

19



OFICINA ESPAÑOLA DE
PATENTES Y MARCAS

ESPAÑA



11 Número de publicación: **2 998 552**

51 Int. Cl.:

G16B 20/40 (2009.01)

G16B 30/00 (2009.01)

G16H 10/40 (2008.01)

G16H 40/40 (2008.01)

12

TRADUCCIÓN DE PATENTE EUROPEA

T3

86 Fecha de presentación y número de la solicitud internacional: **04.12.2020** **PCT/US2020/063462**

87 Fecha y número de publicación internacional: **10.06.2021** **WO21113749**

96 Fecha de presentación y número de la solicitud europea: **04.12.2020** **E 20834038 (0)**

97 Fecha y número de publicación de la concesión europea: **09.10.2024** **EP 4070318**

54 Título: **Sistemas y procedimientos de automatización de llamadas de expresión de RNA en una canalización para la predicción del cáncer**

30 Prioridad:

04.12.2019 US 201962943712 P

45 Fecha de publicación y mención en BOPI de la traducción de la patente:
20.02.2025

73 Titular/es:

TEMPUS AI, INC. (100.00%)
600 West Chicago Avenue, Suite 510
Chicago, IL 60654, US

72 Inventor/es:

BELL, JOSHUA, SK;
IGARTUA, CATHERINE y
DREWS, JOSHUA

74 Agente/Representante:

GONZÁLEZ PECES, Gustavo Adolfo

ES 2 998 552 T3

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

DESCRIPCIÓN

Sistemas y procedimientos de automatización de llamadas de expresión de RNA en una canalización para la predicción del cáncer

Campo técnico

- 5 La presente divulgación se refiere en general al uso de información de secuencia de ARN para realizar el control de calidad en lotes de muestras de secuenciación de ARN.

Antecedentes

- 10 Las métricas de calidad de ARN de una sola muestra implantadas actualmente en los procesos de control de calidad en bioinformática se diseñaron para detectar la calidad general de la muestra y la integridad del transcriptoma. Sin embargo, hay muchas fuentes potenciales de error que pueden afectar a la calidad de las muestras y, en concreto, hay efectos que pueden repercutir en los resultados de lotes enteros de muestras de secuenciación. Por ejemplo, los cambios en el protocolo de preparación de bancos (por ejemplo, reactivos, lote de sondas de captura o instrumental) o los cambios en la canalización de datos bioinformáticos (por ejemplo, la versión de un programa) pueden dar lugar a cambios sutiles en el transcriptoma, denominados efectos de lote, e influir en el rendimiento del modelo y en la interpretación clínica de los procesos posteriores (por ejemplo, determinar si un gen está sobreexpresado o subexpresado en relación con muestras secuenciadas previamente, o modelos de aprendizaje automático para diagnosticar tumores de origen desconocido). Estos efectos de lote sólo son detectables aprovechando los patrones de datos entre muestras y requieren la aplicación de nuevos procesos entre el cartografiado del ARN y el análisis posterior.
- 15 El documento BÜTTNER MAREN *et al.*, "A test metric for assessing single-cell RNA-Seq batch correction", NATURE METHODS, NATURE PUB. GROUP, NUEVA YORK, vol. 16, n.º 1, 20 de diciembre de 2018 (20-12-2018), páginas 43-49 divulga una prueba para el efecto de lote de los k vecinos más próximos fácil de usar, robusta y sensible para la cuantificación de efectos de lote.
- 20 El documento NYAMUNDANDA GIFT *et al.*, "A Novel Statistical Method to Diagnose, Quantify and Correct Batch Effects in Genomic Studies", SCIENTIFIC REPORTS, vol. 7, n.º 1, 7 de septiembre de 2017 (07-09-2017) divulga un novedoso procedimiento estadístico, que determina el efecto de lote para evaluar el efecto de lote basado en el análisis probabilístico de componentes principales y covariables ("probabilistic principal component and covariates analysis", PCCA).
- 25 El documento POLANSKI KRZYSZTOF *et al.*, "BBKNN: fast batch alignment of single cell transcriptomes", BIOINFORMATICS, 10 de agosto de 2019 (10-08-2019) divulga un algoritmo de integración de datos basado en gráficos extremadamente rápido para combinar diversos conjuntos de datos mediante la eliminación de los efectos de lote técnicos.
- 30

En la técnica son necesarios procedimientos mejorados para realizar automáticamente la detección en un control de calidad y el análisis de los efectos de lote de muestras de secuenciación de ARN heterogéneas con alto rendimiento.

