADN y ARN muestran típicamente un descenso de 2-4 ºC en la Tf para un desapareamiento interno. Con la cadena principal de PNA no iónica, el descenso es más cercano a 7-9 ºC. Esto puede proporcionar una mejor diferenciación de secuencia. De forma similar, debido a su naturaleza no iónica, la hibridación de las bases unidas a estas cadenas principales es relativamente insensible a la concentración salina.

Un ácido nucleico útil en la invención puede contener un resto de azúcar no natural en la cadena principal. Las modificaciones de azúcares ejemplares incluyen pero sin limitación modificaciones 2' tales como adición de halógeno, alquilo, alquilo sustituido, SH, SCH_3, OCN, Cl, Br, CN, CF_3, OCF_3, SO_2CH_3, OSO_2, SO_3, CH_3, ONO_2, NO_2, N_3, NH_2, sililo sustituido y similares. También pueden realizarse modificaciones similares en otras posiciones en el azúcar, particularmente en la posición 3' del azúcar en el nucleótido 3' terminal o en oligonucleótidos con enlace 2'-5' y la posición 5' del nucleótido 5' terminal. Los ácidos nucleicos, análogos de nucleósidos o análogos de nucleótidos que tienen modificaciones de azúcares pueden modificarse adicionalmente para incluir un grupo de bloqueo reversible, marcador de enlace peptídico o ambos. En las realizaciones en las que están presentes las modificaciones 2' anteriormente descritas, la base puede tener un marcador con enlaces peptídicos.

Un ácido nucleico usado en la invención también 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 uracilo, adenina, citosina o guanina. Las bases no nativas ejemplares que pueden incluirse en un ácido nucleico, que bien tienen una cadena principal nativa o una estructura análoga incluyen, sin limitación, inosina, xantanina, hipoxantanina, isocitosina, isoguanina, 5-metilcitosina, 5-hidroximetil citosina, 2-aminoadenina. 6-metil adenina, 6-metil guanina, 2-propil guanina, 2-propil adenina, 2-tiouracilo, 2-tiotimina, 2-tiocitosina, 15-halouracilo, 15-halocitosina, 5-propinil uracilo, 5-propinil citosina. 6-azo uracilo, 6-azo citosina. 6-azo timina, 5-uracilo, 4-tiouracilo, 8-halo adenina o guanina, 8-amino adenina o guanina, 8-tiol adenina o guanina, 8-tioalquil adenina o guanina, 8-hidroxil adenina o guanina, uracilo o citosina 5-halo substituido, 7-metilguanina, 7-metiladenina, 8-azaguanina, 8-azaadenina, 7-desazaguanina, 7-desazaadenina, 3-desazaguanina, 3-desazaadenina o similares. Una realización particular puede utilizar isocitosina e isoguanina en un ácido nucleico para reducir la hibridación no específica, como se describe en general en la Patente de Estados Unidos nº 5.681.702.

Una base no nativa usada en un ácido nucleico puede tener actividad de formación de pares de bases universal, en la que es capaz de formar pares de bases con cualquier otra base de origen natural. Las bases ejemplares que tienen actividad de formación de pares de bases universal incluyen 3-nitropirrol y 5-nitroindol. Otras bases que pueden usarse incluyen las que tienen actividad de formación de pares de bases con un subconjunto de las bases de origen natural tales como inosina, que forma pares de bases con citosina, adenina o uracilo.

En un aspecto de la invención, el TSO puede incluir una parte 3' que incluye al menos 2, al menos 3, al menos 4, al menos 5 o 2, 3, 4, o 5, o 2-5 guanosinas, o análogos de guanosina que forman pares de bases con citosina. La presencia de una pluralidad de guanosinas o análogos de guanosina que forman pares de bases con citosina) permite que el TSO hibride de forma transitoria con las citosinas expuestas en el extremo 3' de la primera cadena de ADNc. Esto provoca que la transcriptasa inversa cambie de molde y continúe la síntesis de una cadena complementaria del TSO. En un aspecto de la invención, el extremo 3' del TSO puede bloquearse, por ejemplo por un grupo fosfato 3', para evitar que el TSO actúe como un cebador durante la síntesis de ADNc.

En un aspecto, el ARNm se libera de las células por lisis celular. Si la lisis se consigue parcialmente por

calentamiento, entonces el CDS y/o el TSO pueden añadirse a cada muestra de ARNm individual durante la lisis celular, ya que esto ayudará a la hibridación de los oligonucleótidos. En algunos aspectos, puede añadirse transcriptasa inversa después de la lisis celular para evitar la desnaturalización de la enzima.

En algunos aspectos de la invención, puede incorporarse un marcador en el ADNc durante su síntesis. Por ejemplo, el CDS y/o el TSO pueden incluir un marcador, tal como una secuencia de nucleótidos particular, que puede ser de al menos 4, al menos 5, al menos 6, al menos 7, al menos 8, al menos 9, al menos 10, al menos 15 o al menos 20 nucleótidos de longitud. Por ejemplo, el marcador puede ser una secuencia de nucleótidos de 4-20 nucleótidos de longitud, por ejemplo, 4, 5, 6, 7, 8, 9, 10, 15 o 20 nucleótidos de longitud. Como el marcador está presente en el CDS y/o el TSO se incorporará en el ADNc durante su síntesis y puede actuar por lo tanto como un "código de barras" para identificar el ADNc. Tanto el CDS como el TSO pueden incluir un marcador. La CDS y el TSO pueden incluir cada uno un marcador diferente, de modo que la muestra de ADNc marcada comprende una combinación de marcadores. Cada muestra de ADNc generada por el método anterior puede tener un marcador distinto, o una combinación distinta de marcadores, de modo que una vez que se han agrupado las muestras de ADNc marcadas, el marcador puede usarse para identificar de qué célula individual se originó cada muestra de ADNc. Por lo tanto, cada muestra de ADNc puede ligarse a una única célula, incluso después de haberse agrupado las muestras de ADNc marcadas en los métodos descritos en el presente documento.

Antes de agruparse las muestras de ADNc marcadas, la síntesis de ADNc puede detenerse, por ejemplo retirando o inactivando la transcriptasa inversa. Esto evita que la síntesis de ADNc por transcripción inversa continúe en las muestras agrupadas. Las muestras de ADNc marcadas pueden purificarse opcionalmente antes de la amplificación, bien antes o bien después de agruparse.

Las muestras de ADNc agrupadas pueden amplificarse por reacción en cadena de la polimerasa (PCR) incluyendo PCR de emulsión y PCR de cebador individual en los métodos descritos en el presente documento. Por ejemplo, las muestras de ADNc pueden amplificarse por PCR de cebador individual. El CDS puede comprender una secuencia de cebador de amplificación 5' (APS), que posteriormente permite que la primera cadena de ADNc se amplifique por PCR usando un cebador que es complementario de la APS 5'. El TSO también puede comprender una APS 5', que puede ser al menos 70 % idéntica, al menos 80 % idéntica, al menos 90 % idéntica, al menos 95 % idéntica, o 70 %, 80 %, 90 % o 100 % idéntica a la APS 5' en el CDS. Esto significa que las muestras de ADNc agrupadas pueden amplificarse por PCR usando un cebador individual (es decir por PCR de cebador individual), que aprovecha el efecto de supresión de PCR para reducir la amplificación de amplicones contaminantes cortos y dímeros de cebadores (Dai *et al*., J Biotechnol 128(3): 435-43 (2007)). Como los dos extremos de cada amplicón son complementarios, los amplicones cortos formarán horquillas estables, que son malos moldes para PCR. Esto reduce la cantidad de ADNc truncado y mejora el rendimiento de moléculas de ADNc más largas. La APS 5' puede diseñarse para facilitar el procesamiento corriente abajo de la biblioteca de ADNc. Por ejemplo, si la biblioteca de ADNc va a analizarse por un método de secuenciación particular, por ejemplo, la tecnología de secuenciación SOLiD de Applied Biosystems, o el Analizador de Genoma de Illumina, la APS 5' puede diseñarse para ser idéntica a los cebadores usados en estos métodos de secuenciación. Por ejemplo, la APS 5' puede ser idéntica al cebador P1 de SOLiD y/o una secuencia P2 de SOLiD insertada en el CDS, de modo que las secuencias P1 y P2 requeridas para secuenciación por SOLiD sean integrales de la biblioteca amplificada.

Otro método ejemplar para amplificar ADNc agrupado incluye PCR. La PCR es una reacción en la que se realizan copias repetidas de un polinucleótido diana usando un par de cebadores o un conjunto de cebadores que consisten en un cebador cadena arriba y uno cadena abajo, y un catalizador de polimerización, tal como un ADN polimerasa, y típicamente una enzima polimerasa termoestable. Se conocen bien en la técnica métodos para PCR, y se enseñan, por ejemplo, en MacPherson *et al*. (1991) PCR 1 : A Practical Approach (IRL Press at Oxford University Press). Todos los procesos para producir copias repetidas de un polinucleótido, tales como PCR o clonación génica, se denominan colectivamente en el presente documento replicación. También puede usarse un cebador como una sonda en reacciones de hibridación, tales como análisis de transferencia de Southern o Northern.

Para PCR de emulsión, se crea una reacción de PCR de emulsión agitando vigorosamente o removiendo una mezcla de "agua en aceite" para generar millones de compartimentos acuosos de tamaños micrométricos. La biblioteca de ADN se mezcla en una dilución limitante bien con las perlas antes de la emulsificación o directamente en la mezcla de emulsión. La combinación del tamaño del compartimento y la dilución limitante de las perlas y moléculas diana se usa para generar compartimentos que contengan, en promedio, solamente una molécula de ADN y perla (a la dilución óptima muchos compartimentos tendrán perlas sin ninguna diana). Para facilitar la eficacia de amplificación, se incluyen cebadores de PCR tanto cadena arriba (concentración baja, coincide con la secuencia de cebador en la perla) y cadena abajo (alta concentración) en la mezcla de reacción. Dependiendo del tamaño de los compartimentos acuosos generados durante la etapa de emulsión, pueden realizarse hasta $3x10^9$ reacciones de PCR individuales por µl simultáneamente en el mismo tubo. Esencialmente cada compartimento pequeño en la emulsión forma un microrreactor de PCR. El tamaño promedio de un compartimento en una emulsión varía de un diámetro submicrométrico a más de 100 micrómetros, dependiendo de las condiciones de emulsión.

"Identidad", "homología" o "similitud" se usan indistintamente y se refieren a la similitud de secuencia entre dos

moléculas de ácido nucleico. La identidad puede determinarse comparando una posición en cada secuencia que pueda alinearse para fines de comparación. Cuando una posición en la secuencia comparada está ocupada por la misma base o el mismo aminoácido, entonces las moléculas son homólogas en esa posición. Un grado de identidad entre secuencias es una función del número de posiciones coincidentes o idénticas compartidas por las secuencias. Una secuencia no relacionada o no homóloga comparte menos del 40 % de identidad, o como alternativa menos del 25 % de identidad, con una de las secuencias.

Que un polinucleótido tenga un cierto porcentaje (por ejemplo, 60 %, 65 %, 70 %, 75 %, 80 %, 85 %, 90 %, 95 %, 98 % o 99 %) de "identidad de secuencia" con otra secuencia significa que, cuando se alinean, el porcentaje de bases son iguales en la comparación de las dos secuencias. Este alineamiento y el porcentaje de identidad de secuencias u homología pueden determinarse usando programas informáticos conocidos en la técnica, por ejemplo los descritos en Ausubel *et al.*, Current Protocols in Molecular Biology, John Wiley & Sons, Nueva York, N. Y., (1993). Preferentemente, se usan parámetros por defecto para el alineamiento. Un programa de alineamiento es BLAST, usando parámetros por defecto. En particular, los programas son BLASTN y BLASTP, usando los siguientes parámetros por defecto: código genético = convencional; filtro = ninguno; cadena = ambas; punto de corte = 60; expectativa = 10; Matriz = BLOSUM62; Descripciones = 50 secuencias; clasificar por = ALTA PUNTUACIÓN; Bases de datos = no redundantes, GenBank + EMBL + DDBJ + PDB + GenBank CDS translations + SwissProtein + SPupdate + PIR. Pueden encontrarse detalles de estos programas en el Centro Nacional para la Información Biotecnológica.

El método para preparar una biblioteca de ADNc descrito en el presente documento puede comprender además procesar la biblioteca de ADNc para obtener una biblioteca adecuada para secuenciación. Como se usa en el presente documento, una biblioteca es adecuada para secuenciación cuando la complejidad, el tamaño, la pureza o similares de una biblioteca de ADNc son adecuados para el método de exploración deseado. En particular, la biblioteca de ADNc puede procesarse para hacer a la muestra adecuada para cualquier método de exploración de alto rendimiento, tal como tecnología de secuenciación SOLiD Applied Biosystems o Analizador de Genoma de Illumina. Como tal, la biblioteca de ADNc puede procesarse fragmentando la biblioteca de ADNc (por ejemplo con DNasa) para obtener una biblioteca de extremo 5' de fragmentos cortos. Pueden añadirse adaptadores al ADNc, por ejemplo en uno o ambos extremos para facilitar la secuenciación de la biblioteca. La biblioteca de ADNc puede amplificarse adicionalmente, por ejemplo por PCR, para obtener una cantidad suficiente de ADNc para secuenciación.

Las realizaciones de la invención proporcionan una biblioteca de ADNc producida por cualquiera de los métodos descritos en el presente documento. Esta biblioteca de ADNc puede secuenciarse para proporcionar un análisis de expresión génica en células individuales o en una pluralidad de células individuales.

Las realizaciones de la invención también proporcionan un método para analizar la expresión génica en una pluralidad de células individuales, comprendiendo el método las etapas de preparar una biblioteca de ADNc usando el método descrito en el presente documento y secuenciando la biblioteca de ADNc. Un "gen" se refiere a un polinucleótido que contiene al menos una fase abierta de lectura (ORF) que es capaz de codificar un polipéptido o una proteína particular después de transcribirse y traducirse. Puede usarse cualquiera de las secuencias polinucleotídicas descritas en el presente documento para identificar fragmentos mayores o secuencias codificantes de longitud completa del gen con los que se asocian. Se conocen por los expertos en la materia métodos para aislar secuencias de fragmentos mayores.

Como se usa en el presente documento, "expresión" se refiere al proceso por el que se transcriben polinucleótidos a ARNm y/o el proceso por el que el ARNm transcrito se traduce posteriormente en péptidos, polipéptidos o proteínas. Si el polinucleótido deriva de ADN genómico, la expresión puede incluir corte y empalme del ARNm en una célula eucariota.

La biblioteca de ADNc puede secuenciarse por cualquier método de exploración adecuado. En particular, la biblioteca de ADNc puede secuenciarse usando un método de exploración de alto rendimiento, tal como la tecnología de secuenciación SOLiD de Applied Biosystems, o Analizador del Genoma de Illumina. En un aspecto de la invención, la biblioteca de ADNc puede secuenciarse al azar. El número de lecturas puede ser de al menos 10.000, al menos 1 millón, al menos 10 millones, al menos 100 millones, o al menos 1000 millones. En otro aspecto, el número de lecturas puede ser de 10.000 a 100.000, como alternativa de 100.000 a 1 millón, como alternativa de 1 millón a 10 millones, como alternativa de 10 millones a 100 millones, o como alternativa de 100 millones a 1000 millones. Una "lectura" es una longitud de secuencia de ácido nucleico continua obtenida por una reacción de secuenciación.

La "secuenciación al azar" se refiere a un método usado para secuenciar una cantidad muy grande de ADN (tal como el genoma completo). En este método, el ADN para secuenciar se rompe primero en fragmentos más pequeños que pueden secuenciarse individualmente. Las secuencias de estos fragmentos se vuelven a ensamblar después en su orden original basándose en sus secuencias solapantes, produciendo de este modo una secuencia completa. La "rotura" del ADN puede realizarse usando varias técnicas diferentes incluyendo digestión con enzimas

de restricción o corte mecánico. Las secuencias solapantes típicamente se alinean por un ordenador convenientemente programado. Se conocen bien en la técnica métodos y programas para secuenciación al azar de una biblioteca de ADNc.

Una realización del método de la invención se resume en la **Figura 1**. Las células se obtienen de un tejido de interés y se obtiene una suspensión de células individuales. Se coloca una célula individual en un pocillo de una placa de 96 pocillos en mezcla de captura celular. Las células se lisan y se añade mezcla de reacción de transcripción inversa directamente a los lisados sin purificación adicional. Esto da como resultado la síntesis de ADNc a partir de ARNm celular y la incorporación de un marcador en el ADNc. Las muestras de ADNc marcadas se agrupan y amplifican y después se secuencian para producir 100 millones de lecturas. Esto permite la identificación de genes que se expresan en cada célula individual.

Se entiende que también se proporcionan modificaciones que no afectan sustancialmente a la actividad de las diversas realizaciones de la presente invención dentro de la definición de la invención proporcionada en el presente documento. En consecuencia, se pretende que los siguientes ejemplos ilustren pero no limiten la presente invención.

Ejemplo 1

Transcripción inversa marcada en células individuales (STRT)

Una realización del método de la invención puede denominarse “transcripción inversa marcada en células individuales” (STRT) y se describe en detalle a continuación.

Recogida y lisis de células

Se preparó una placa de 96 pocillos que contenía Mezcla de Captura Celular separando en alícuotas 5 µl/pocillo de la Placa Maestra de Captura Celular (véase Tabla 1 posterior) en una placa Thermo-Fast AbGene.

Reactivo	**Para un pocillo**	**Para una placa**	**Concentración final**
STRT-T30-BIO (100 µM)	0,25 µl	27,5 µl	400 nM
Reactivo	**Para un pocillo**	**Para una placa**	**Concentración final**
Tampón de STRT (5x)	12,5 µl	1375 µl	1x
STRT-FW-n (5 µM)	5 µl	(5 µl/pocillo)	400 nM
Agua	44,75	4,9 ml	
Total	**62,5 µl**		

Tabla 1 preparar una Placa Maestra de Captura Celular de STRT.

Se mezclaron 27,5 µl de STRT-T30-BIO (100 µM) con 1375 µl de tampón STRT 5x y 4,9 ml de agua sin Rnasa/Dnasa. Se separaron en alícuotas 57,5 µl de esta solución a casa pocillo de una placa de 96 pocillos y se añadieron 5 µl/pocillos de STRT-FW-n (de la placa de reserva 5 µM), es decir un oligo diferente en cada pocillo.

La secuencia de STRT-T30-BIO (que es un CDS) es:

5’-BIO-AAGCAGTGGTATCA.ACGCAGAGT$_{30}$VN-3’,

y la secuencia de STRT-FW-n (que es un TSO) es:

5’-AAGCAGTGGTATCAACGCAGAGTGGATGCTXXXXXrGrGrG-3’(X=marcador celular)

BtsCI→2/0

n es 1-96 y cada oligonucleótido tiene un marcador celular distinto, de modo que se añade un oligonucleótido diferente a cada pocillo que contiene una única célula.

Se cultivaron celular madre embrionarias de ratón (R1) sin células de alimentación, se tripsinizaron, se clarificaron mediante un tamiz celular y se resuspendieron en PBS 1x. Las células se seleccionaron después por FACS en la Placa de Captura, colocándose una única célula en cada pocillo. La Placa de Captura de transfirió a un termociclador de PCR y se incubó a 72 ºC durante 2 minutos, y después se enfrío a 4 ºC durante 5 minutos para permitir que se produjera la hibridación. El detergente en tampón de STRT ayuda a reducir la adsorción de ARNm y ADNc a las paredes del tubo de reacción durante etapas posteriores, y también mejora la lisis de las células. La

etapa de calentamiento provoca que la célula se lise completamente y libere su ARN. Cuando la temperatura se reduce, el cebador de oligo (dT) hibrida.

Transcripción inversa

Se añadieron 5 µl/pocillo de mezcla de RT (véase Tabla 2 posterior) y la placa se incubó a 42 ºC durante 45 minutos, sin tapa calentada.

Reactivo	**Para 96 reacciones**	**Para una reacción**	**Concentración final**
Tampón de STRT (5x)	110 µl	1 µl	1x
DTT (20 mM)	110 µl	1 µl	2 mM
dNTP (10 mM)	110 µl	1 µl	1 mM
Agua sin Dnasa/Rnasa	209 µl	1 µl	
Superscript 11 RT (200 U/µl)	11 µl	1 µl de 20 U/µl (diluir en tampón RT 1x)	2,5 U/µl
Volumen total	**550 µl**	**5 µl**	

Tabla 2 Composición de mezcla de RT

Cuando se añade la mezcla de RT, la enzima transcriptasa inversa (Superscript II RT) sintetiza una primera cadena y el oligo de cambio de molde marcado introduce una secuencia de cebador cadena arriba.