35 **Sumario de la invención**

La invención se expone en las reivindicaciones adjuntas.

Sumario de la divulgación

- 40 Dados los antecedentes anteriores, se necesitan sistemas y procedimientos mejorados para realizar el control de calidad de lotes (por ejemplo, la evaluación de la calidad) de muestras de ARN. Ventajosamente, la presente divulgación proporciona soluciones a estas y otras deficiencias de la técnica. Por ejemplo, en algunas realizaciones, los sistemas y procedimientos descritos en el presente documento proporcionan un control de calidad automatizado de lotes enteros de muestras de secuenciación de ARN (por ejemplo, realizando así un análisis de control de calidad más rápido que el disponible actualmente). Del mismo modo, en algunas realizaciones, los procedimientos y sistemas descritos en el presente documento mejoran los sistemas y procedimientos de diagnóstico que utilizan datos de expresión de ARN, por ejemplo, para la oncología de precisión, mediante la identificación de los efectos de lote que de otro modo no se identificarían a través de las métricas normales de control de calidad de una sola muestra.
- 45

- Un aspecto de la presente divulgación proporciona un procedimiento para realizar el control de calidad. El procedimiento se realiza en un sistema informático que tiene uno o más procesadores, y una memoria que almacena uno o más programas para su ejecución por dichos uno o más procesadores. El procedimiento consiste en obtener (a), en formato electrónico, un conjunto de datos de lotes que comprende, para cada muestra respectiva de un lote de muestras, una pluralidad correspondiente de lecturas de secuencia derivadas de la muestra respectiva mediante secuenciación de ARN dirigida o del transcriptoma completo y los metadatos correspondientes para la muestra respectiva. El procedimiento continúa determinando (b), para el conjunto de datos de lotes, un lote de referencia emparejado por cohorte, en el que el lote de referencia emparejado por cohorte está equilibrado en cuanto al
- 50