La **Figura 2** muestra la síntesis de ADNc por cambio de molde. El extremo 5' del ADNc (que corresponde al extremo 3' del ARNm) puede controlarse añadiendo una cola (que es oligo dT en este caso) al cebador de síntesis de ADNc (CDS). El extremo 3' del ADNc puede controlarse usando el oligo de cambio de molde (TSO). Cuando la transcriptasa inversa alcanza el extremo 5' del molde de ARNm, preferentemente añade 2-5 citosinas. El oligo de cambio de molde, que tiene 2-5 guanosinas, hibrida de forma transitoria, y la transcriptasa inversa después cambia de molde y sintetiza la cadena complementaria. Por este mecanismo, ambos extremos del ADNc pueden controlarse de forma arbitraria.

La estructura de un TSO típico se muestra en la **Figura 3**. En este TSO particular, la secuencia de cambio de molde 3' incluye tres riboguaninas (rG). El marcador celular se muestra como "XXXXX" y puede tener en general cualquier longitud o composición de nucleótidos. Puede insertarse una secuencia arbitraria en el extremo 5' del TSO, después de la APS 5', o después del marcador celular, pero no en el extremo 3'.

La estructura de un CDS típico se muestra en la **Figura 4**. La RCS es oligo dT con un nucleótido de anclaje (V = A, C, G degradado). El marcador celular "XXXXX" puede tener cualquier longitud o composición de nucleótidos. Adicionalmente, pueden insertarse secuencias arbitrarias en el extremo 5', después de la APS 5' o después del marcador celular.

Purificación de ADNc

Se añadieron 50 µl de PBI (Kit de Purificación de PCR Qiaquick) a cada pocillo para inactivar la transcriptasa inversa. El PBI inactiva la transcriptasa inversa y después se agrupó ADNc de todos los pocillos. La adición de PBI antes del agrupamiento evita que se produzca síntesis de ADNc una vez que se han agrupado las muestras de ADNc. El ADNc agrupado se cargó en una única columna de Qiaquick y el ADNc purificado se eluyó en 30 µl de tampón EB a un tubo de Polialómero Beckman. La etapa de purificación retira los cebadores (<40 pb) así como proteínas y otros residuos.

Amplificación de ADNc de longitud completa

El ADNc se amplificó por PCR añadiendo los reactivos mostrados en la Tabla 3.

Reactivo	**Para un tubo:**	**Concentración final**
Agua sin Rnasa/Dnasa	54 µl	
Tampón de PCR Advantage2 (10x)	10 µl	1x
dNTP (10 mM)	2 µl	200 µM
STRT-PCR (10 µM)	2 µl	200 µM
Mezcla de ADN Polimerasa Advantage2 (50x)	2 µl	1x

Reactivo	Para un tubo:	Concentración final
Volumen total	**100** µl	

Tabla 3. Reactivos usados para amplificación de ADNc de longitud completa.

La secuencia de STRT-PCR es:
5'-BIO-AAGCAGTGGTATCAACGCAGAGT-3'

Se realizó PCR usando una tapa calentada de la siguiente menara: 1 min. a 95 ºC, 25 ciclos de [5 s a 95 ºC, 5 s a 65 ºC 6 min. a 68 ºC] 4 ºC para siempre.

Se transfirieron 30 µl de la reacción a un nuevo tubo de PCR, marcado "Optimización". Los 70 µl restantes se almacenaron a 4 ºC hasta más tarde. Se retiraron 10 µl del tubo de Optimización y el resto de la muestra se procesó durante tres ciclos más. Esto se repitió para obtener alícuotas de 25, 28 y 31 ciclos. Se usó un gel de agarosa al 2 % de diagnóstico para determinar el número de ciclos óptimo (que es el ciclo justo antes de la saturación de la PCR), así como para visualizar el intervalo de tamaños del producto (véase **Figura 5**). Típicamente, el número óptimo de ciclos fue de aproximadamente 28. Los 70 µl restantes de reacción se procesaron para alcanzar el número óptimo de ciclos (además de los 25 ciclos ya procesados).

El producto de PCR se purificó usando una columna Qiaquick (kit de purificación de PCR) y se eluyó en 50 µl de EB en un tubo de polialómero de Beckman. La concentración esperada en ese estadio fue de aproximadamente 20-40 ng/µl (1-2 µg de rendimiento total).

Tratamiento con DNasa

La muestra se trató con DNasaI en presencia de Mn^{2+} para generar roturas de doble cadena y reducir el tamaño. En primer lugar, se mezclaron los siguientes componentes en el orden mostrado en la Tabla 4.

Reactivo	Volumen	Concentración final
Molde de ADNc	50 µl	8 - 16 ng/µl
Agua	42,8 µl	
Tampón de DNasaI 10x	11,6 µl	
$MnCl_2$ 100 mM (*)	11,6 µl	10 mM
Volumen total	**116** µl	

(*) Es crucial añadir $MnCl_2$ lo último a la reacción, ya que de otro modo la BSA presente en el tampón precipitará.

Tabla 4. Composición de mezcla de reacción para tratamiento con DNasa.

Se preparó DNasa I diluida (0,01 unidades/µl) justo antes de su uso de la siguiente manera: 40 µl de tampón de DNasa I 10x, 318 µl de agua, 40 µl de $MnCl_2$ 100 mM y 2 µl de DNasaI (2 U/µl).

Se añadieron 4 µl de esta DNasa I diluida a la mezcla de reacción descrita en la Tabla 4, y se incubó a TA durante exactamente 10 minutos. La reacción se detuvo después añadiendo 600 µl de PBI.

La muestra se purificó en una columna Qiaquick y se eluyó en 30 µl de EB.

Captura de perlas y reparación de extremos/traslación de muesca

Los fragmentos se unieron a continuación con perlas para capturar extremos 5' y 3', y después se trataron con TaqExpress para reparar extremos deshilachados y muescas. Se lavaron 30 µl de Estreptavidina MyOne Cl Dynabeads dos veces en B y W 2x (Dynal), después se añadió a la muestra tratada con DNasa, se incubó durante 10 minutos, y después se lavó 3x en B y W 1x. Aproximadamente el 10 % de la muestra se unió a las perlas (es decir aproximadamente 30 - 60 ng), ya que los fragmentos internos no se biotinilaron.

Las perlas se lavaron una vez en tampón TaqExpress 1x y se resuspendieron en la mezcla de reacción mostrada en la Tabla 5:

Reactivo	Volumen	Concentración final
Tampón de TaqExpress 10x (azul, con $MgCl_2$)	4 µl	1x
dNTP(10 mM)	0,8 µl	200 µM
Agua	33 µl	
TaqExpress (5 U/µl)	2 µl	0,25 U/µl
Volumen total	**116 µl**	

Tabla 5. Composición de la mezcla de reacción usada para reparación de extremos/traslación de muescas.

La reacción se incubó a 37 ºC durante 30 minutos, y después se lavó tres veces en tampón de NEB4 1x.

Liberación de fragmentos y ligamiento del adaptador RDV/FDV

Los fragmentos se liberaron por digestión con BtsCI, y simultáneamente se ligaron con los adaptadores de FDV y RDV. Las perlas se resuspendieron después en la mezcla de reacción mostrada en la Tabla 6.

Reactivo	Volumen	Concentración final
Tampón NEB4 10x	4 µl	1x
ATP (10 mM)	4 µl	1 mM
Adaptador STRT-RDV-A (10 µM)	4 µl	1 µM
Adaptador STRT-FDV (10 µM)	4 µl	1 µM
Agua	26 µl	
ADN Ligasa T4 (5 U/µl: Invitrogen)	2 µl	0,25 U/µl
BtsCI (20 U/µl)	2 µl	1 U/µl
Volumen total	**40 µl**	

Tabla 6. Mezcla de reacción para resuspensión de las perlas.

La secuencia de STRT-FDV, realizada hibridando STRT-ADP1U y STRT-ADP1L, fue:

5'-----CCACTACGCCTCCGCTTTCCTCTCTATGGGCAGTCGGTGATCT-3'
3'-PHO-GGTGATGCGGAGGCGAAAGGAGAGATACCCGTCAGCCACTA-PHO-5'

La secuencia de STRT-RDV-A, realizada por hibridación de STRT-ADP2U-T y STRT-ADP2L fue:

5'-----AACTGCCCCGGGTTCCTCATTCTCTT-3'
3'-PHO-TTGACGGGGCCCAAGGAGTAAGAGA-PHO-5'

Las perlas se incubaron durante 30 minutos a 37 ºC. La reacción se detuvo añadiendo 200 µl de PBI, mientras que las perlas se mantuvieron en el imán. El sobrenadante se cargó en una columna Qiaquick, se purificó y se eluyó en 30 µl de EB en un tubo de polialómero de Beckman. La concentración del ADNc fue de aproximadamente 1 - 2 ng/µl.

Amplificación por PCR de bibliotecas

Se prepararon ocho reacciones usando alícuotas de 4, 2, 1, 1/2, 1/4, 1/8, 1/16 y 1/32 µl de la biblioteca adaptada, cada una en 4 µl. Cada biblioteca se amplificó usando la mezcla de reacción de PCR mostrada en la Tabla 7.

Reactivo	Para un tubo:	Concentración final
Biblioteca de ADNc (serie de diluciones)	4 µl	(20 pg/µl)
Tampón azul TaqExpress (10x)	5 µl	1x
dNTP (10 mN)	1 µl	200 µM
SOLID-P1 (10 nM)	2 µl	400 nM
SOLID-P2 (10 µM)	2 µl	400 nM
Reactivo	**Para un tubo:**	**Concentración final**
Agua sin Rnasa/Dnasa	35,7 µl	0,25 U/µl

Reactivo	Para un tubo:	Concentración final
Polimerasa TaqExpress (25 U/µl)	0,3 µl	0,15 U/µl
Volumen total	**50 µl**	

Tabla 7. Mezcla de reacción de PCR para amplificación de biblioteca de ADNc.

La secuencia de SOLID-P1 fue:

5'-CCACTACGCCTCCGCTTTCCTCTCTATG-3'

La secuencia de SOLID-P2 fue:

5'-CTGCCCCGGGTTCCTCATTCTCT-3'

La PCR se procesó con tapa calentada: 5 min. a 94 ºC, 18 ciclos de [15 s a 94 ºC, 15 s a 68 ºC], 5 min. a 70 ºC.

Las ocho reacciones se cargaron en un E-gel 2 %, 10 µl + 10 µl de agua para determinar qué reacción estaba a punto de saturarse (véase **Figura 6**).

Después se realizó una nueva reacción de PCR usando el número óptimo de ciclos y material de partida. Por ejemplo, si 1/4 µl era óptimo a 18 ciclos, entonces se realizaron 14 ciclos.

El producto de PCR se cargó en un E-gel al 2 %, se escindió una región de 125 - 200 pb del gel y se purificó por Kit de Extracción en Gel Qiagen (véase **Figura 7**). El ADNc purificado se eluyó en 50 µl de EB.

La biblioteca de ADNc se preparó ahora para secuenciación por SOLiD, y podría ir directamente a PCR de emulsión.

Para verificar la calidad de la biblioteca de ADNc, se clonó una alícuota usando el kit de clonación TOPO TA de Invitrogen, y se secuenció por secuenciación de Sanger. La **Figura 8** muestra un resultado típico que demuestra la presencia de secuencias de cebadores para SOLiD (P1 y P2; subrayado), el marcador específico de célula (encuadrado), y las 2-5 G (sombreadas en una caja gris) del mecanismo de cambio de molde. A partir de 22 secuencias Sanger, 7 no eran mapeables en nada en GenBank. Todas excepto una de estas fueron ligamientos erróneos de los adaptadores de SOLiD, que pueden rediseñarse para evitar que suceda esto. En experimentos separados, no se encontró ningún adaptador con ligamientos erróneos después de bloquear sus extremos romos con fosfato 3'. Como alternativa, los extremos 3' no ligantes podrían bloquearse usando didesoxinucleótidos o diseñando una cadena protruyente incompatible con los extremos ligantes de los adaptadores.

De las 15 secuencias restantes, una era un ARN ribosómico (45S), que no estaba poliadenilado. Se produjo probablemente debido a colocación de cebadores errónea interna durante la síntesis de primera cadena. Las 14 lecturas restantes fueron todas de ARNm poliadenilado, en la orientación correcta y con marcadores celulares correctos.

Para resumir este conjunto de datos, 15 de 22 lecturas fueron maleables y 14 de estas 15 fueron marcadores de transcrito correctos. Todos los transcritos vistos en el conjunto de datos de secuencia de Sanger se enumeran a continuación:

Gen	Longitud (de ARNm)
Proteína ribosómica L35	452
B2_Mm2	~200
Tubulina beta 2c	1 561
B2 Mm1	~195
Gen de RIKEN 1110008L16	3 127
Sod2	661
Chchd2	910
mt-Cox2	947
Hnrnpab	2 545
Proteína ribosómica L24	558
Proteína ribosómica S18	524
RIKEN 2700060E02	941
B2_Mm1	~195

Gen	Longitud (de ARNm)
Proteína ribosómica S28	356

Como se esperaba, esta lista estaba dominada por genes altamente expresados como proteínas ribosómicas. Estaban presentes varios transcritos largos en esta muestra, lo que indica que no hubo ningún desvío fuerte (si hubo alguno) hacia ARNm cortos.

Resulta interesante que se observaron tres copias de repeticiones de B2 (de subfamilias Mm1 y Mm2). Estas son repeticiones de familia SINE expresadas a partir de un promotor pol III (no pol II como la mayoría de los ARNm), pero con fuertes señales de poliadenilación. Se ha mostrado que se expresan a niveles extremadamente altos en células ES, que comprendían juntas más del 10 % de todo el ARNm. Aún resulta más interesante que alcancen el máximo justo antes de la fase S en células en división, y por lo tanto es una indicación temprana de que usando este método será posible caracterizar el ciclo celular en células primarias desincronizadas.

Control de calidad por PCR en tiempo real cuantitativa

Para verificar que las bibliotecas eran representativas del contenido de ARNm de la población celular de ES original, se realizó PCR en tiempo real cuantitativa frente a un conjunto de marcadores con respecto a pluripotencialidad, así como marcadores para tejidos diferenciados. Se comparó una biblioteca de ADNc preparada de acuerdo con métodos clásicos a partir de 1 µg de ARN total (~100.000 células) con la biblioteca preparada a partir de 96 células individuales usando el protocolo de STRT.

Se detectaron marcadores bien conocidos de pluripotencialidad, tales como Sox2, Oct4 y Nanog a niveles similares en ambas muestras, mientras que se detectaron marcadores de diferenciación de capas germinales tales como Brachyury, Gata4 y Eomes solamente a niveles muy bajos en ambas muestras (véase **Figura 9**). La correlación cuantitativa fue buena (coeficiente de correlación de Pearson 0,84), con la excepción de Plk1, que no se detectó en células individuales en este experimento.

Reactivos usados

Reactivo	**Fuente**
B y W 2x	Tris HCl 10 mM pH 7,5, EDTA 1 mM, NaCl 2 M
Reserva de Tween-20 10 %	Preparar Tween-20 10 %, filtrar a 0,45 µM
Estreptavidina CI MyOne Dynabeads	Invitrogen
Tampón STRT 5x	Tris-HCl 100 mM pH 8, KCl 375 mM, $MgCl_2$ 30 mM, Tween-20 0,1 %
Agua sin DNasa/RNasa	Ambion
DTT 20 mM	
dNTP 10 mM y 1 mM	NEB / In Vitro
Superscript II RT	Invitrogen
Tampón de PCR Advantage2 10x	Clontech
Mezcla de polimerasa Advantage2 50x	Clontech
Tampón de Dnasa I 10x	Tris 0,5 M pH 7,5, BSA 0,5 mg/ml
$MnCl_2$ 100 mM	Reserva 1 M de Sigma
Placa de 96 pocillos Thermo-Fast	AbGene
Tiras de tubos Thermo-Strip	AbGene
Tubos de polialómero de 1,5 ml	Beckman-Coulter
E-gel al 2 % con SYBR Safe	Invitrogen
Kit de Purificación de PCR Qiaquick	Qiagen
Kit de Extracción en Gel Qiaquick	Qiagen
Endonucleasa de restricción BtsCI	NEB
ADN ligasa T4	Invitrogen
ATP 10 mM	
Tampón de NEB4 10x	NEB

Reactivo	Fuente
Polimerasa TaqExpress 5 U/µl	Genetix, Reino Unido
Tampón TaqExpress 10x (azul)	Genetix, Reino Unido

Tabla 8. Lista de reactivos usados en el método descrito anteriormente.

Ejemplo II

Caracterización del paisaje transcripcional de células individuales por ARN-Sec altamente múltiple

El entendimiento del desarrollo y mantenimiento de tejidos se ha visto ayudado en gran medida por el análisis de expresión génica a gran escala. Sin embargo, los tejidos son invariablemente complejos, consistentes en múltiples tipos celulares en una diversidad de estados moleculares. Como resultado, el análisis de expresión de un tejido confunde los patrones de presión verdaderos de sus tipos celulares constituyentes. Se describe en el presente documento una nueva estrategia, denominada perfil de expresión de células individuales al azar, que se usó para acceder a dichas muestras complejas. Es un método sencillo y altamente múltiple usado para generar cientos de perfiles de expresión de ARN-Sec de células individuales. Las células se agrupan después basándose en sus perfiles de expresión, formando un mapa celular bidimensional en el que pueden proyectarse datos de expresión. El mapa celular resultante integra tres niveles de organización: la población completa de células, las subpoblaciones funcionalmente distintas que contiene, y las células individuales en sí mismas, todas sin la necesidad de marcadores conocidos para clasificar los tipos celulares. La viabilidad de la estrategia se demuestra analizando los transcriptomas completos de 436 células individuales de tres tipos distintos. Esta estrategia permite el descubrimiento y análisis imparcial de tipos celulares de origen natural durante el desarrollo, fisiología del adulto y enfermedad.

Métodos

Cultivo celular

Se cultivaron células ES RI como se ha descrito previamente (Moliner *et al.*, Stem Cells Dev. 17: 233-243 (2008)). Se cultivaron células MEF y Neuro-2A en DMEM con FBS al 10 %, penicilina/estreptomicina 1x, Glutamax 1x y 2-mercaptoetanol 0,05 mM. Todos los reactivos de cultivo fueron de Gibco.

PCR en tiempo real cuantitativa (Q-PCR)

Se aisló ARN usando Trizol (Invitrogen) y se transcribió de forma inversa 1 µg de ARN total con Superscript III (Invitrogen) y cebador de oligo (dT). Se mezcló una Mezcla Maestra Verde de SYBR (Applied Biosystems) y una cantidad de ADNc correspondiente a 5 ng de ARN con 4 pmoles de cebadores (Eurofins MWG Operon, Alemania) en un volumen total de 10 µl, y se analizó en un termociclador en tiempo real 7900HT (Applied Biosystems). Se usó una serie de diluciones del molde para determinar la eficacia de cebadores.

Transcripción inversa marcada con células individuales (STRT)

Las células se disociaron enzimáticamente usando TrypLE Express (Invitrogen), se lavaron y se resuspendieron en solución salina tamponada con fosfato (PBS). Se recogió una única célula en cada pocillo de una placa de captura de 96 pocillos (AbGene Thermo-Fast 96 cat. nº 0600) por clasificación celular activada por fluorescencia (FACS), y la placa se congeló inmediatamente en hielo seco. La FACS se usó solamente para recoger células individuales y para rechazar células muertas y residuos basándose en la dispersión de la luz; no se usó ningún indicador de fluorescencia, y por lo tanto las células recogidas representarían una muestra aleatoria de la población.

La placa de captura celular contenía una única célula por pocillo en 5 µl de tampón *STRT* (Tris-HCl 20 mM pH 8,0, KCl 75 mM, $MgCl^2$ 6 mM, Tween-20 0,02 %) con STRT-T30-BIO 400 nM (5'-biotina-AAGCAGTGGTATCAACGCAGAGT_{30}VN-3'; este y todos los otros oligos fueron de Eurofins MWG Operon) y STRT-FW-n 400 nM (5'-AAGCAGTGGTATCAACGCAGAGTGGATGCTXXXXXrGrGrG-3', en la que "rG" indica un ribonucleótido guanina y "XXXXX" era un código de barras). Cada pocillo de la placa de captura contenía un oligo auxiliar de cambio de molde diferente (STRT-FW-n) con un código de barras distinto. Por ejemplo, el pocillo A01 recibió STRT-FW-1 con la secuencia 5'-AAGCAGTGGTATCAACGCAGAGTGGATGCTCAGAArGrGrG-3' que tenía una secuencia de código de barras CAGAA. Los 96 códigos de barras y las secuencias de oligo auxiliares se proporcionan en la Tabla 9.

Código de barras	Nombre del oligo	Secuencia (GGG) = ribonucleótidos
CAGAA	STRT-FW-1	AAGCAGTGGTATCAACGCAGAGTGGATGCTCAGAA(GGG)
CATAC	STRT-FW-2	AAGCAGTGGTATCAACGCAGAGTGGATGCTCATAC(GGG)
CAAAG	STRT-FW-3	AAGCAGTGGTATCAACGCAGAGTGGATGCTCAAAG(GGG)
CACAT	STRT-FW-4	AAGCAGTGGTATCAACGCAGAGTGGATGCTCACAT(GGG)
CATCA	STRT-FW-5	AAGCAGTGGTATCAACGCAGAGTGGATGCTCATCA(GGG)
CAGCC	STRT-FW-6	AAGCAGTGGTATCAACGCAGAGTGGATGCTCAGCC(GGG)
CACCG	STRT-FW-7	AAGCAGTGGTATCAACGCAGAGTGGATGCTCACCG(GGG)
CAACT	STRT-FW-8	AAGCAGTGGTATCAACGCAGAGTGGATGCTCAACT(GGG)
CAAGA	STRT-FW-9	AAGCAGTGGTATCAACGCAGAGTGGATGCTCAAGA(GGG)
CACGC	STRT-FW-10	AAGCAGTGGTATCAACGCAGAGTGGATGCTCACGC(GGG)
CATGT	STRT-FW-11	AAGCAGTGGTATCAACGCAGAGTGGATGCTCATGT(GGG)
CACTA	STRT-FW-12	AAGCAGTGGTATCAACGCAGAGTGGATGCTCACTA(GGG)
CAATC	STRT-FW-13	AAGCAGTGGTATCAACGCAGAGTGGATGCTCAATC(GGG)
CATTG	STRT-FW-14	AAGCAGTGGTATCAACGCAGAGTGGATGCTCATTG(GGG)
CAGTT	STRT-FW-15	AAGCAGTGGTATCAACGCAGAGTGGATGCTCAGTT(GGG)
CCTAA	STRT-FW-16	AAGCAGTGGTATCAACGCAGAGTGGATGCTCCTAA(GGG)
CCGAC	STRT-FW-17	AAGCAGTGGTATCAACGCAGAGTGGATGCTCCGAC(GGG)
CCAAT	STRT-FW-18	AAGCAGTGGTATCAACGCAGAGTGGATGCTCCAAT(GGG)
CCGCA	STRT-FW-19	AAGCAGTGGTATCAACGCAGAGTGGATGCTCCGCA(GGG)
CCTCC	STRT-FW-20	AAGCAGTGGTATCAACGCAGAGTGGATGCTCCTCC(GGG)
CCACG	STRT-FW-21	AAGCAGTGGTATCAACGCAGAGTGGATGCTCCACG(GGG)
CCAGC	STRT-FW-22	AAGCAGTGGTATCAACGCAGAGTGGATGCTCCAGC(GGG)
CCTGG	STRT-FW-23	AAGCAGTGGTATCAACGCAGAGTGGATGCTCCTGG(GGG)
CCGGT	STRT-FW-24	AAGCAGTGGTATCAACGCAGAGTGGATGCTCCGGT(GGG)
CCATA	STRT-FW-25	AAGCAGTGGTATCAACGCAGAGTGGATGCTCCATA(GGG)
CCGTG	STRT-FW-26	AAGCAGTGGTATCAACGCAGAGTGGATGCTCCGTG(GGG)
CCTTT	STRT-FW-27	AAGCAGTGGTATCAACGCAGAGTGGATGCTCCTTT(GGG)
CGAAA	STRT-FW-28	AAGCAGTGGTATCAACGCAGAGTGGATGCTCGAAA(GGG)
CGCAC	STRT-FW-29	AAGCAGTGGTATCAACGCAGAGTGGATGCTCGCAC(GGG)
CGGAG	STRT-FW-30	AAGCAGTGGTATCAACGCAGAGTGGATGCTCGGAG(GGG)
CGTAT	STRT-FW-31	AAGCAGTGGTATCAACGCAGAGTGGATGCTCGTAT(GGG)
CGCCA	STRT-FW-32	AAGCAGTGGTATCAACGCAGAGTGGATGCTCGCCA(GGG)
CGACC	STRT-FW-33	AAGCAGTGGTATCAACGCAGAGTGGATGCTCGACC(GGG)
CGTCG	STRT-FW-34	AAGCAGTGGTATCAACGCAGAGTGGATGCTCGTCG(GGG)
CGGCT	STRT-FW-35	AAGCAGTGGTATCAACGCAGAGTGGATGCTCGGCT(GGG)
CGGGA	STRT-FW-36	AAGCAGTGGTATCAACGCAGAGTGGATGCTCGGGA(GGG)
CGTGC	STRT-FW-37	AAGCAGTGGTATCAACGCAGAGTGGATGCTCGTGC(GGG)
CGAGG	STRT-FW-38	AAGCAGTGGTATCAACGCAGAGTGGATGCTCGAGG(GGG)
CGCGT	STRT-FW-39	AAGCAGTGGTATCAACGCAGAGTGGATGCTCGCGT(GGG)
CGTTA	STRT-FW-40	AAGCAGTGGTATCAACGCAGAGTGGATGCTCGTTA(GGG)
CGGTC	STRT-FW-41	AAGCAGTGGTATCAACGCAGAGTGGATGCTCGGTC(GGG)
CGCTG	STRT-FW-42	AAGCAGTGGTATCAACGCAGAGTGGATGCTCGCTG(GGG)
CGATT	STRT-FW-43	AAGCAGTGGTATCAACGCAGAGTGGATGCTCGATT(GGG)
CTCAA	STRT-FW-44	AAGCAGTGGTATCAACGCAGAGTGGATGCTCTCAA(GGG)
CTAAC	STRT-FW-45	AAGCAGTGGTATCAACGCAGAGTGGATGCTCTAAC(GGG)
CTTAG	STRT-FW-46	AAGCAGTGGTATCAACGCAGAGTGGATGCTCTTAG(GGG)
CTGAT	STRT-FW-47	AAGCAGTGGTATCAACGCAGAGTGGATGCTCTGAT(GGG)

Código de barras	Nombre del oligo	Secuencia (GGG) = ribonucleótidos
CTACA	STRT-FW-48	AAGCAGTGGTATCAACGCAGAGTGGATGCTCTACA(GGG)
CTGCG	STRT-FW-49	AAGCAGTGGTATCAACGCAGAGTGGATGCTCTGCG(GGG)
CTTCT	STRT-FW-50	AAGCAGTGGTATCAACGCAGAGTGGATGCTCTTCT(GGG)
CTTGA	STRT-FW-51	AAGCAGTGGTATCAACGCAGAGTGGATGCTCTTGA(GGG)
CTGGC	STRT-FW-52	AAGCAGTGGTATCAACGCAGAGTGGATGCTCTGGC(GGG)
CTCGG	STRT-FW-53	AAGCAGTGGTATCAACGCAGAGTGGATGCTCTCGG(GGG)
CTAGT	STRT-FW-54	AAGCAGTGGTATCAACGCAGAGTGGATGCTCTAGT(GGG)
CTGTA	STRT-FW-55	AAGCAGTGGTATCAACGCAGAGTGGATGCTCTGTA(GGG)
CTTTC	STRT-FW-56	AAGCAGTGGTATCAACGCAGAGTGGATGCTCTTTC(GGG)
CTATG	STRT-FW-57	AAGCAGTGGTATCAACGCAGAGTGGATGCTCTATG(GGG)
CTCTT	STRT-FW-58	AAGCAGTGGTATCAACGCAGAGTGGATGCTCTCTT(GGG)
GACAA	STRT-FW-59	AAGCAGTGGTATCAACGCAGAGTGGATGCTGACAA(GGG)
GAAAC	STRT-FW-60	AAGCAGTGGTATCAACGCAGAGTGGATGCTGAAAC(GGG)
GATAG	STRT-FW-61	AAGCAGTGGTATCAACGCAGAGTGGATGCTGATAG(GGG)
GAGAT	STRT-FW-62	AAGCAGTGGTATCAACGCAGAGTGGATGCTGAGAT(GGG)
GAACA	STRT-FW-63	AAGCAGTGGTATCAACGCAGAGTGGATGCTGAACA(GGG)
GAGCG	STRT-FW-64	AAGCAGTGGTATCAACGCAGAGTGGATGCTGAGCG(GGG)
GATCT	STRT-FW-65	AAGCAGTGGTATCAACGCAGAGTGGATGCTGATCT(GGG)
GATGA	STRT-FW-66	AAGCAGTGGTATCAACGCAGAGTGGATGCTGATGA(GGG)
GAGGC	STRT-FW-67	AAGCAGTGGTATCAACGCAGAGTGGATGCTGAGGC(GGG)
GACGG	STRT-FW-68	AAGCAGTGGTATCAACGCAGAGTGGATGCTGACGG(GGG)
GAAGT	STRT-FW-69	AAGCAGTGGTATCAACGCAGAGTGGATGCTGAAGT(GGG)
GAGTA	STRT-FW-70	AAGCAGTGGTATCAACGCAGAGTGGATGCTGAGTA(GGG)
GATTC	STRT-FW-71	AAGCAGTGGTATCAACGCAGAGTGGATGCTGATTC(GGG)
GAATG	STRT-FW-72	AAGCAGTGGTATCAACGCAGAGTGGATGCTGAATG(GGG)
GACTT	STRT-FW-73	AAGCAGTGGTATCAACGCAGAGTGGATGCTGACTT(GGG)
GCAAA	STRT-FW-74	AAGCAGTGGTATCAACGCAGAGTGGATGCTGCAAA(GGG)
GCCAC	STRT-FW-75	AAGCAGTGGTATCAACGCAGAGTGGATGCTGCCAC(GGG)
GCGAG	STRT-FW-76	AAGCAGTGGTATCAACGCAGAGTGGATGCTGCGAG(GGG)
GCTAT	STRT-FW-77	AAGCAGTGGTATCAACGCAGAGTGGATGCTGCTAT(GGG)
GCACC	STRT-FW-78	AAGCAGTGGTATCAACGCAGAGTGGATGCTGCACC(GGG)
GCTCG	STRT-FW-79	AAGCAGTGGTATCAACGCAGAGTGGATGCTGCTCG(GGG)
GCGCT	STRT-FW-80	AAGCAGTGGTATCAACGCAGAGTGGATGCTGCGCT(GGG)
GCGGA	STRT-FW-81	AAGCAGTGGTATCAACGCAGAGTGGATGCTGCGGA(GGG)
GCTGC	STRT-FW-82	AAGCAGTGGTATCAACGCAGAGTGGATGCTGCTGC(GGG)
GCAGG	STRT-FW-83	AAGCAGTGGTATCAACGCAGAGTGGATGCTGCAGG(GGG)
GCCGT	STRT-FW-84	AAGCAGTGGTATCAACGCAGAGTGGATGCTGCCGT(GGG)
GCTTA	STRT-FW-85	AAGCAGTGGTATCAACGCAGAGTGGATGCTGCTTA(GGG)
GCGTC	STRT-FW-86	AAGCAGTGGTATCAACGCAGAGTGGATGCTGCGTC(GGG)
GCCTG	STRT-FW-87	AAGCAGTGGTATCAACGCAGAGTGGATGCTGCCTG(GGG)
GCATT	STRT-FW-88	AAGCAGTGGTATCAACGCAGAGTGGATGCTGCATT(GGG)
GGTAA	STRT-FW-89	AAGCAGTGGTATCAACGCAGAGTGGATGCTGGTAA(GGG)
GGCAG	STRT-FW-90	AAGCAGTGGTATCAACGCAGAGTGGATGCTGGCAG(GGG)
GGAAT	STRT-FW-91	AAGCAGTGGTATCAACGCAGAGTGGATGCTGGAAT(GGG)
GGTCC	STRT-FW-92	AAGCAGTGGTATCAACGCAGAGTGGATGCTGGTCC(GGG)
GGACG	STRT-FW-93	AAGCAGTGGTATCAACGCAGAGTGGATGCTGGACG(GGG)
GGCCT	STRT-FW-94	AAGCAGTGGTATCAACGCAGAGTGGATGCTGGCCT(GGG)

Código de barras	Nombre del oligo	Secuencia (GGG) = ribonucleótidos
GGCGA	STRT-FW-95	AAGCAGTGGTATCAACGCAGAGTGGATGCTGGCGA(GGG)
GGAGC	STRT-FW-96	AAGCAGTGGTATCAACGCAGAGTGGATGCTGGAGC(GGG)

Tabla 9. Códigos de barras y secuencias oligonucleotídicas auxiliares.

La placa de captura celular se descongeló y se calentó después para lisar las células (20 ºC durante 5 minutos, 72 ºC durante 4 minutos, 10 ºC durante 5 minutos en un termociclador). Se añadieron 5 µl de mezcla de transcripción inversa (DTT 4 mM, dNTP 2 mM, Superscript II 5 U/µl en tampón de *STRT*) a cada pocillo y la placa se incubó (10 ºC durante 10 minutos, 42 ºC durante 45 minutos) para completar la transcripción inversa y el cambio de molde.

Para purificar el ADNc y retirar los cebadores que no habían reaccionado, se añadieron a cada pocillo 50 µl de PB (Kit de Purificación de PCR Qiaquick, Qiagen) las 96 reacciones se agruparon y se purificaron sobre una única columna de Qiaquick. El ADNc se eluyó en 30 µl de EB en un tubo de polialómero de 1,5 ml (Beckman).

La muestra de ADNc de 96 pocillos completa se amplificó en un único tubo en 100 µl de dNTP 200 µM, cebador de STRT-PCR 200 µM (5'-biotina-AAGCAGTGGTATCAACGCAGAGT-3'; Eurofins MWG Operon), Mezcla de ADN Polimerasa Advantage2 1x (Clontech) en tampón de PCR Advantage2 1x (Clontech) con 1 min. a 94 ºC seguido de 25 ciclos de 15 s a 95 ºC, 30 s a 65 ºC, 3 min. a 68 ºC, con tapa calentada. Se visualizó una alícuota en un E-gel de agarosa 1,2 % (Invitrogen) y la muestra se amplificó 1-5 ciclos adicionales si fue necesario. El producto se purificó (Kit de Purificación de PCR Qiaquick, Qiagen) y se cuantificó por fluorímetro (Qubit, Invitrogen). Las producciones típicas fueron de 0,5 - 1 µg total. Las alícuotas se tomaron en este estadio para análisis de micromatrices y Q-PCR.

Preparación de muestras para secuenciación de alto rendimiento

Se fragmentó ADNc amplificado por DNasa I en presencia de Mn^{2+}, lo que provoca una preferencia por roturas de doble cadena. Se fragmentaron 50 µl de ADNc en tampón de DNasa I complementado con $MnCl^2$ 10 mM y DNasa I diluida a 0,0003 U/µl en un volumen total de 120 µl durante exactamente seis minutos a temperatura ambiente. La reacción se detuvo mediante la adición de 600 µl de PB (Kit de Purificación de PCR Qiaquick, Qiagen), se purificó y se eluyó en 30 µl de EB en un tubo de polialómero (Beckman).

Se inmovilizaron fragmentos 3' y 5' en 30 µl de perlas paramagnéticas recubiertas con estreptavidina (Dynabeads MyOne CI, Invitrogen), después volvieron a suspenderse en 30 µl de tampón TaqExpress (Genetix, Reino Unido). Los extremos se repararon y se generaron salientes A individuales incubando las perlas en 40 µl de dNTP 200 µM, 0,25 U/µl. TaqExpress (Genetix, Reino Unido) en tampón de TaqExpress a 37 ºC durante 30 minutos, seguido de tres lavados en tampón NE 4 (New England Biolabs).

Se liberaron fragmentos 5' que contenían códigos de barras e insertos de ADNc de las perlas por digestión con BtsCI, y se ligaron simultáneamente adaptadores para generar una muestra adecuada para secuenciación en el Analizador del Genoma Illumina. Las perlas volvieron a suspenderse en 40 µl de ATP 1 mM, adaptador SOLEXA-ADP1 1 µM (5'-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT-3' y 3'-PHO-TTACTATGCCGCTGGTGGCTCTAGATGTGAGAAAGGGATGTGCTGCGAGAAGGCTA-PHO-5'), adaptador SOLEXA-ADP2 1 µM (5'-CAAGCAGAAGACGGCATACGAGCTCTTCCGATCT-3' y 3'-PHO-GTTCGTCTTCTGCCGTATGCTCGAGAAGGCTAG-PHO-5'), ADN ligasa T4 0,25 U/µl (Invitrogen), BtsCI 1 U/µl (New England Biolabs) en tampón NE 4 1x, y se incubó 30 minutos a 37 ºC. Las perlas se retiraron y el sobrenadante se purificó usando AmPure (Agencourt) y se eluyeron en 40 µl de EB (Qiagen).

La muestra se cargó en un E-gel SizeSelect 2 % y se recogió el intervalo de 200 - 300 pb. Se amplificó una alícuota de 4 µl en 50 µl de volumen total que contenía dNTP 200 µM, 400 nM de cada cebador (5'-AATGATACGGCGACCACCGA-3' y 5'-CAAGCAGAAGACGGCATACGAG-3') y polimerasa Phusion 0,15 U/µl en tampón HF Phusion (New England Biolabs) con 30 s a 98 ºC, 14-18 ciclos de [10 s a 98 ºC, 30 s a 65 ºC, 30 s a 72 ºC] seguido de 5 min. a 70 ºC. Se usaron las amplificaciones de ensayo para determinar el número mínimo de ciclos necesario. La muestra amplificada se purificó por Purificación por PCR Qiaquick seguido de un E-gel SizeSelect 2 % recogiendo de nuevo la región de 200 - 300 pb. La concentración se midió por Qubit (Invitrogen) y fue típicamente de 5 ng/µl. Las alícuotas se clonaron (TOPO, Invitrogen) y se secuenciaron por secuenciación de Sanger para verificar la calidad de la muestra y determinar la longitud de fragmento promedio. Basándose en esta información, la concentración molar pudo determinarse con precisión y fue en general por encima de 10 nM. Se realizó formación de grupos y secuenciación por síntesis en un Analizador de Genoma *IIx* de acuerdo con los protocolos del fabricante (Illumina, Inc., San Diego, Estados Unidos) en un proveedor de servicios comercial (Fasteris SA, Ginebra. Suiza).

Mapeo, cuantificación y visualización

Se clasificaron lecturas sin procesar por código de barras (primeras cinco bases) y se recortaron para retirar hasta cinco G 5' introducidas por cambio de molde, y A 3' que aparecieron en ocasiones cuando una lectura se extendió hasta la cola de poli(A). Solamente se permitieron códigos de barras exactos, y los códigos de barras se diseñaron para que ningún error individual convirtiera un código de barras convirtiera un código de barras válido en otro. Las lecturas se mapearon después en el genoma de ratón usando Bowtie (Langmead *et al.*, Genome Biol. 10: R25 (2009)) con los ajustes por defecto. Se descartaron las lecturas no mapeadas. Después, para cada característica anotada en el ensamblaje de NCBI 37.1, todas las lecturas de mapeo se contaron para generar un recuento sin procesar. Es decir, todas las lecturas que se mapearon en cualquiera de los exones de un gen se asignaron a ese gen, las isoformas no se distinguieron. Finalmente, las lecturas sin procesar para cada célula se normalizaron a transcritos por millón (t.p.m.). Los pocillos con menos de 1000 lecturas mapeadas se omitieron de análisis adicional; supuestamente estas incluían casos en los que el instrumento de FACS no había conseguido acertar con la gota de reactivo mientras se seleccionaban las células.

Para visualizar células en un paisaje bidimensional, se calcularon en primer lugar todas las similitudes por pares. Se usó la distancia de Bray-Curtis como una métrica de similitud porque tendía a manejar bien el ruido en genes poco expresados. La correlación convencional produjo resultados similares, pero con algunas células descolocadas más (datos no mostrados). Después se construyó un gráfico de similitud dejando que los nodos representen células, y conectando cada célula con sus cinco células más similares (para mayor claridad, las células con menos de 10.000 lecturas se omitieron, ya que eran susceptibles de generar bordes engañosos). Por lo tanto cada nodo (célula) tuvo cinco bordes salientes y diversos números de bordes entrantes. Después se usó una distribución dirigida por fuerza para proyectar la gráfica en dos dimensiones, revelando la estructura interna basada en similitudes célula-célula. La función GraphPlot del programa Mathematica (Wolfram Research Inc., Estados Unidos) se usó con la opción "Inclusión Eléctrica en Resorte".