- emplazamiento del tejido, la pureza del tumor, el tipo de cáncer, el procedimiento de recolección, la identidad del secuenciador y/o la fecha de secuenciación. El procedimiento realiza una o más pruebas de control de calidad global de lotes en el conjunto de datos de lotes utilizando al menos el lote de referencia emparejado por cohorte. El procedimiento elimina (d) las muestras respectivas del conjunto de datos de lotes que no superan una o más pruebas de control de calidad global de lotes, o marca para la inspección manual las muestras respectivas que no superan una o más pruebas de control de calidad global de lotes.
- En algunas realizaciones, la determinación de un conjunto de datos de referencia emparejado por cohorte para el conjunto de datos de lotes comprende extraer, para cada muestra en el lote de muestras: i) una pluralidad respectiva de características de secuencia a partir de la pluralidad respectiva de lecturas de secuencia, obteniendo de este modo una pluralidad de características de secuencia del lote, y ii) una pluralidad respectiva de características de metadatos de la muestra, obteniendo de este modo una pluralidad de características de metadatos del lote. En algunas realizaciones, la determinación del conjunto de datos de referencia emparejado por cohorte comprende además la selección, a partir de un conjunto de datos de referencia, basada al menos en parte en la pluralidad de características de procesamiento y secuencia de muestras del lote o en la pluralidad de características de metadatos del lote, del conjunto de datos de referencia emparejado por cohorte que comprende una pluralidad de muestras de referencia.
- En algunas realizaciones, el procedimiento comprende además realizar, para cada muestra respectiva en el lote de muestras, a partir de la pluralidad correspondiente de lecturas de secuencia, una o más pruebas de control de calidad de una sola muestra en la muestra respectiva; y eliminar las muestras respectivas del lote de muestras que no superan cualquiera de dichas una o más pruebas de control de calidad de una sola muestra, o marcar para inspección manual las muestras respectivas que no superan cualquiera de dichas una o más pruebas de control de calidad de una sola muestra.
- En algunas realizaciones, dichas una o más pruebas de control de calidad global de lotes comprenden pruebas para uno o más efectos de lote de un conjunto que comprende procedimientos de secuenciación y análisis de la canalización de datos bioinformáticos.
- En algunas realizaciones, el procedimiento comprende además determinar una combinación lineal o no lineal de la pluralidad de características de secuencia del lote y la pluralidad de características de metadatos del lote sometiendo la pluralidad de características de secuencia del lote y la pluralidad de características de metadatos del lote a una técnica de reducción de dimensiones.
- En algunas realizaciones, el procedimiento comprende además usar el lote de referencia emparejado por cohorte para ajustar cada muestra en el conjunto de datos de lotes a una o más covariables de confusión antes de realizar (c) dichas una o más pruebas de control de calidad global de lotes.
- En algunas realizaciones, el procedimiento comprende además proporcionar, para cada muestra en el lote de muestras, un informe de muestra respectivo, en el que cada informe de muestra respectivo comprende al menos uno de un conjunto de llamadas de expresión, una o más terapias emparejadas, o uno o más ensayos clínicos emparejados.
- En algunas realizaciones, el procedimiento se lleva a cabo en un sistema informático que comprende un servidor en la nube. En algunas realizaciones, dichas una o más pruebas de control de calidad global de lotes comprenden un primer módulo, y dichas una o más pruebas de control de calidad de una sola muestra comprenden un segundo módulo.
- Otras realizaciones se dirigen a sistemas, dispositivos para consumidores portátiles y medios legibles por ordenador asociados con los procedimientos descritos en el presente documento. Cualquier realización descrita en el presente documento, cuando sea aplicable, puede aplicarse a cualquier aspecto de los procedimientos descritos en el presente documento.
- Los expertos en la materia apreciarán sin inconvenientes otros aspectos y ventajas de la presente divulgación a partir de la siguiente descripción detallada, en la que sólo se muestran y describen realizaciones ilustrativas de la presente divulgación. Como será evidente, la presente divulgación incluye otras realizaciones diferentes, y sus distintos detalles pueden modificarse en diversos aspectos obvios, todo ello sin apartarse de la divulgación. Por consiguiente, los dibujos y la descripción deben considerarse de carácter ilustrativo y no restrictivo.
- Breve descripción de los dibujos**
- La figura 1 ilustra un diagrama de bloques de un dispositivo informático de ejemplo, de acuerdo con algunas realizaciones de la presente divulgación.
- Las figuras 2A y 2B proporcionan colectivamente un diagrama de flujo de procesos y características para realizar el control de calidad de un lote de muestras de secuencias de ARN, en el que los bloques opcionales se indican con recuadros discontinuos, de acuerdo con algunas realizaciones de la presente divulgación.