Resultados

Se presentan datos de 436 células individuales recogidas de tres tipos celulares de ratón diferentes: células madre embrionarias (ES RI, Wood *et al.*, Nature 365: 87-89 (1993)), una línea celular tumoral de neuroblastoma (Neuro-2A, Olmsted *et al.*, Proc. Nati. Acad. Sci U.S.A. 65: 129-136 (1970)) y fibroblastos embrionarios (MEF). Brevemente, cada muestra se preparó seleccionando células individuales por separación de células activadas por fluorescencia (FACS) en los pocillos de una placa de PCR de 96 pocillos precargada con tampón de lisis; calentando la placa hasta completar la lisis, añadiendo después reactivos de transcripción inversa para generar un ADNc de primera cadena. Para incorporar un código de barras específico de pocillo (y por lo tanto específico de célula), se usó el mecanismo de cambio de molde de transcriptasa inversa (Schmidt *et al.*, Nucleic Acids Res. 27: e31 (1999)) por el que un oligo auxiliar dirige la incorporación de una secuencia específica en el extremo 3' de la molécula de ADNc (**Figura 10A**). Se usó oligo auxiliar diferente en cada pocillo, con códigos de barras de cinco bases distintos y secuencia cebadora universal. Después de la síntesis de ADNc, las 96 reacciones se agruparon, se purificaron y se amplificaron por PCR de un único cebador en un único tubo. Por lo tanto se redujo la desviación de amplificación de célula a célula, y el número de ciclos de PCR se pudo mantener bajo ya que la amplificación comenzaba a partir de 96 veces más material. Las muestras amplificadas se adaptaron después para secuenciación usando métodos convencionales. El procedimiento se nombró "STRT" (transcripción inversa marcada en célula individual). Para más detalles, véase la sección de Métodos y la **Figura 11**.

Se obtuvieron típicamente 5-12 millones de lecturas sin procesar por carril de secuenciación en un Analizador de Genoma Illumina IIx, y cada muestra se analizó en hasta ocho carriles (pero típicamente uno o dos). Se retiraron las lecturas que carecían de un código de barras apropiado, principalmente provocado por errores en la preparación de muestras tales como adaptadores ligados erróneamente. De las 79±11 % (media±d.t) lecturas, aproximadamente tres cuartos (75±12 %) podrían colocarse en el genoma de ratón permitiendo hasta dos errores de secuenciación, dando como resultado aciertos hasta 14.718±3.006 características distintas (incluyendo ARNm, ARN mitocondrial y repeticiones expresadas). Estos resultados se resumen en la Tabla 10.

Bibliotecas	*Especies*	*Fuente celular*	*Carriles*	*Longitud (pb)*	*Lecturas sin procesar*	*Con código de barras*	*Mapeadas*	*Características*	Características acumuladas
L006	Ratón	ESR1	8	1x36 7x50	46.130.389	38.284.613(83 %)	21.356.066(56 %)	13.168	13.168
L013	Ratón	ESR1	1	36	6.482.712	5.328.081 (82 %)	3.199.050(60 %)	19.328	21.285
L019A	Ratón	ESR1	1	36	4.302.649	2.588.536 (60 %)	2.344.394(91 %)	12.229	
L019B	Ratón	ESR1 (*)	1	36	5.134.171	3.323.243 (65 %)	2.562.559 (77 %)	14.794 (**)•	23.152
L023	Ratón	Neuro2A	1	50	10.062.086	8.588.048 (86 %)	5.963.143(69 %)	11.464	23.735
L026A	Ratón	MEF	1	36	4.278.888	3.729.229 (87 %)	2.862.951 (77 %)	12.924	
L026B	Ratón	MEF(*)	1	36	3.912.144	3.398.519(87 %)	2.599.074 (76 %)	13.832 (**)	24.288

(*) estas muestras se amplificaron independientemente de bibliotecas “A” correspondientes, pero no se controló su calidad por secuenciación de Sanger en la preparación
(**) acumulativo con la muestra previa (es decir A y B juntos)

Tabla 10. Sumarios de bibliotecas. Los porcentajes se refieren a porcentaje de la columna precedente a no ser que se indique de otro modo. Las “características” incluyen ARNm, ARNm mitocondrial y familias de retrotransposones expresadas. Las “características acumuladas” siguen el número acumulado de características entre todos los lib.

Los aciertos abarcaron algunos transcritos (**Figura 10B**), pero se localizaban más habitualmente en una región de aproximadamente 200 - 1500 pb desde el extremo 3' de cada gen, como se ilustra en la **Figura 10C**. Esto era esperable porque se usó un cebador de oligo(dT) para generar ADNc a partir del extremo 3' y el ARNm probablemente se degradaba parcialmente durante la lisis celular por hidrólisis de alta temperatura en presencia de Mg^{2+}. El fondo de por ejemplo contaminación de ADN genómico, valorado por aciertos para la secuencia no anotada, fue mínimo ($<10^{-3}$ lecturas por millón por kilobase), como se ve claramente en la **Figura 10B** y **Figura 10C**. Obsérvese la escasez de aciertos para la cadena inversa y para intrones tanto para Pou5f1 como para Nanog; fueron similares otros loci.

Escrutando los datos mapeados, no se encontraron pruebas de colocación de cebadores errónea u otras reacciones secundarias indeseadas. Los experimentos de control mostraron que ARN, transcriptasa inversa y oligo de cambio de molde se requerían individualmente para producir productos (datos no mostrados). La amplia mayoría de lecturas mapeadas tuvieron un código de barras orientado apropiadamente, lo que indica que se iniciaron desde el cebador de oligo dT y se cambiaron de molde correctamente. No se encontraron pruebas de un motivo complementario de ninguno de los cebadores cerca de los sitios de mapeo de lectura o de hecho de ningún otro motivo, excepto por una desviación de T general débil en pocos casos (**Figura 12**). Por otro lado, hubo frecuentes puntos calientes de cambio de molde (**Figura 13**), que indican restricciones estructurales en el ARN que afectan a la síntesis de ADNc y/o cambio de molde en sitios particulares. Los puntos calientes fueron uniformes entre células, y por lo tanto no deberían afectar a comparaciones cuantitativas.

Para caracterizar la complejidad de las muestras, y para determinar la profundidad de secuenciación requerida para tomar muestras de la mayoría de complejidad disponible, se estudió la tasa de "nuevo descubrimiento" en función de la profundidad de lectura. En otras palabras, se determinó el número de moléculas nuevas, distintas, que se descubrieron a medida que se añadieron más secuencias. Debería observarse que, como máximo se generó un clon amplificable de cada molécula de ARN poliadenilada y este clon después se amplificó y secuenció a partir de su extremo 5'. Por lo tanto las lecturas que se mapeaban en distintas localizaciones deben haberse generado de distintas moléculas de ARNm. Por otro lado, las lecturas que se mapean en la misma localización pueden haberse generado por coincidencia a partir de dos moléculas de ARNm, o pueden representar copias del clon inicial de muestra. El número de lecturas de mapeo distinto fue por lo tanto un límite inferior en la verdadera complejidad de muestras. Como se muestra en la **Figura 14A**, ninguna de las muestras presentadas en el presente documento se secuenciaron hasta la saturación, incluso a 21 millones de lecturas mapeadas, pero la producción de nuevas moléculas tendió a ralentizarse después de 5-10 millones de lecturas. Proyectando las curvas hasta profundidad de lectura infinita, la mayoría de las muestras parecerían contener al menos tres millones de moléculas distintas, o aproximadamente 30.000 por célula. Suponiendo 10^5 – 10^6 moléculas de ARNm por célula, esto sugeriría que el método convirtió con éxito 3-30 % del ARNm en lecturas mapeables, como un límite inferior (muy) conservativo. La verdadera proporción fue probablemente significativamente mayor, ya que se generarían muchas lecturas coincidentes (descartadas en este análisis) a partir de puntos calientes de cambio de molde, como se ha mencionado anteriormente.

Por el contrario, la tasa de descubrimiento de características distintas se redujo rápidamente, y el 86 % de todas las características distintas se detectan en el Primer 14 % de lecturas (**Figura 14B**). Esto sugiere que el método recuperó con éxito la mayoría de características expresadas presentes en las muestras, incluso a profundidad de lectura relativamente baja.

Se requiere con frecuencia información de cadena para asignar apropiadamente lecturas a unidades transcripcionales, ya que los genes frecuentemente solapan en cadenas opuestas. Por ejemplo, más de 3000 genes humanos solapan de esta manera (Yelin *et al.*, Nat. Biotechnol. 21: 379-386 (2003)). Debido a que el mecanismo de cambio de molde usado para introducir un código de barras sucede direccionalmente, el número de cadenas podría conservarse durante todo el protocolo. Para confirmar esto, se examinó el genoma mitocondrial, que se expresa como un único transcrito largo de una cadena (la cadena H) y se corta posteriormente para escindir transcritos de ARNt localizados entre genes codificantes de proteínas. Solamente se poliadenilan después genes codificantes de proteínas. Se genera un único transcrito codificante de proteína, ND6, de la cadena L, pero se expresa muy débilmente y está irregularmente poliadenilado (Slomovic *et al.*, Mol. Cell. Biol. 25: 6427-6435 (2005)). Como se muestra en la **Figura 10D**, se observó una especificidad de cadena muy fuerte (> de 99 % de lecturas en la cadena H) y no se detectó expresión significativa de genes de ARNt, lo que confirma que el método era específico de poli(A). El número pequeño de aciertos en la cadena L se produjo principalmente cerca del promotor de cadena L, lo que puede explicarse por la poliadenilación de transcritos de cadena L abortados (Slomovic *et al.*, mencionado anteriormente). De forma similar, la expresión aparente de ND6 en la cadena errónea se explica probablemente por la poliadenilación natural de ND5 cadena debajo de su fase abierta de lectura (Slomovic *et al.*, mencionado anteriormente). La especificidad de cadena permitió asignar de forma inequívoca lecturas a transcritos expresados, incluso en casos en los que se coexpresaron dos genes solapantes (**Figura 15**).

En la escala mayor de los cromosomas nucleares, los aciertos se distribuyeron aproximadamente igual en las cadenas directa e inversa. La densidad de lectura se correlacionó fuertemente con la densidad génica como se muestra para el cromosoma 19 en la **Figura 10E**, lo que indica de nuevo que la mayoría de las lecturas se originaron

específicamente de transcritos expresados y se mapearon con precisión en el genoma.

Para generar una medida cuantitativa de la expresión génica, se contó el número de aciertos para cada característica anotada, normalizado a transcritos por millón (t.p.m.). Suponiendo 10^5 a 10^6 transcritos por célula, 1 a 10 t.p.m. corresponde a una única molécula de ARNm por célula. No se usó la longitud del transcrito (como en la medida de RPKM (Mortazavi *et al.*, Nat. Methods 5: 621-628 (2008)) para normalizar porque se generó una única molécula de extremo 3' amplificable para cada molécula de ARNm introducida, independientemente de su longitud. Una ventaja de este enfoque fue la falta de desviación frente a transcritos cortos (de los que deben tomarse muestras con más profundidad para generar un valor de RPKM detectable) o transcritos largos (que podrían de otro modo suprimirse durante la PCR). De hecho, y en contraste con ARN-Sec convencional Oshlack *et al.*, Biol. Direct 4: 14 (2009)), no se observó desviación dependiente de la longitud para transcritos mayores de 800 nucleótidos (**Figura 16**). Los transcritos menores de aproximadamente 200 nucleótidos se detectaron menos, probablemente debido a que solamente se seleccionaron en gel muestras por encima de 100 pb. Adicionalmente, los transcritos de aproximadamente 600 nucleótidos estaban ligeramente sobrerrepresentados, posiblemente debido a la mayor eficacia de cambio de molde en el extremo 5' de ARNm (Schmidt *et al.*, Nucleic Acids Res. 27: e31 (1999)) o debido a la presencia de algunos genes muy altamente expresados en este intervalo (por ejemplo Dppa5 y Rps14).

Los niveles de expresión abarcaron cuatro órdenes de magnitud en células individuales (aproximadamente 1 – 10.000 t.p.m.), expresados la mayoría de los genes a niveles bajos (<100 t.p.m.; **Figura 17A**). Dada la profundidad relativamente superficial de la secuenciación usada en el presente documento, los genes expresados por debajo de 10 t.p.m. fueron generalmente indetectables debido solamente al límite de toma de muestras. Ya que se agrupó ADNc de una única célula antes de la amplificación, las producciones de diferentes células no pudieron normalizarse posteriormente. Como consecuencia, se tomaron muestras desiguales de las células y varió el límite de detección. Por ejemplo, compárense dos células con muestras tomadas a 500.000 lecturas (**Figura 17B**) y 100.000 lecturas (**Figura 17C**). En el primer caso, el límite de detección aparente fue aproximadamente 10 t.p.m., mientras que en el segundo caso los genes por debajo de 100 t.p.m. generalmente no se detectaron. Sin embargo, en ambos casos, los genes por encima del límite de detección se cuantificaron de forma reproducible en células individuales (el coeficiente de variación fue de 46 % a 500.000 lecturas; y 72 % a 100.000 lecturas). La extensión de este análisis a todas las células y genes mostró que la sensibilidad abordaba el límite teórico impuesto por la profundidad de toma de muestras (**Figura 17D**); la diferencia puede explicarse por pérdidas en transcripción inversa, cambio de molde y manipulación de muestras. Los niveles de expresión medidos fueron generalmente precisos, como se determinó por comparación con Q-PCR (**Figura 17E**), e hibridación de micromatrices (**Figura 18**). De acuerdo con informes publicados basados en Q-PCR (Bengtsson *et al.*, Genome Res. 15: 1388-1392 (2005)), la abundancia de ARNm de Actb mostró una distribución aproximadamente normal logarítmica entre células (**Figura 19**). Se expresión ARN polimerasa II (subunidad grande) a 25±123 t.p.m. en células ES, comparable a las 27 RPKM halladas por ARN-Sec (Cloonan *et al.*, Nat. Methods 5: 613-619 (2008)) y a las 33±79 t.p.m. halladas en células CHO por detección directa in situ (suponiendo 300.000 transcritos por célula) (Raj *et al.*, PLoS Biol 4: 309 (2006)).

Se visualizaron las relaciones célula-célula en un mapa bidimensional, de modo que las células más estrechamente relacionadas se localizarían cerca entre sí. De esta manera, los tipos celulares basados solamente en los datos de expresión pudieron detectarse y distinguirse, sin basarse en marcadores preexistentes. Un análisis de componentes principales convencional (PCA) reveló tres grupos distintos de células, como se esperaba (**Figura 20**). Sin embargo, se consiguió una separación más completa en grupos de tipos celulares distintos usando un método basado en gráfica (véase **Métodos**). Brevemente, se construyó un gráfico con nodos que representaban células, y bordes que representaban similitud de patrón de expresión célula-célula (**Figura 21A**). Se usó una distribución dirigida por fuerza para proyectar la gráfica en dos dimensiones. En este caso de ensayo usando solamente dos tipos celulares (células ES y Neuro2A), se consiguió una separación casi perfecta (**Figura 21B**). Un mapa mayor que incorporaba MEF y células ES adicionales mostró buena separación (**Figura 21C**); lo que demuestra que los perfiles de expresión de células individuales contenían suficiente información para distinguir tipos celulares de novo. El análisis tanto de PCA como basado en gráfico distinguió claramente los tipos celulares ensayados en el presente documento, pero el método basado en gráfico generó grupos más homogéneos, bien separados. Ambos métodos agruparon con precisión células ES preparadas independientemente juntas distintas de los otros tipos celulares, lo que muestra que los grupos no representaban artefactos de preparación de muestras.

Los datos de expresión génica se proyectaron en el mapa, lo que proporcionó un modo fácil de entender rápidamente los patrones de expresión génica en ambas células individuales y en los grupos que representan tipos celulares (**Figura 22**). Se expresaron claramente de forma específica un conjunto de marcadores de células ES bien conocidos (Dppa5, Sox2, Sal14, PouSf1, Nanog, Zfp42, Zic3, Esrrb) en células ES, aunque sus niveles de expresión variaron ampliamente entre células (obsérvese la escala de color logarítmica). Se expresaron más ampliamente algunos genes importantes para pluripotencialidad (Klf4, Myc y Klf2). La potencia del análisis de células individuales a gran escala resultó evidente en el hecho de que aunque no todas las células expresaron todos los marcadores, los patrones de actividad génica fueron altamente uniformes al nivel de grupo. Por ejemplo, incluso un gen citoesquelético altamente expresado como Actb no se detectó siempre, pero su expresión en cada uno de los tres grupos principales fue evidente. Consecuentemente, los factores de transcripción menos expresados característicos de células ES no se detectaron en algunas células ES individuales, pero el patrón global de expresión en el grupo de

células ES fue inequívoco y coherente con su identidad como células ES. En general, a medida que los niveles de expresión promedio se redujeron de 45 000 t.p.m. (mt_Rnr2) a 1700 (Actb), 850 (Rp14), 73 (KLras) y 0 t.p.m. (Calb1), el número de células de expresión también se redujo, lo que refleja la naturaleza estocástica de la expresión génica así como los límites de la sensibilidad del método.

La representación del mapa celular demostró que (1) las células individuales mostraban patrones de expresión altamente variables, pero su patrón global de expresión era suficiente para agrupar células de un tipo juntas como un grupo; (2) una vez que se formó un grupo de células, que representa un tipo celular definido, los patrones de expresión génica (al nivel de grupo) fueron inequívocos. Por lo tanto, el perfil de expresión de células individuales al azar es una estrategia eficaz para acceder a datos de expresión de una única célula en poblaciones heterogéneas de células.

Análisis

Se describe en el presente documento un método fiable y preciso para obtener perfiles de transcripción de ARN-Sec de cientos de células individuales, y se muestra que pueden usarse perfiles de expresión de células individuales para formar grupos específicos de tipos celulares. Esto permite el análisis de patrones específicos de tipos celulares de expresión génica tanto al nivel de células individuales como al nivel de población, sin la necesidad de marcadores conocidos o incluso un conocimiento previo de que existe un cierto tipo celular. Esa estrategia general puede extenderse para estudiar todos los tipos de muestras mixtas. Por ejemplo, podría aplicarse para controlar la aparición de tipos celulares específicos durante la organogénesis, sin la necesidad de purificar esos tipos celulares usando marcadores de superficie celular. De forma similar, podría usarse para estudiar poblaciones pequeñas de células madre incluidas en tejidos adultos, tales como las células madre que mantienen las criptas intestinales. El método también podría aplicarse a enfermedad, incluyendo la caracterización de muestras celulares de tumores heterogéneas o las células cancerosas en circulación poco habituales que pueden contribuir a la metástasis.

Lo que une todas estas líneas de investigación científica dispares es la necesidad de separar poblaciones heterogéneas de células. En la actualidad, la separación se consigue principalmente por aislamiento físico de las células basándose en marcadores de superficie celular conocidos, o por marcaje genético de las células deseadas de modo que puedan aislarse basándose, por ejemplo, en la expresión de GFP. Sin embargo, el uso de marcadores previamente conocidos evita el descubrimiento de nuevos tipos celulares, y siempre produce el riesgo de dar como resultado datos mixtos si los marcadores no eran verdaderamente específicos. Por el contrario, los métodos descritos en el presente documento han mostrado que pueden separarse células de distintos tipos simplemente por ordenador, siempre que se generen grandes números de perfiles de expresión de células individuales.

Resulta importante, por lo tanto, que se requiere un método escalable, de muy alto rendimiento, para realización de perfiles de expresión de células individuales. Por lo tanto, se desarrolló un método para preparar una muestra de ADNc de una célula individual con código de barras a partir 96 células en una única etapa de incubación. Como consecuencia, se pudieron agrupar 96 células y tratarse como una única muestra a lo largo del procedimiento, lo que aumentó en gran medida el rendimiento y redujo el coste. También se puede reducir el desvío de amplificación, ya que las 96 células se amplificaron en un único tubo cerrado. El procedimiento completo tardó dos días en realizarse, de 96 células vivas a muestras finalizadas cargadas en el analizador de genoma. El coste, incluyendo todos los reactivos y consumibles para generar 10-15 millones de lecturas de 36 pb usando servicios comerciales, fue de aproximadamente 3500 $ (es decir, aproximadamente 35 $/célula).

Los datos generados en el presente documento fueron en un gran número de células individuales, cada una analizada a una profundidad relativamente superficial de cobertura. Esto permitió la generación de datos en muchas más células individuales de lo que se ha presentado nunca en un único estudio (no se ha publicado ningún experimento de transcriptoma de células individuales con más de una docena de células), y producir un mapa celular con alta resolución. De hecho, siempre que se tome muestra de cada célula con suficiente profundidad para agrupar correctamente, tendría con frecuencia más sentido analizar un gran número de células que analizar cada célula con más profundidad. Cuantas más células se añadan, más precisos serán los datos agregados obtenidos de cada tipo celular distinto (grupo), y mejor será la resolución en el “espacio de tipo celular”. Por ejemplo, se tomaron muestras de muchas de las células ES del presente documento a menos de 100.000 lecturas/célula, pero en total se identificaron 160 células ES en el mapa celular, lo que comprende más de 1,5 millones de lecturas. La toma de muestras de un gran número de células será especialmente importante cuando el enfoque se aplique a tejidos complejos, en los que algunos tipos de células pueden estar presentes solamente en una pequeña minoría. Además, a medida que los costes de secuenciación continúan reduciéndose, el balance entre el número de células y el número de lecturas será menos apremiante.

Se prevé el uso de realización de perfiles transcripcionales de células individuales a muy gran escala para construir un mapa detallado de tipos celulares de origen natural, lo que proporcionaría acceso sin precedentes a la maquinaria genética activa en cada tipo de célula en cada estadio del desarrollo. Puede usarse la misma estrategia para diseccionar la heterogeneidad mutacional de neoplasias al nivel de células individuales.

REIVINDICACIONES

1. Un método para preparar una biblioteca de ADNc a partir de una pluralidad de células individuales, comprendiendo el método las etapas de:

(i) liberar ARNm de cada célula individual para proporcionar una pluralidad de muestras de ARNm individuales, en las que el ARNm en cada muestra de ARNm individual es de una única célula;
(ii) sintetizar una primera cadena de ADNc a partir del ARNm en cada muestra de ARNm individual e incorporar un marcador definido o una combinación de marcadores definida en cada muestra de ADNc individual para proporcionar una pluralidad de muestras de ADNc marcadas, en la que cada muestra de ADNc tiene un marcador o una combinación de marcadores definido, en la que el ADNc en cada muestra de ADNc marcada es complementario al ARNm de una única célula;
(iii) agrupar las muestras de ADNc marcadas; y
(iv) amplificar las muestras de ADNc agrupadas para generar una biblioteca de ADNc que comprende ADNc bicatenario.

2. El método de acuerdo con la reivindicación 1, en el que en la etapa (ii) el marcador se incorpora en el ADNc durante su síntesis.

3. El método de acuerdo con la reivindicación 1 ó 2, en el que la síntesis de la primera cadena de ADNc en la etapa (ii) se dirige por un cebador de síntesis de ADNc (CDS) que incluye una secuencia complementaria de ARN (RCS) que es al menos parcialmente complementaria de uno o más ARNm en una muestra de ARNm individual.

4. El método de acuerdo con cualquiera de las reivindicaciones 1-3, en el que la RCS es al menos parcialmente complementaria a una parte de la primera cadena de ADNc, de modo que sea capaz de dirigir la síntesis de una segunda cadena de ADNc usando la primera cadena de ADNc como molde, o
en el que se añade un oligonucleótido de cambio de molde (TSO) a cada muestra de ARNm individual, en el que dicho TSO comprende una parte que es al menos parcialmente complementaria a una parte en el extremo 3' de la primera cadena de ADNc, en el que opcionalmente el CDS o el TSO incluye un marcador, en el que preferentemente el marcador es una secuencia de nucleótidos de 4-20 nucleótidos de longitud, o
en el que opcionalmente tanto el CDS como el TSO incluyen un marcador, en el que preferentemente el CDS y el TSO incluyen cada uno un marcador diferente, de modo que la muestra de ADNc marcada comprenda una combinación de marcadores.

5. El método de acuerdo con cualquiera de las reivindicaciones 1-4, en el que la primera cadena de ADNc incluye una parte 3' que comprende una pluralidad de citosinas o análogos de citosina que forman pares de bases con guanosina, en el que opcionalmente el TSO incluye una parte 3' que comprende una pluralidad de guanosinas o análogos de guanosina que forman pares de bases con citosina, en el que preferentemente las guanosinas son ribonucleósidos o monómeros de ácido nucleico bloqueados.

6. El método de acuerdo con cualquiera de las reivindicaciones 3-5, en el que el CDS comprende una secuencia cebadora de amplificación en 5' (APS) y una RCS en 3'.

7. El método de acuerdo con la reivindicación 6, en el que la RCS en 3' comprende un oligo(dT), una secuencia específica de familia génica, una secuencia aleatoria o una secuencia semialeatoria no autocomplementaria, y/o en el que el TSO incluye una APS en 5', en el que opcionalmente el CDS y la APS en 5' del TSO es al menos un 80 % idéntica a la APS en 5' del CDS, o en la que el CDS y la APS en 5' del TSO es 100 % idéntica a la APS en 5' del CDS.

8. El método de acuerdo con cualquiera de las reivindicaciones 1-7, en el que las células se lisan para liberar ARNm y/o en el que el ARNm se purifica después de la etapa (i), y/o en el que la síntesis de ADNc a partir de ARNm se detiene antes de agruparse las muestras de ADNc marcadas, y/o en el que las muestras de ADNc marcadas se purifican antes de la amplificación del ADNc.

9. El método de acuerdo con la reivindicación 1, en el que en la etapa (iv) las muestras de ADNc agrupadas se amplifican por PCR, en el que opcionalmente las muestras de ADNc agrupadas se amplifican por PCR de emulsión, en el que preferentemente las muestras de ADNc agrupadas se amplifican por PCR de un único cebador.

10. El método de acuerdo con cualquiera de las reivindicaciones 1-9, en el que el método comprende además procesar la biblioteca de ADNc para obtener una biblioteca adecuada para secuenciación.

11. El método de acuerdo con la reivindicación 10, en el que el procesamiento comprende fragmentar la biblioteca de ADNc y/o en el que el procesamiento incluye la etapa de añadir un adaptador al ADNc y/o en el que se amplifica la biblioteca de ADNc.

12. Una biblioteca de ADNc producida por el método de cualquiera de las reivindicaciones 1 a 11.

13. Un método para analizar la expresión génica en una pluralidad de células individuales, comprendiendo el método las etapas de:

(i) preparar una biblioteca de ADNc de acuerdo con el método de cualquiera de las reivindicaciones 1 a 11; y
(ii) secuenciar la biblioteca de ADNc.

14. El método de acuerdo con la reivindicación 13, en el que la secuenciación es por secuenciación al azar, en el que opcionalmente la biblioteca de ADNc se secuencia para obtener al menos 10.000, al menos 1 millón, al menos 10 millones, al menos 100 millones, o al menos 1000 millones de lecturas, en el que una lectura es una longitud de ácido nucleico continuo obtenida por una reacción de secuenciación.

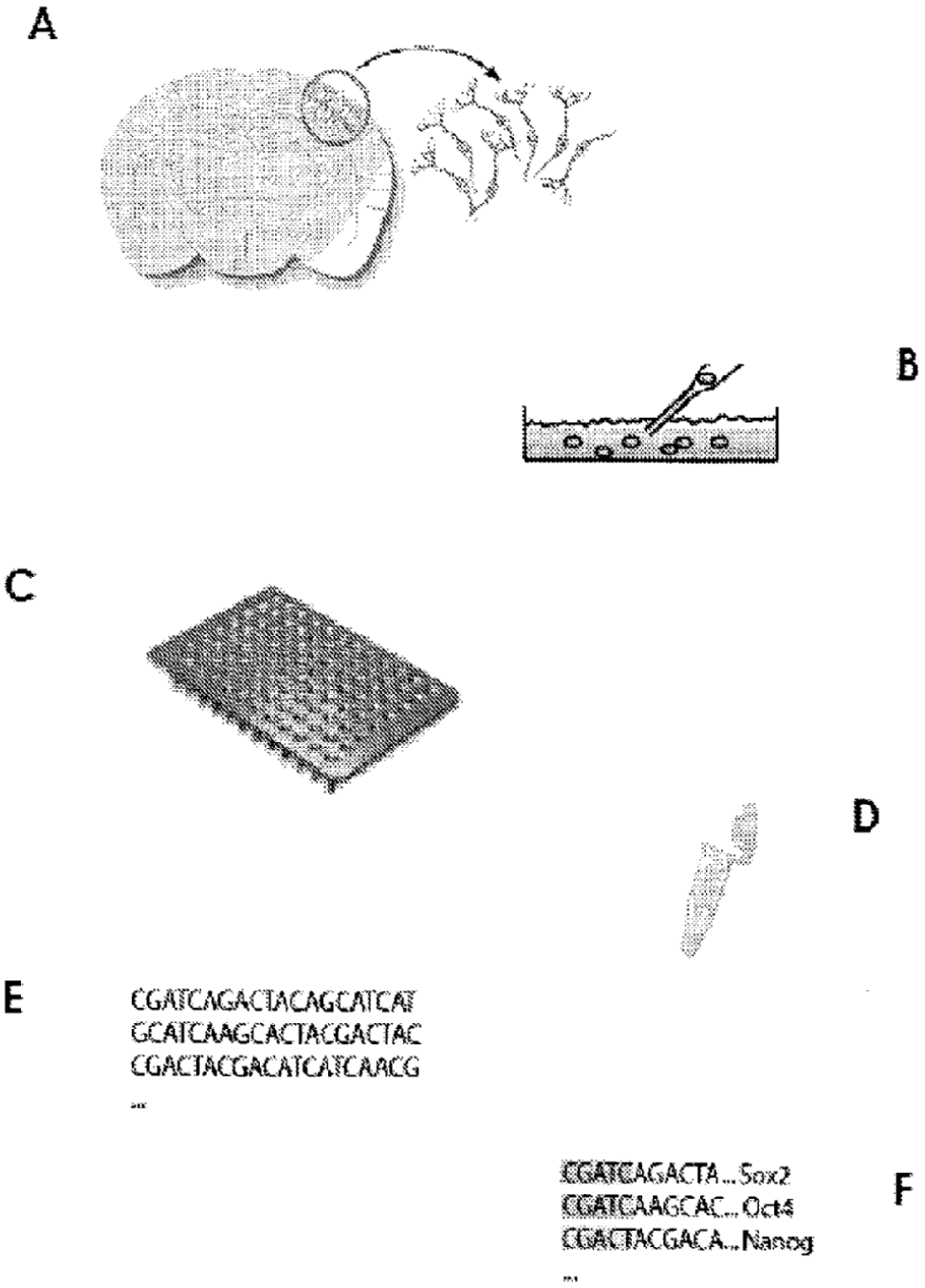
A
B
C
D
E
CGATCAGACTACAGCATCAT
GCATCAAGCACTACGACTAC
CGACTACGACATCATCAACG
...
CGATCAGACTA...Sox2
CGATCAAGCAC...Oct4
CGACTACGACA...Nanog
...
F

Figura 1

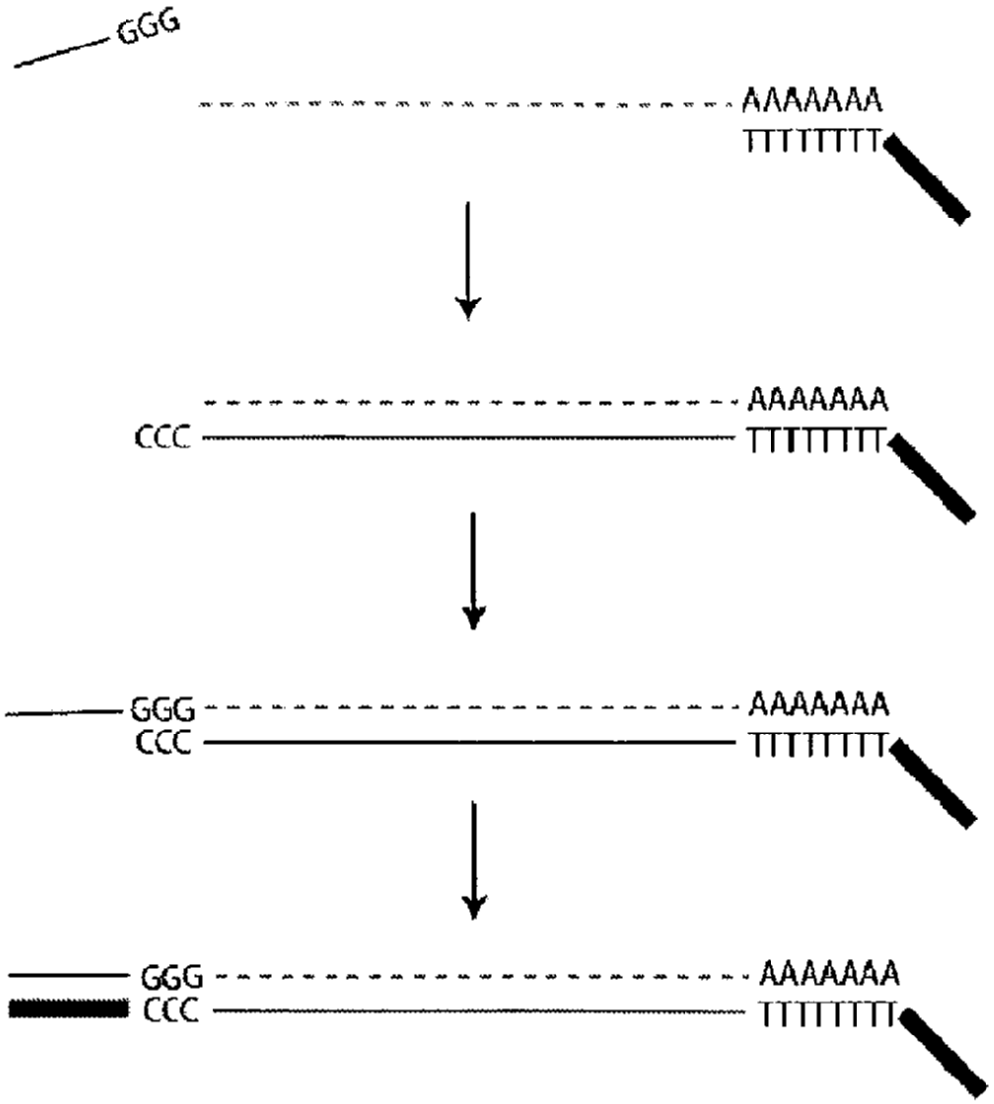
GGG
AAAAAAA
TTTTTTTT
AAAAAAA
CCC
TTTTTTTT
GGG
AAAAAAA
CCC
TTTTTTTT
GGG
AAAAAAA
CCC
TTTTTTTT

Figura 2

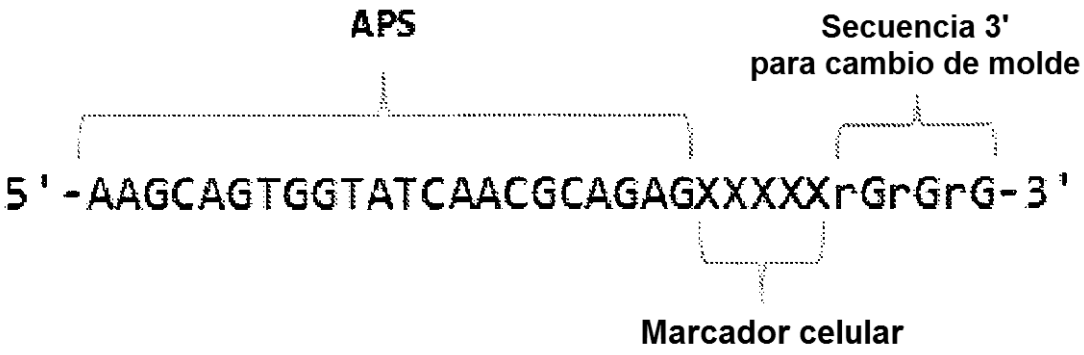
APS
Secuencia 3'
para cambio de molde
5'-AAGCAGTGGTATCAACGCAGAGXXXXXrGrGrG-3'
Marcador celular

Figura 3

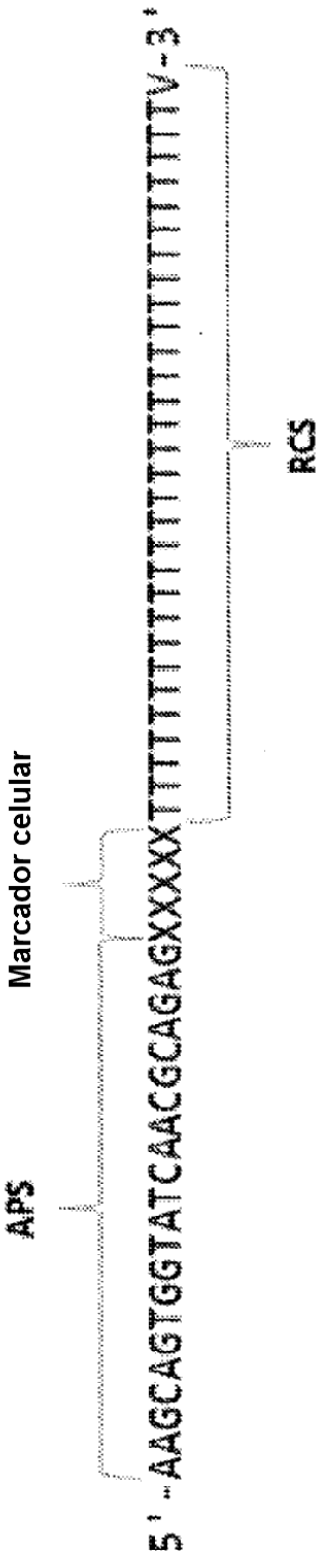
APS
Marcador celular
5' - AAGCAGTGGTATCAACGCAGAGXXXXTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTV - 3'
RCS

Figura 4

Figura 5

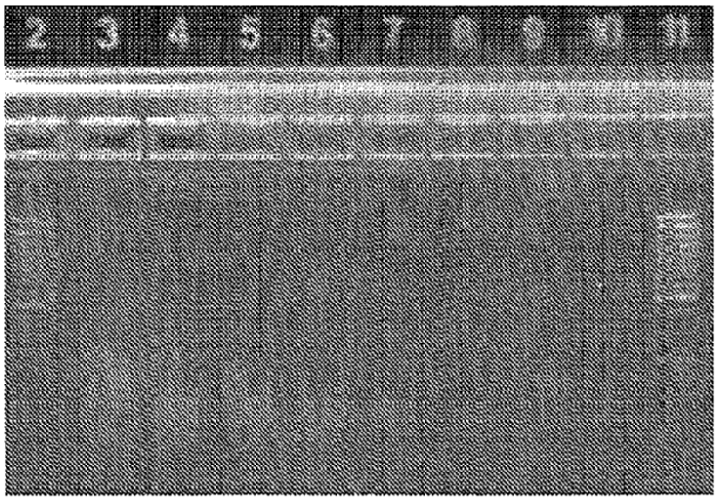

Figura 6

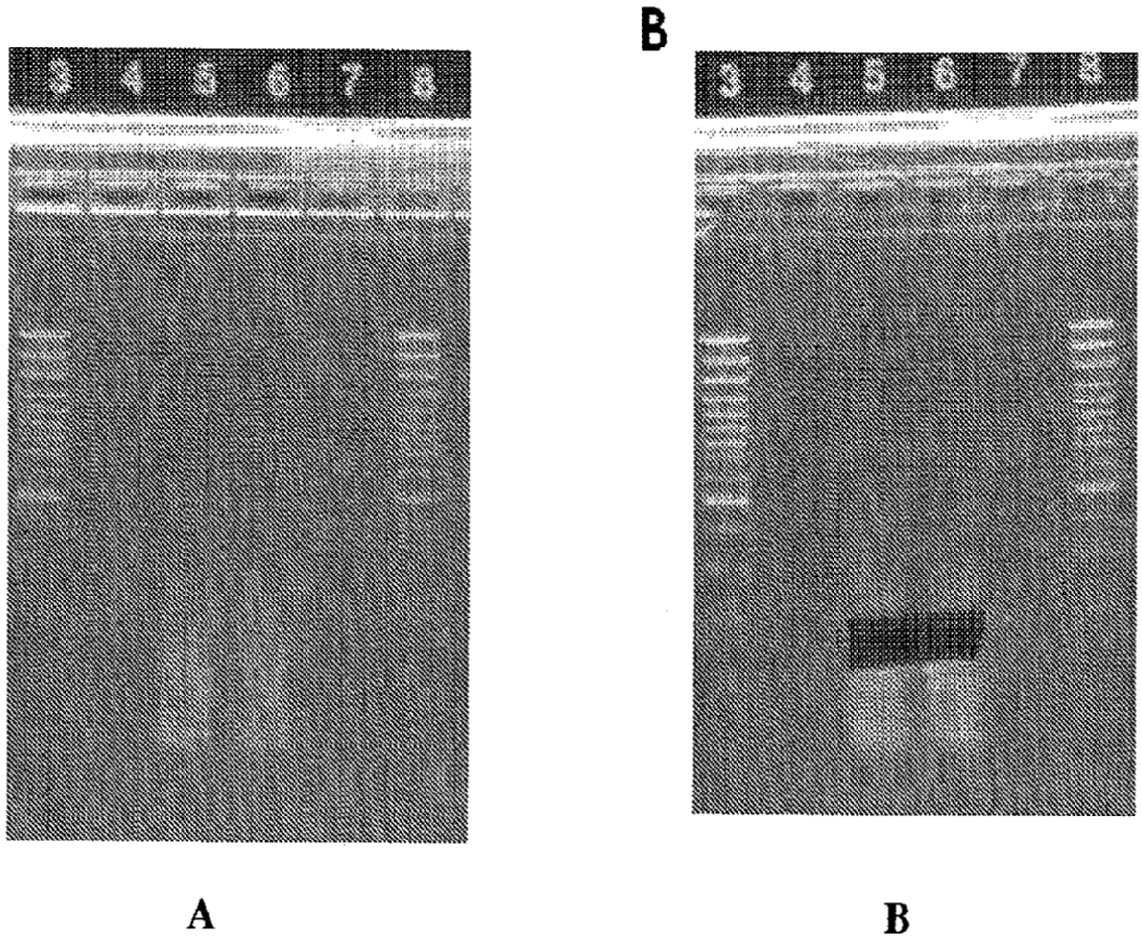

Figura 7

GTGGCTCTGATGCATGCTCGAGCGGCCGCCAGTGTGATGGATATCTGCAGAATTCGCCCTTCCACTAC
GCCTCCGCTTTTC
CTCTCTATGGGCAGTCGGTGATCTCAGAACGCGGATCCAAGACATGGGCGCTACTTGACTGTGGCTGC
CGTGTTCAGGGGC
CGCATGAAGAGAAGAATGAGGAACCCGGGGGCAGAAGGGCGAATTCCAGCACACTGGCGGCCGTTACTAGT
GGATCCGAGCTCG
GTACCAAGCTTGGCGTAATCATGGTCATAGCTGTTTCCTGTGTGAAATTGTTATCCGCTCACAATTCC
ACACAACATACG