- 5 Las figuras 3A, 3B y 3C muestran colectivamente la evaluación de los efectos de lote técnicos en muestras de ARN recogidas con tubos PAX o EDTA, de acuerdo con algunas realizaciones de la presente divulgación. La figura 3A ilustra la incrustación UMAP de cohortes agrupadas y muestras emparejadas por tejido. Las figuras 3B y 3C ilustran que la prueba de la U de Mann Whitney encuentra programáticamente la diferencia entre las muestras emparejadas recogidas con tubos PAX o EDTA en ambas coordenadas UMAP. Las coordenadas UMAP no están ordenadas, y la coordenada que detecta el efecto de lote es arbitraria.
- 10 Las figuras 4A, 4B y 4C muestran colectivamente la evaluación de los efectos de lote técnicos en las muestras alineadas con las canalizaciones de los programas bioinformáticos Kallisto o STAR, de acuerdo con algunas realizaciones de la presente divulgación. La figura 4 ilustra la incrustación UMAP de muestras emparejadas por cohorte y por tejido. Las figuras 4B y 4C ilustran los resultados de una prueba de la U de Mann Whitney en cada una de las coordenadas UMAP.
- 15 Las figuras 5A, 5B y 5C muestran colectivamente la distribución de las tasas de falsos descubrimientos ("false discovery rates", FDR) para fuentes comunes de efectos de lote técnicos. Para cada clase técnica (celda de flujo (figura 5A), canalización (figura 5B) y secuenciador (figura 5C)), se analizaron los efectos de lote técnicos para 15 submuestras por característica y se calcularon las FDR corregidas por Benjamini Hochberg para todas las submuestras. Las distribuciones presentadas representan la mediana de la FDR en las submuestras.
- 20 Las figuras 6A y 6B ilustran colectivamente los resultados del análisis de reducción de dimensiones PCA de aproximadamente 100 muestras de cáncer emparejadas para las que se determinó la expresión de ARN antes y después de un cambio de producción, como se describe en el ejemplo 2. Los efectos de lote técnicos derivados del cambio de producción se identifican en el tercer término de componente principal (PC3) de la figura 6A. Este efecto de lote técnico puede eliminarse aplicando un factor de corrección a los datos de expresión de ARN obtenidos tras el cambio de producción, como se ilustra en la figura 6B.
- 25 La figura 7 ilustra la incrustación UMAP del análisis del transcriptoma de cánceres hematológicos a partir de muestras biológicas recogidas mediante tres metodologías.
- La figura 8 ilustra la incrustación UMAP del análisis del transcriptoma de cánceres a partir de muestras biológicas en las que la extracción del ARN fue realizada por el médico o inmediatamente antes de la secuenciación del ARN.
- 30 La figura 9 ilustra la incrustación PCA (PC8) del análisis del transcriptoma de las mismas muestras de cáncer utilizando dos lotes de las mismas sondas de captura. Las muestras emparejadas se ilustran con una línea que une los respectivos términos de PC.
- La figura 10 ilustra la incrustación UMAP del análisis del transcriptoma de cánceres realizados en grupos de una sola muestra, 3x o 6x. El grupo de muestras emparejadas se indica con círculos.
- 35 La figura 11 ilustra la incrustación UMAP del análisis del transcriptoma de cánceres realizado tras 7-9 ciclos de amplificación por PCR tras el enriquecimiento. El grupo de muestras emparejadas se indica con círculos.
- La figura 12 ilustra la incrustación UMAP del análisis del transcriptoma de cánceres realizado con diferentes molaridades de carga del secuenciador (0,7 uM, 1 uM y 1,5 uM). El grupo de muestras emparejadas se indica con círculos.
- 40 La figura 13 ilustra la incrustación UMAP del análisis del transcriptoma de cánceres realizado con diferentes reactivos químicos de secuenciación. El grupo de muestras emparejadas se indica con las líneas entre los puntos.
- La figura 14 ilustra un ejemplo de canalización de datos para la obtención de perfiles de expresión de ARN, de acuerdo con algunas realizaciones de la presente divulgación.
- 45 La figura 15 ilustra un ejemplo de procedimiento para realizar el control de calidad de un lote de muestras, de acuerdo con algunas realizaciones de la presente divulgación.
- La figura 16 ilustra un ejemplo de procedimiento para validar un cambio en la canalización de datos bioinformáticos, por ejemplo, un proceso de expresión de ARN, de acuerdo con algunas realizaciones de la presente divulgación.
- 50 La figura 17 ilustra un ejemplo de procedimiento para ampliar una base de datos de referencia, de acuerdo con algunas realizaciones de la presente divulgación.