Figura 8

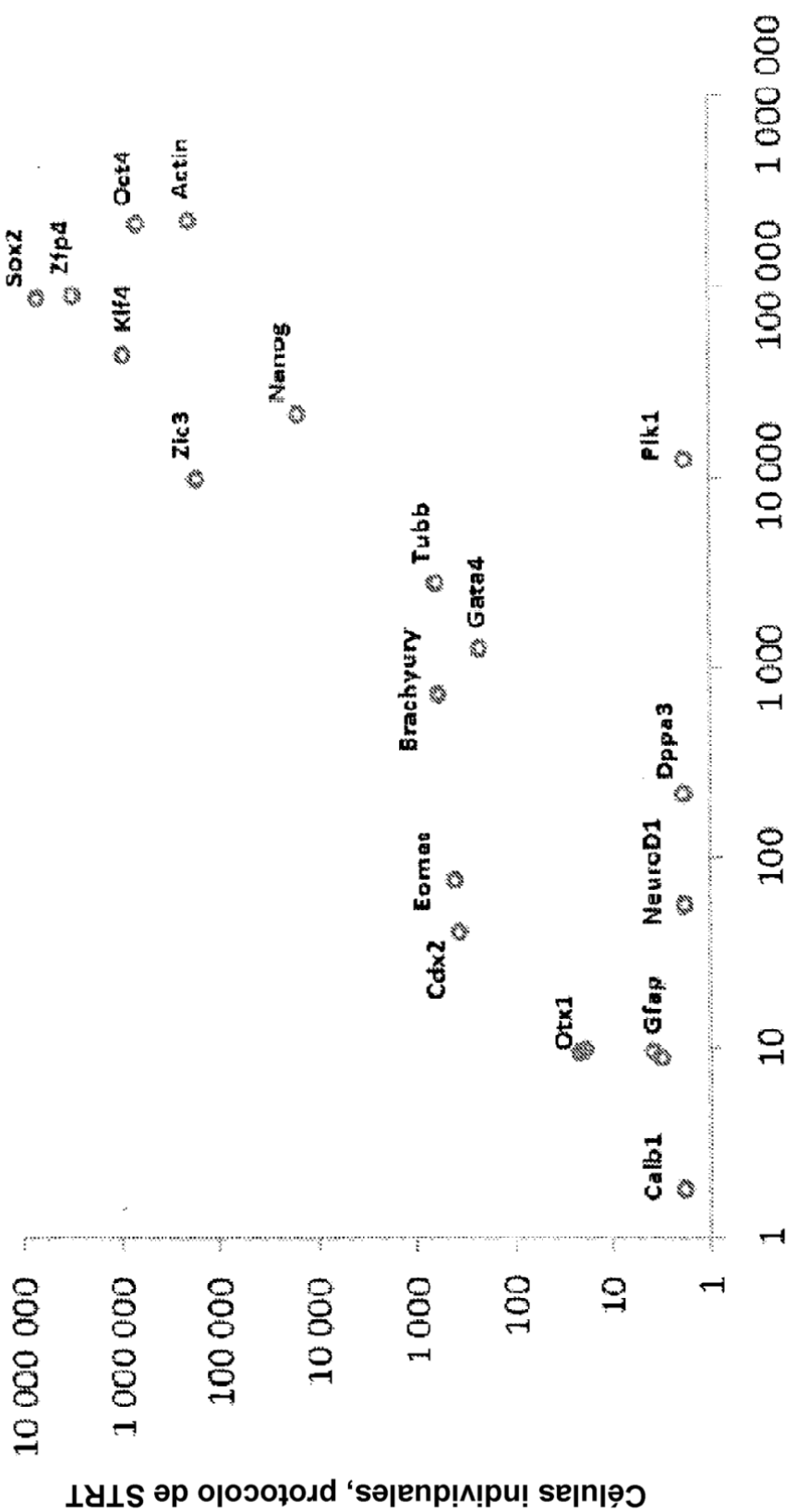
Células individuales, protocolo de STRT
10 000 000
1 000 000
100 000
10 000
1 000
100
10
1
Sox2
Zfp4
Klf4
Oct4
Actin
Nanog
Zic3
Pik1
Tubb
Gata4
Brachyury
Eomes
Cdx2
Otx1
Gfap
NeuroD1
Dppa3
Calb1
1
10
100
1 000
10 000
100 000
1 000 000
Células ES, 1 µg de ARN

Figura 9

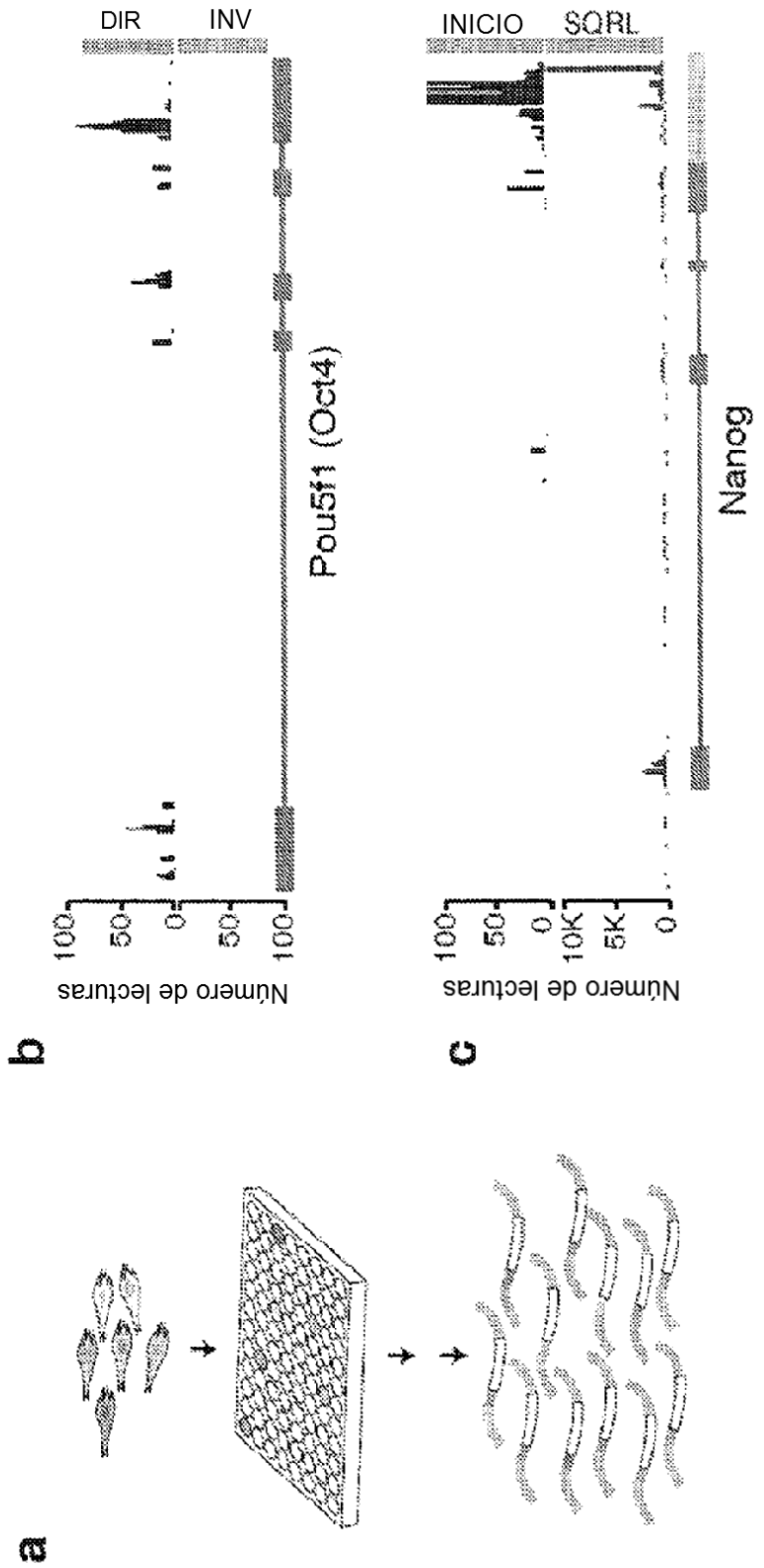
a
b
DIR
INV
Pou5f1 (Oct4)
Número de lecturas
100
50
0
50
100
c
INICIO
SQRL
Nanog
Número de lecturas
100
50
0
10K
5K
0

Figura 10A-C

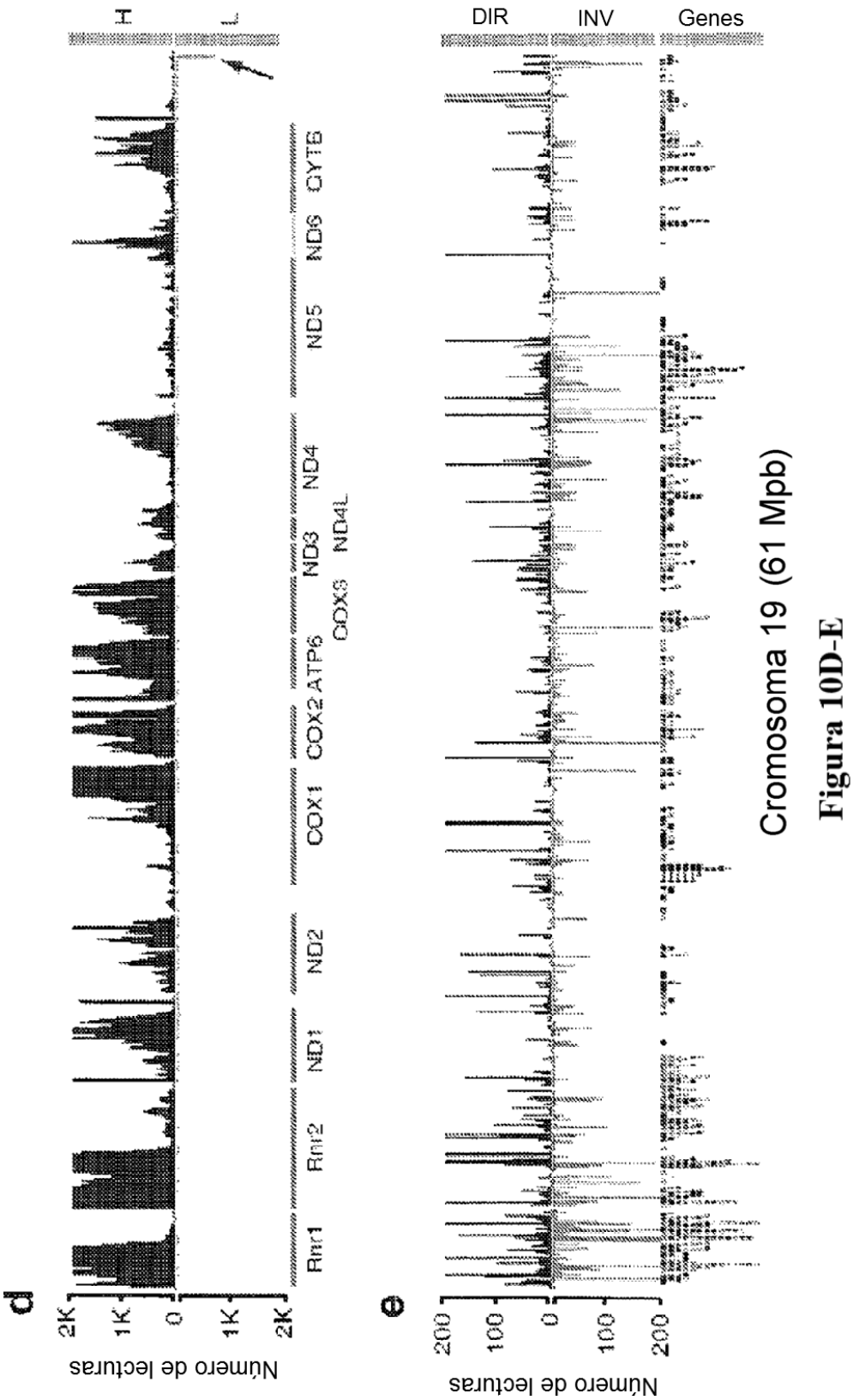
d
H
L
Rnr1
Rnr2
ND1
ND2
COX1
COX2
ATP6
COX3
ND3
ND4
ND4L
ND5
ND6
CYTB
2K
1K
0
1K
2K
Número de lecturas
e
DIR
INV
Genes
200
100
0
100
200
Número de lecturas
Cromosoma 19 (61 Mpb)

Figura 10D-E

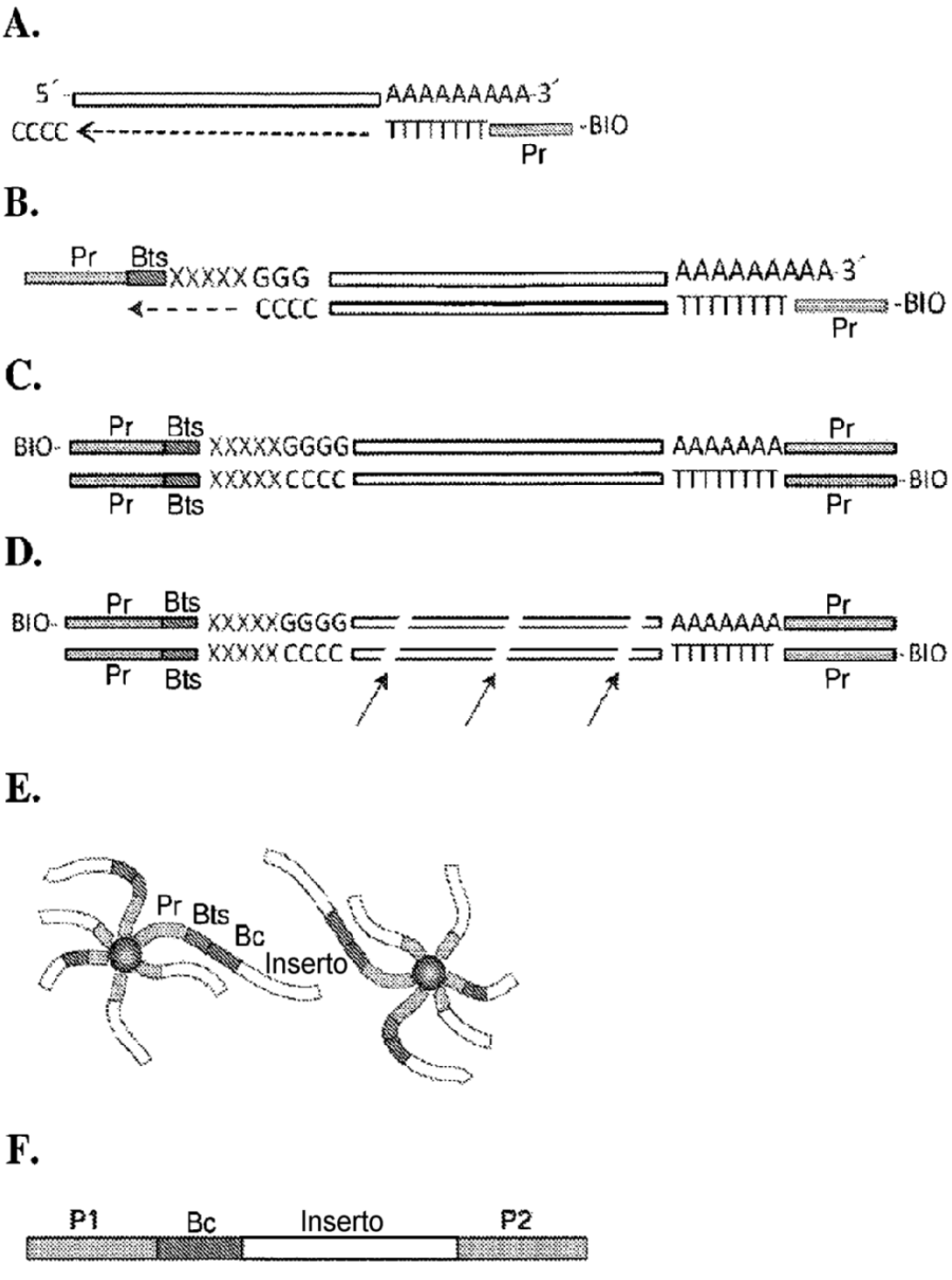
A.
5´
AAAAAAAAA-3´
CCCC
TTTTTTTT
-BIO
Pr
B.
Pr
Bts
XXXXX GGG
AAAAAAAAA-3´
CCCC
TTTTTTTT
-BIO
Pr
C.
BIO-
Pr
Bts
XXXXXGGGG
AAAAAAA
Pr
XXXXX CCCC
TTTTTTTT
-BIO
Pr
Bts
Pr
D.
BIO-
Pr
Bts
XXXXXGGGG
AAAAAAA
Pr
XXXXX CCCC
TTTTTTTT
-BIO
Pr
Bts
Pr
E.
Pr
Bts
Bc
Inserto
F.
P1
Bc
Inserto
P2

Figura 11

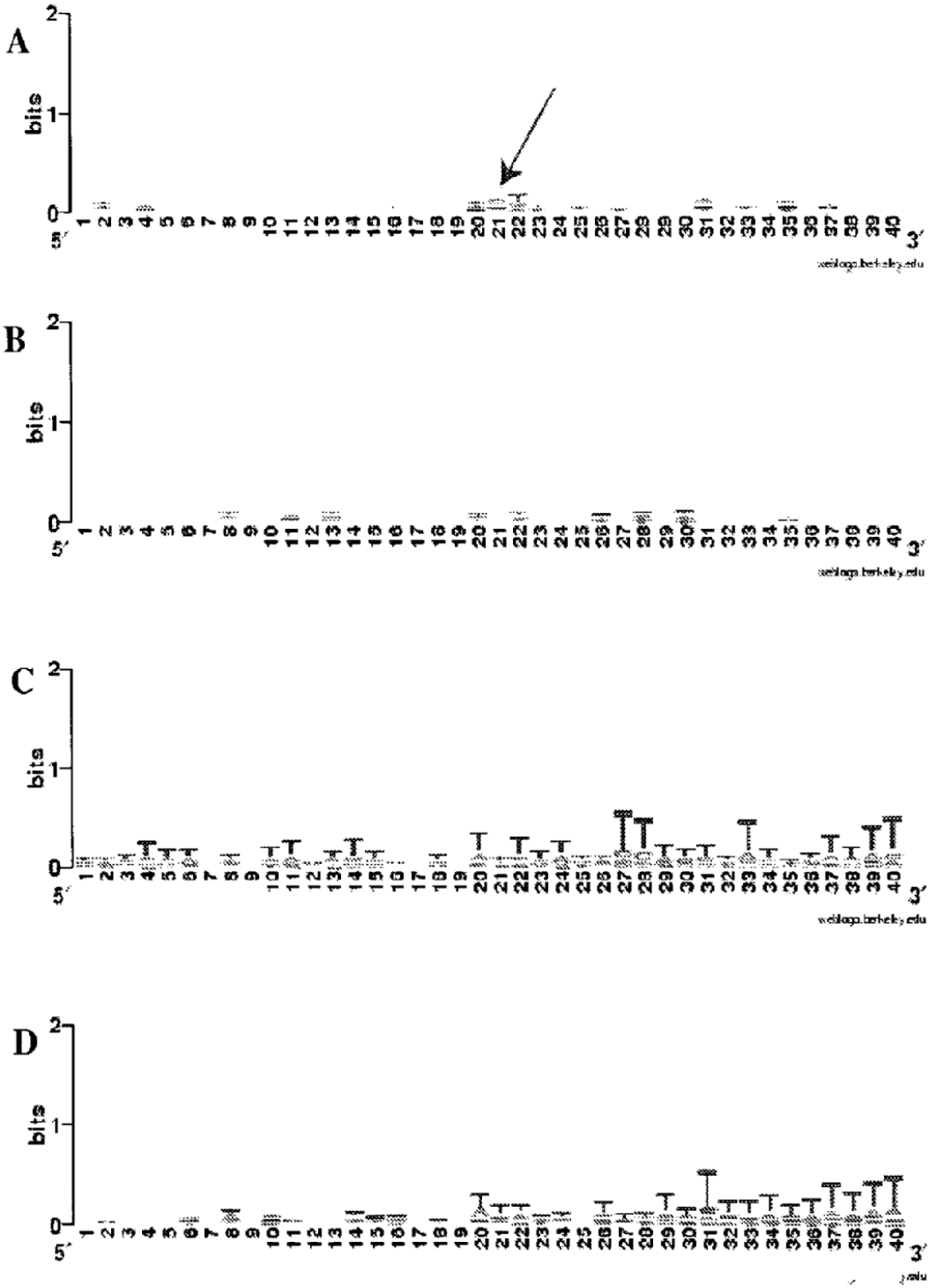

Figura 12

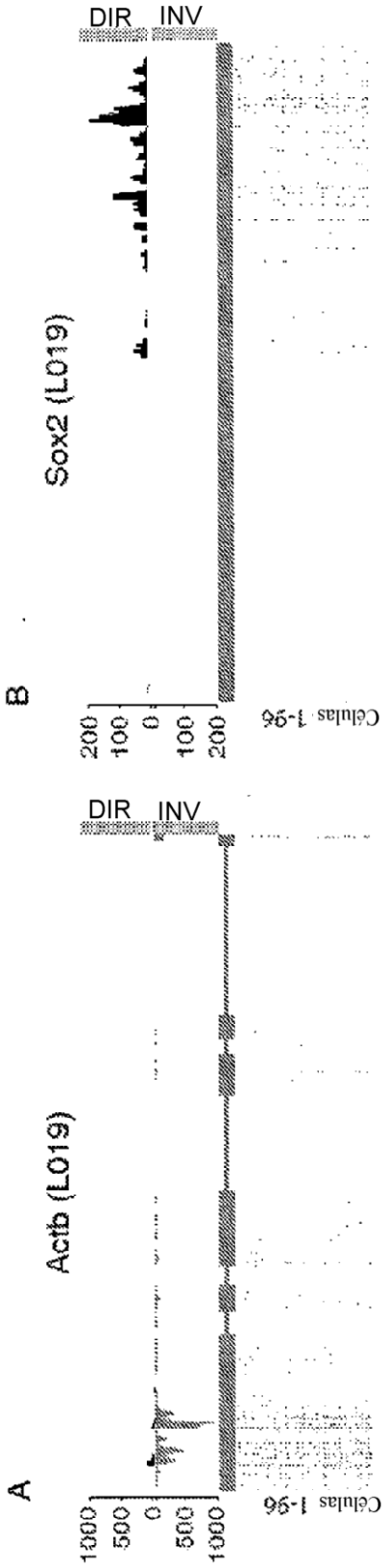
A
Actb (L019)
DIR
INV
1000
500
0
500
1000
Células 1-96
B
Sox2 (L019)
DIR
INV
200
100
0
100
200
Células 1-96

Figura 13

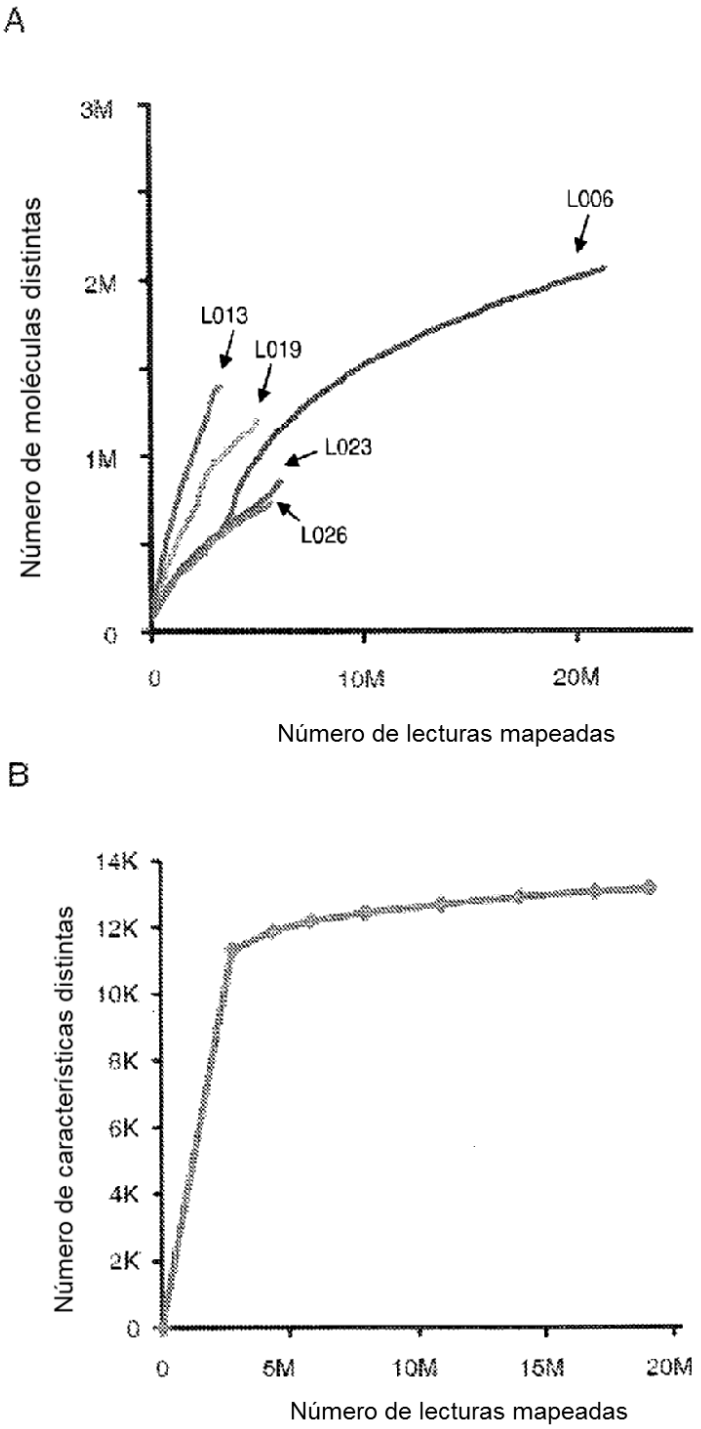
A
3M
2M
1M
0
Número de moléculas distintas
L006
L013
L019
L023
L026
0
10M
20M
Número de lecturas mapeadas
B
14K
12K
10K
8K
6K
4K
2K
0
Número de características distintas
0
5M
10M
15M
20M
Número de lecturas mapeadas

Figura 14

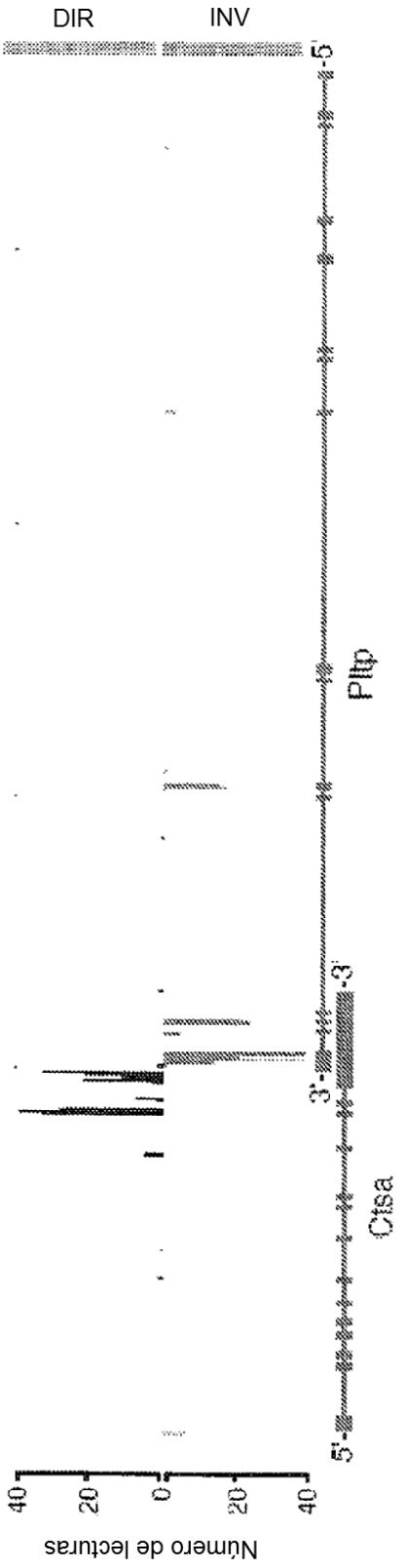

Figura 15

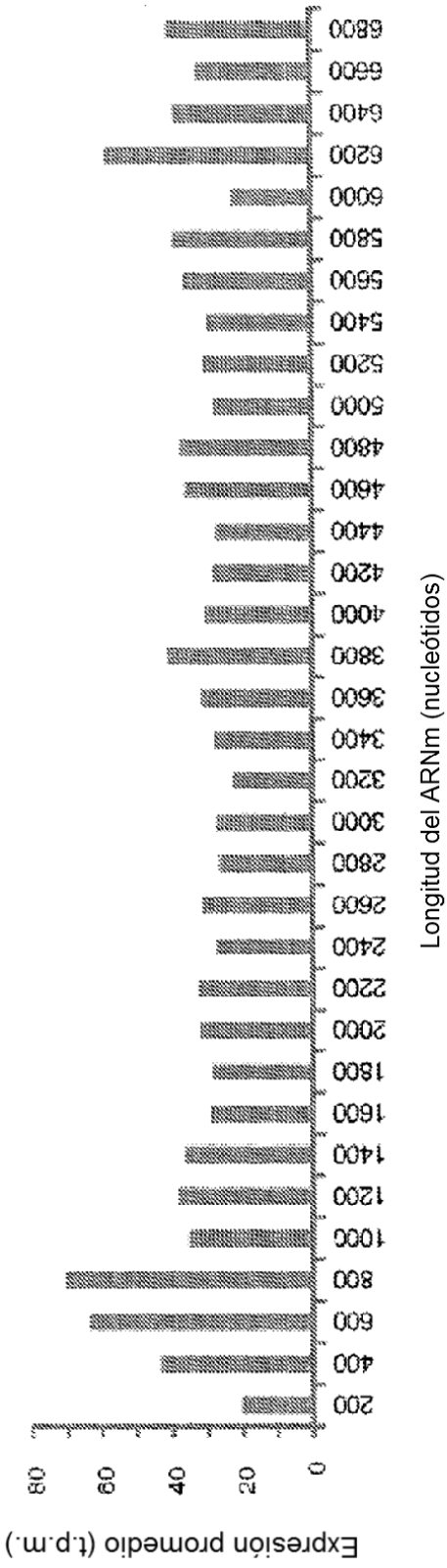
Expresión promedio (t.p.m.)
0
20
40
60
80
200
400
600
800
1000
1200
1400
1600
1800
2000
2200
2400
2600
2800
3000
3200
3400
3600
3800
4000
4200
4400
4600
4800
5000
5200
5400
5600
5800
6000
6200
6400
6600
6800
Longitud del ARNm (nucleótidos)

Figura 16

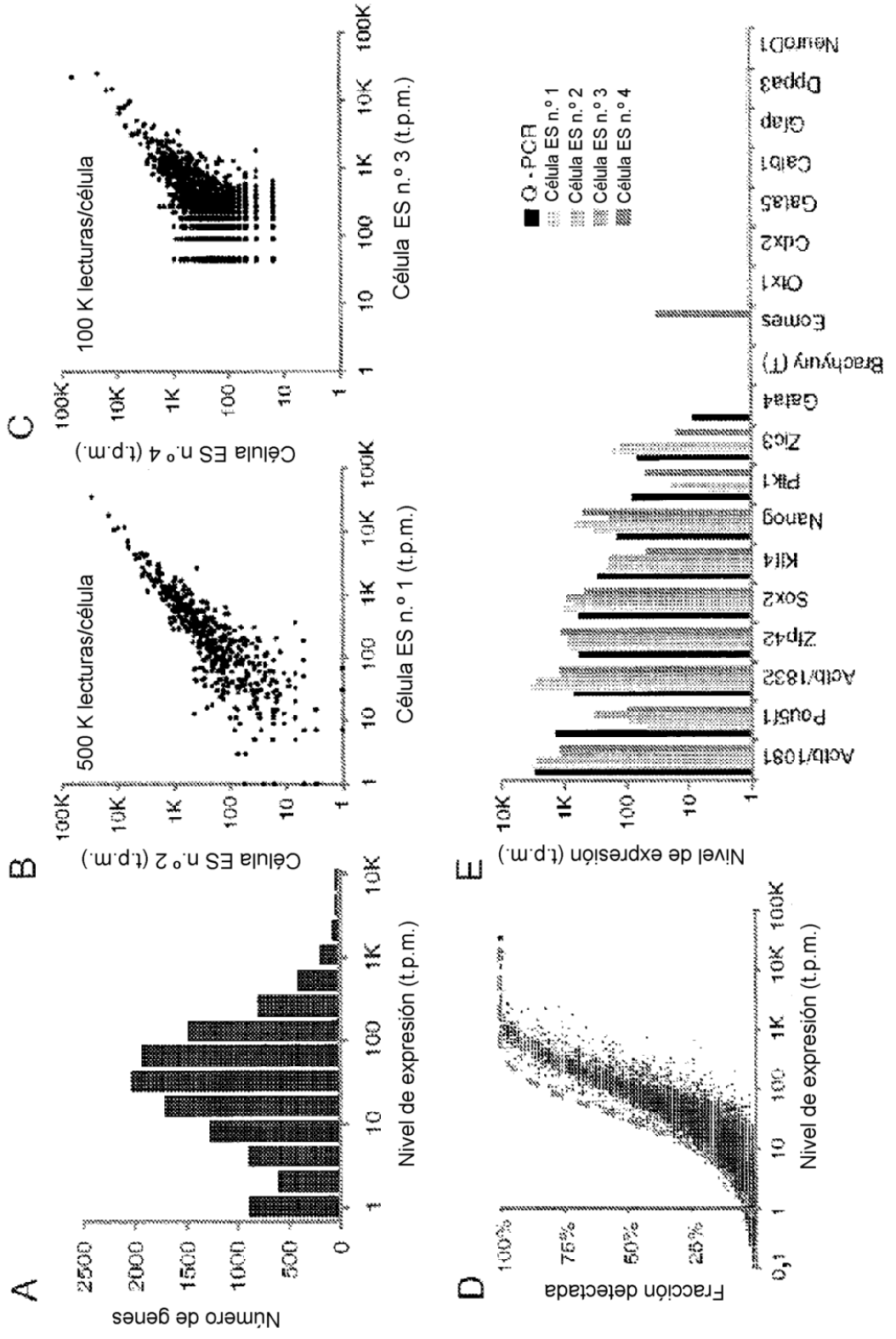
A
Número de genes
Nivel de expresión (t.p.m.)
B
Célula ES n.º 2 (t.p.m.)
500 K lecturas/célula
Célula ES n.º 1 (t.p.m.)
C
Célula ES n.º 4 (t.p.m.)
100 K lecturas/célula
Célula ES n.º 3 (t.p.m.)
D
Fracción detectada
Nivel de expresión (t.p.m.)
E
Nivel de expresión (t.p.m.)
Q-PCR
Célula ES n.º 1
Célula ES n.º 2
Célula ES n.º 3
Célula ES n.º 4
Actb/1081
Pou5f1
Actb/1832
Zfp42
Sox2
Klf4
Nanog
Plk1
Zic3
Gata4
Brachyury (T)
Eomes
Otx1
Cdx2
Gata5
Cdh1
Gfap
Dppa3
NeuroD1

Figura 17

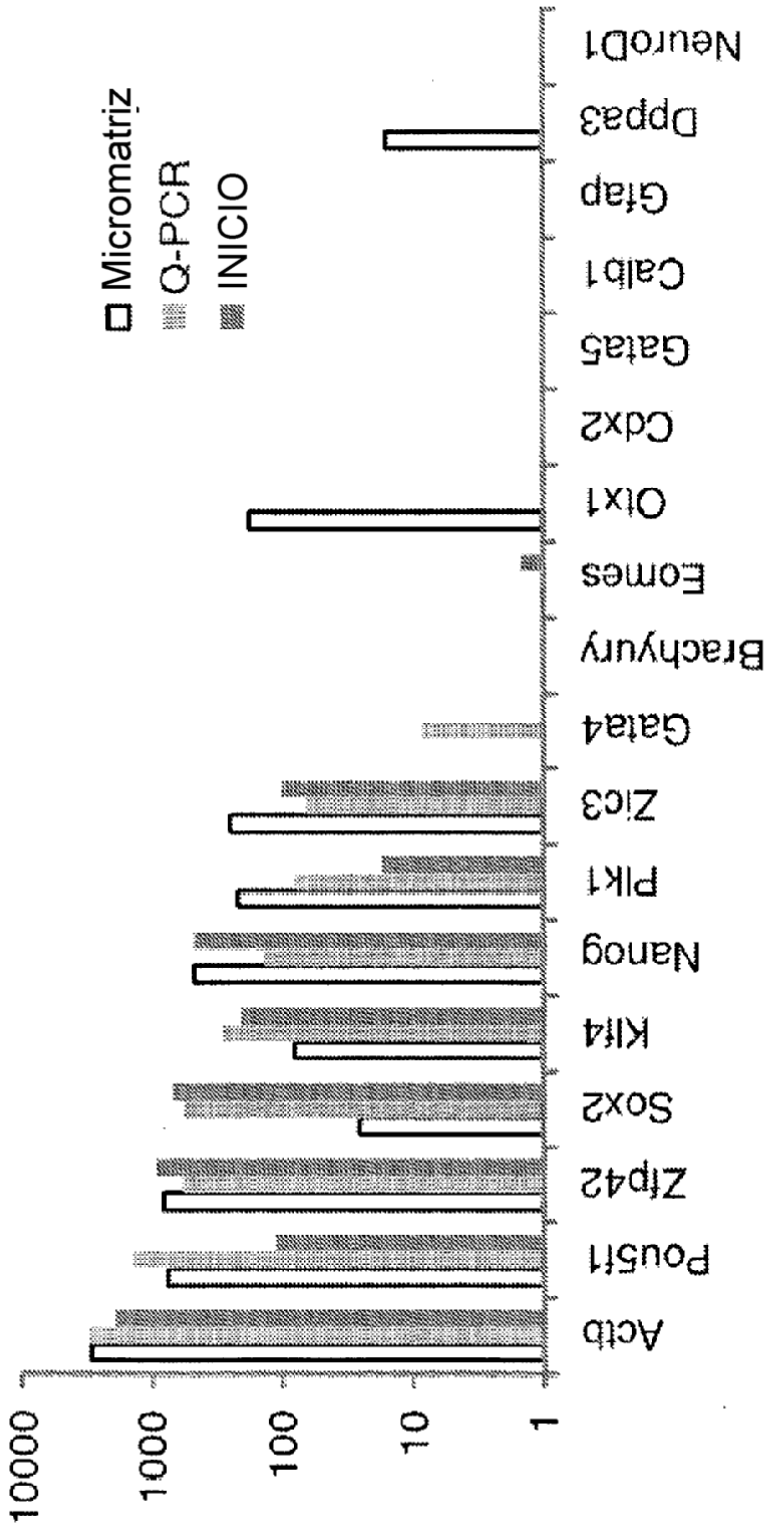
Micromatriz
Q-PCR
INICIO
10000
1000
100
10
1
Actb
Pou5f1
Zfp42
Sox2
Klf4
Nanog
Plk1
Zic3
Gata4
Brachyury
Eomes
Otx1
Cdx2
Gata5
Calb1
Gtap
Dppa3
NeuroD1

Figura 18

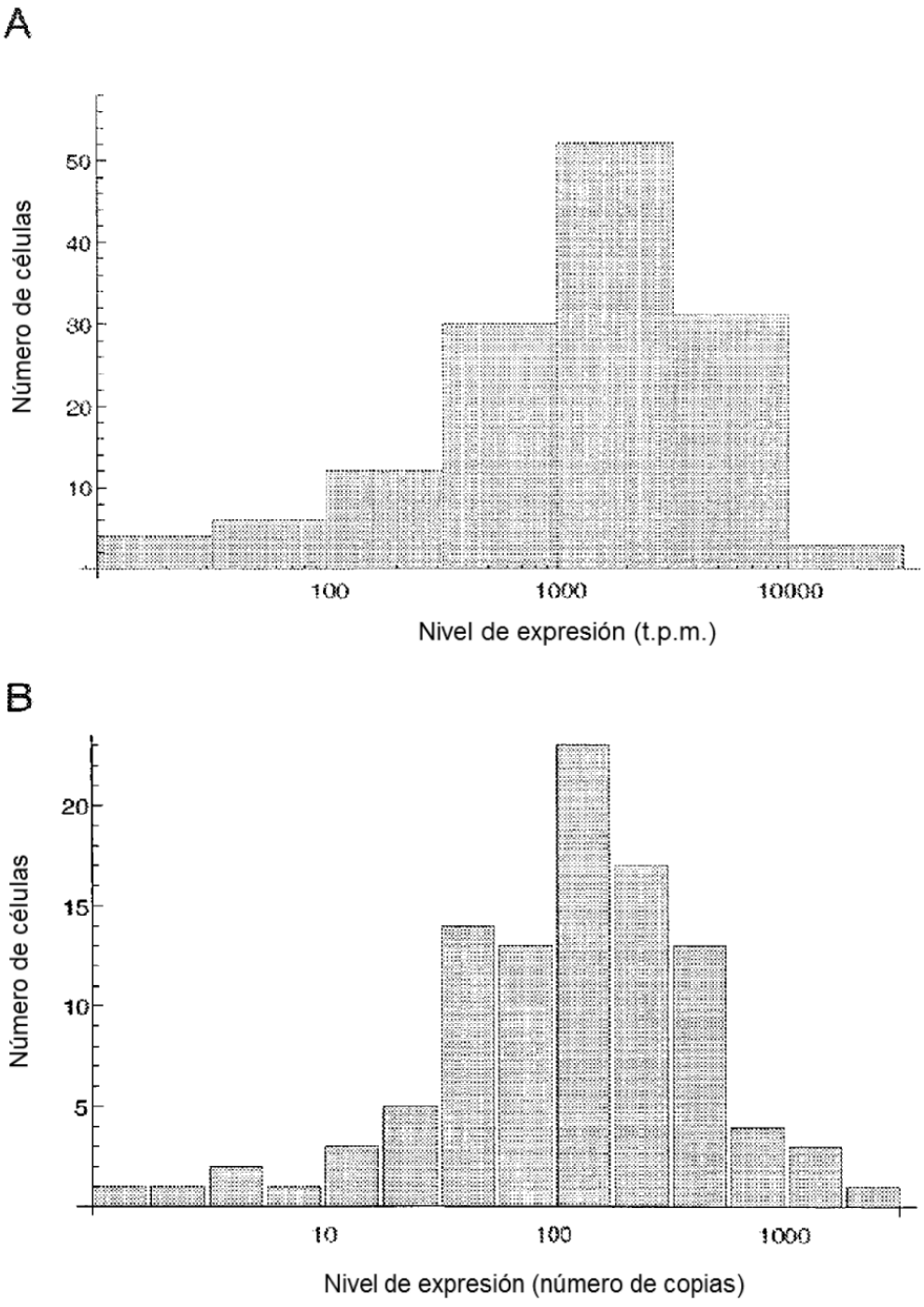
A
Número de células
50
40
30
20
10
100
1000
10000
Nivel de expresión (t.p.m.)
B
Número de células
20
15
10
5
10
100
1000
Nivel de expresión (número de copias)

Figura 19

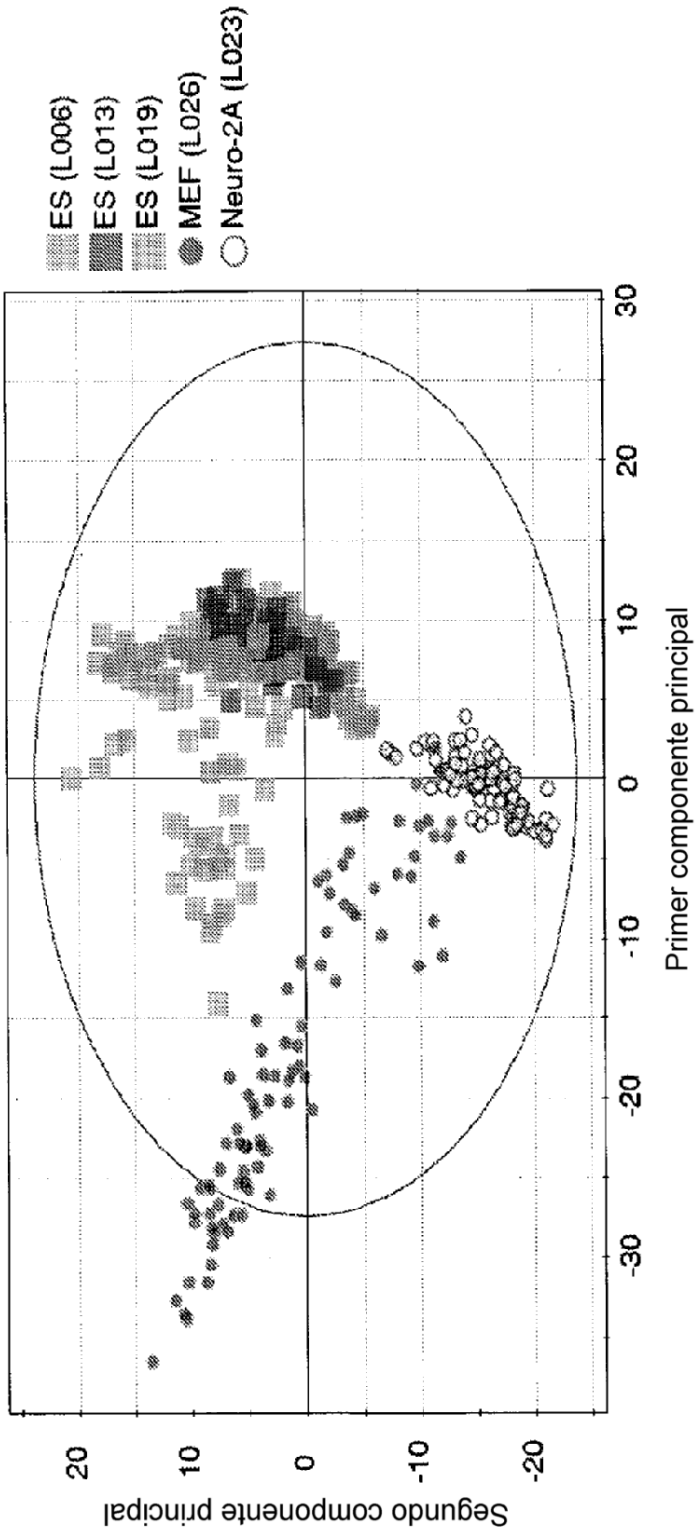
ES (L006)
ES (L013)
ES (L019)
MEF (L026)
Neuro-2A (L023)
Segundo componente principal
20
10
0
-10
-20
30
20
10
0
-10
-20
-30
Primer componente principal

Figura 20

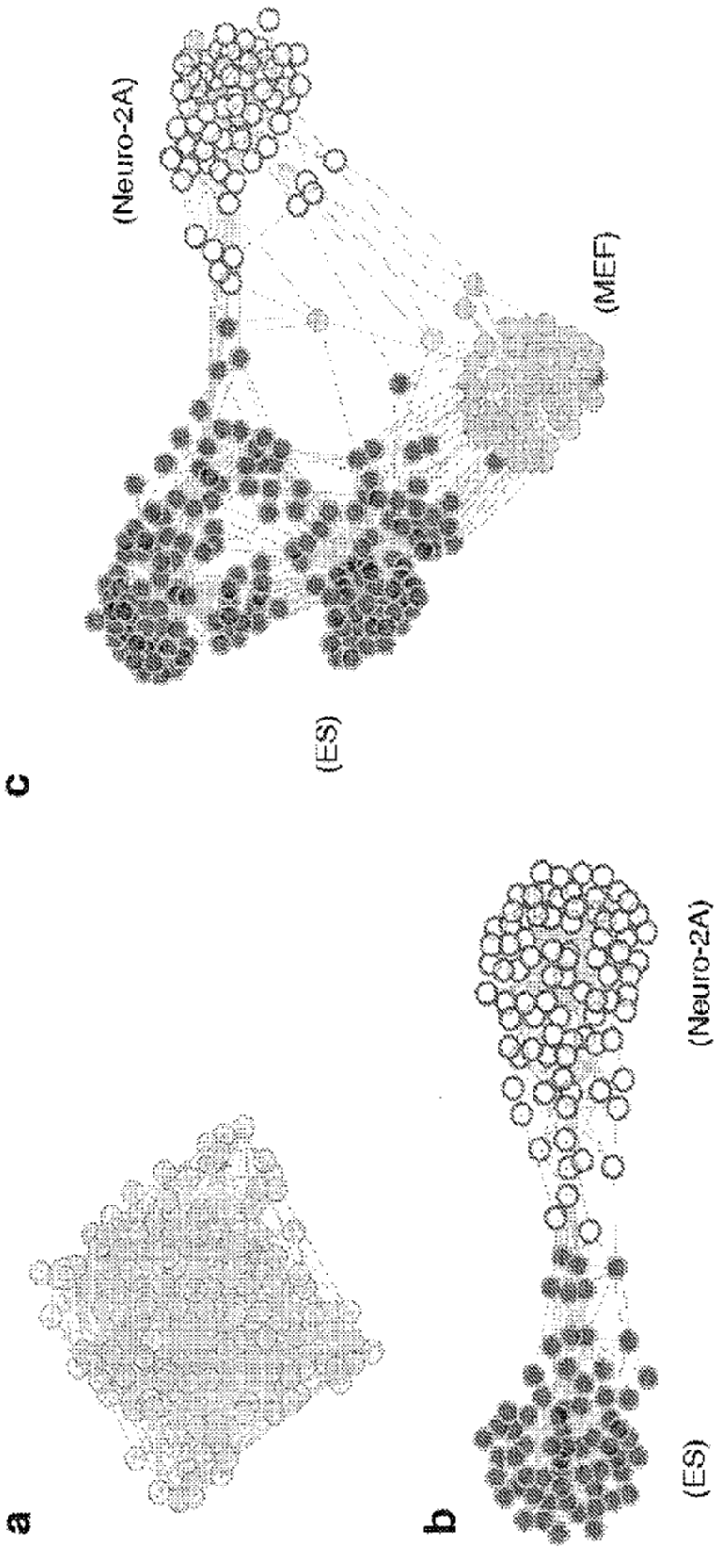

Figura 21

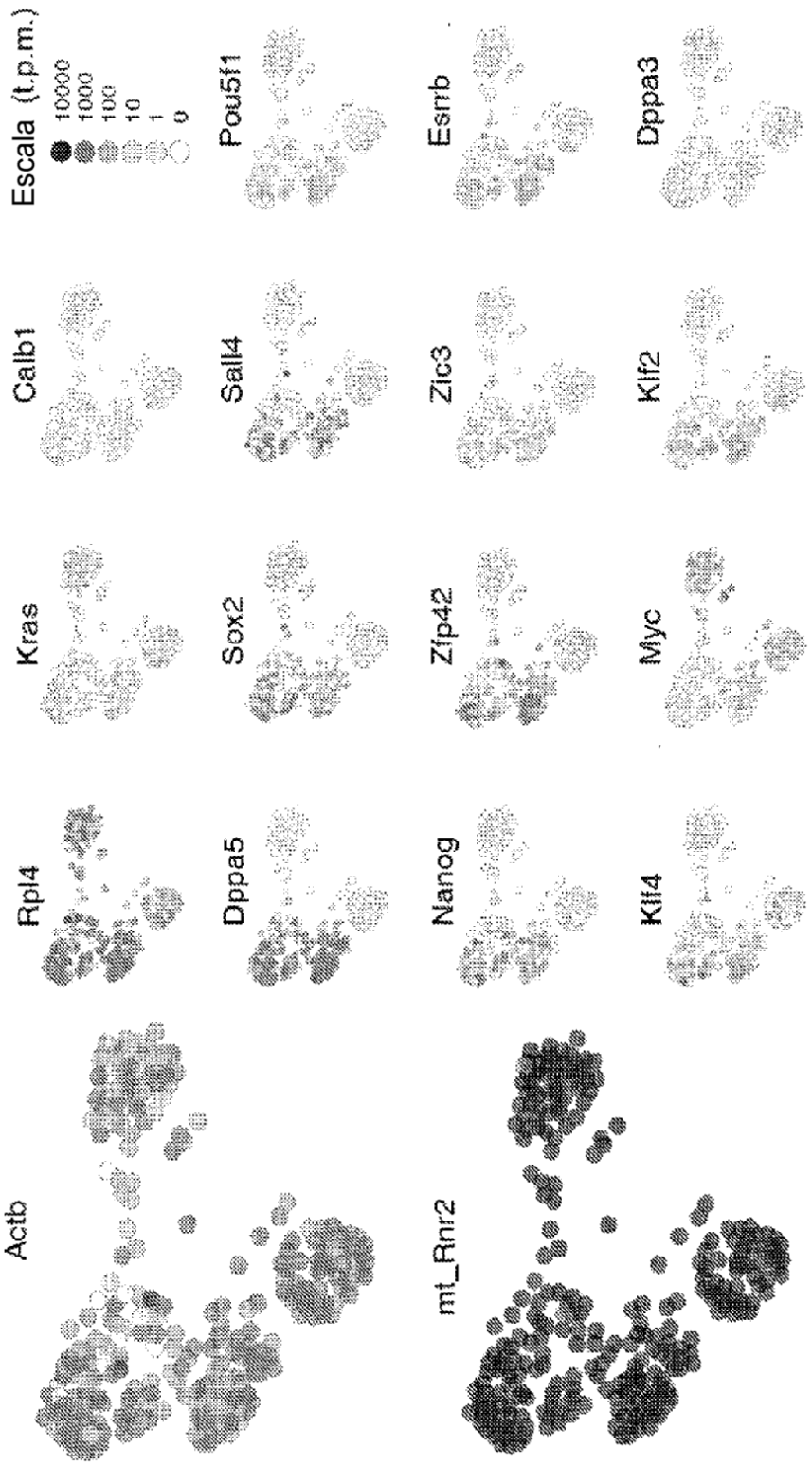
Escala (t.p.m.)
10000
1000
100
10
1
0
Actb
mt_Rnr2
Rpl4
Dppa5
Kras
Sox2
Nanog
Klf4
Zfp42
Myc
Calb1
Sall4
Zic3
Klf2
Pou5f1
Esrrb
Dppa3

Figura 22

REFERENCIAS CITADAS EN LA DESCRIPCIÓN

Esta lista de referencias citadas por el solicitante es únicamente para la comodidad del lector. No forma parte del documento de la patente europea. A pesar del cuidado tenido en la recopilación de las referencias, no se pueden excluir errores u omisiones y la EPO niega toda responsabilidad en este sentido.

Documentos de patentes citados en la descripción

- US 4683195 A **[0020]**
- US 4683202 A **[0020]**
- US 2007028313 A **[0025]**
- US 5962272 A **[0033]**
- US 5681702 A **[0038]**
- US 61164759 B **[0132]**
- US 2010028361 W **[0132]**

Literatura diferente de patentes citada en la descripción

- **CLOONAN et al.** *Nat Methods,* 2008, vol. 5 (7), 613-9 **[0004]**
- **BRAIL et al.** *Mutat Res,* 1999, vol. 406 (2-4), 45-54 **[0005]**
- **LEVSKY et al.** *Science,* 2002, vol. 297 (5582), 836-40 **[0005]**
- **BENGTSSON et al.** *Genome Res,* 2005, vol. 15 (10), 1388-92 **[0005] [0011]**
- **ESUMI et al.** *Nat Genet,* 2005, vol. 37 (2), 171-6 **[0005]**
- **ESUMI et al.** *Neurosci Res,* 2008, vol. 60 (4), 439-51 **[0005] [0025] [0027] [0033]**
- **OZSOLAK et al.** *Nature,* 2009, vol. 461, 814-818 **[0006]**
- **CLOONAN et al.** *Nat. Methods,* 2008, vol. 5, 613-619 **[0006] [0011] [0124]**
- **MORTAZAVI et al.** *Nat. Methods,* 2008, vol. 5, 621-628 **[0006] [0123]**
- **WANG et al.** *Nature,* 2008, vol. 456, 470-476 **[0006]**
- **ESUMI et al.** *Neurosci. Res.,* 2008, vol. 60, 439-451 **[0007]**
- **KURIMOTO et al.** *Nucleic Acids Res.,* 2006, vol. 34, 42 **[0007]**
- **BENGTSSON et al.** *Genome Res,* 2005, vol. 15, 1388-1392 **[0007] [0124]**
- **RAJ et al.** *PLoS Biol,* 2006, vol. 4, 309 **[0007] [0008] [0124]**
- **LAGUNAVICIUS et al.** *RNA,* 2009, vol. 15, 765-771 **[0007]**
- **WARREN et al.** *Proc. Natl. Acad. Sci. U.S.A.,* 2006, vol. 103, 17807-17812 **[0007]**
- **TANIGUCHI et al.** *Nat Methods,* 2009, vol. 6, 503-506 **[0007]**
- **CHUBB et al.** *Curr Biol,* 2006, vol. 16, 1018-1025 **[0008]**
- **INNIS et al.** PCR protocols: a guide to method and applications. Academic Press, 1990 **[0020]**
- **WU et al.** *Genomics,* 1989, vol. 4, 560-569 **[0020]**
- **KURIMOTO et al.** *Nucleic Acids Res,* 2006, vol. 34 (5), e42 **[0025] [0027]**
- **KURIMOTO et al.** *Nat Protoc,* 2007, vol. 2 (3), 739-52 **[0027]**
- **DAI et al.** *J Biotechnol,* 2007, vol. 128 (3), 435-43 **[0027] [0044]**
- **MACPHERSON et al.** PCR 1 : A Practical Approach. IRL Press at Oxford University Press, 1991 **[0045]**
- **AUSUBEL et al.** Current Protocols in Molecular Biology. John Wiley & Sons, 1993 **[0048]**
- **MOLINER et al.** *Stem Cells Dev,* 2008, vol. 17, 233-243 **[0102]**
- **LANGMEAD et al.** *Genome Biol.,* 2009, vol. 10, R25 **[0113]**
- **WOOD et al.** *Nature,* 1993, vol. 365, 87-89 **[0115]**
- **OLMSTED et al.** *Proc. Natl. Acad. Sci U.S.A.,* 1970, vol. 65, 129-136 **[0115]**
- **SCHMIDT et al.** *Nucleic Acids Res.,* 1999, vol. 27, e31 **[0115] [0123]**
- **YELIN et al.** *Nat. Biotechnol,* 2003, vol. 21, 379-386 **[0121]**
- **SLOMOVIC et al.** *Mol. Cell. Biol,* 2005, vol. 25, 6427-6435 **[0121]**
- **OSHLACK et al.** *Biol. Direct,* 2009, vol. 4, 14 **[0123]**