Los números de referencia semejantes se refieren a partes correspondientes a lo largo de las diversas vistas de los dibujos, de acuerdo con algunas realizaciones de la presente divulgación.

Descripción detallada

La divulgación en el presente documento proporciona procedimientos mejorados para realizar análisis de control de calidad de lotes (por ejemplo, valor de celdas de flujo enteras) de muestras de secuencias de ARN. Las pruebas de control de calidad ("quality control", QC) se realizan en lotes de muestras de secuenciación de ARN que incluyen lecturas de secuenciación y metadatos de muestras. Los procedimientos descritos en el presente documento sirven para garantizar una calidad de datos suficiente para realizar informes automatizados de llamadas de expresión de ARN y para detectar efectos de lote que puedan influir en los informes clínicos. Estos procedimientos también sirven para garantizar la coherencia de la calidad de los datos, lo cual es importante para compararlos a lo largo del tiempo. Los procedimientos de control de calidad descritos en el presente documento permiten la revisión y el análisis automáticos de toda una celda de flujo de muestras.

Beneficios

La presente divulgación proporciona un procedimiento novedoso que evalúa los efectos de lote técnicos en un conjunto de muestras de transcriptoma (por ejemplo, una celda de flujo) agrupándolas con un conjunto de muestras de referencia validadas emparejadas por el tipo de cáncer y el emplazamiento del tejido (por ejemplo, un lote de referencia emparejado por cohorte). Estos procedimientos suponen una mejora con respecto a la técnica anterior, ya que permiten realizar análisis globales simultáneos (por ejemplo, realizados en todo el conjunto de muestras de un lote) y análisis de control de calidad de una sola muestra. Estos procedimientos de control de calidad benefician a los pacientes, ya que proporcionan análisis rápidos y precisos de la calidad de las muestras y, por tanto, permiten mejorar y agilizar el diagnóstico y el tratamiento de los pacientes.

Los laboratorios que realizan grandes volúmenes de secuenciación de ARN deben ser extremadamente cautelosos con los efectos de lote técnicos si las muestras se van a comparar a lo largo de periodos de tiempo prolongados, lo cual es un imperativo para los análisis de los transcriptomas del cáncer y la determinación de la respuesta de los pacientes a los tratamientos. Los cambios en los reactivos, los protocolos o las tecnologías utilizadas en la extracción de ácidos nucleicos, la preparación de bancos y la secuenciación pueden alterar los transcriptomas de forma que invaliden o compliquen las comparaciones de muestras de diferentes lotes, lo que exige una supervisión continua de la calidad y la coherencia de las muestras. Este seguimiento puede resultar especialmente difícil cuando se analizan muestras procedentes de tejidos distintos, ya que el tipo de tumor es el principal determinante biológico de la variación del transcriptoma en el cáncer. Esto significa que se espera que las muestras de cáncer de cerebro y de hígado, por ejemplo, difieran tanto por lo que se refiere al transcriptoma que su comparación no sea informativa para la detección del efecto de lote. El hecho de que los procedimientos descritos en el presente documento proporcionen una correspondencia de cohortes entre las muestras de referencia y las muestras de cada celda de flujo individual hace que estas métricas de control de calidad sean más precisas que los procedimientos anteriores.

Definiciones

La terminología utilizada en la presente divulgación tiene por objeto únicamente describir realizaciones concretas y no pretende ser limitativa de la invención. Tal como se utilizan en la descripción de la invención y en las reivindicaciones adjuntas, se entiende que las formas en singular "un/una" y "el/la" incluyen también las formas en plural, a menos que el contexto indique claramente lo contrario. También se entenderá que el término "y/o", tal como se utiliza en el presente documento, se refiere y abarca todas y cada una de las combinaciones posibles de uno o más de los elementos enumerados asociados. Se entenderá además que los términos "incluye", "comprende" o cualquier variación de los mismos, cuando se utilicen en la presente memoria descriptiva, especifican la presencia de las características, números enteros, etapas, operaciones, elementos y/o componentes indicados, pero no excluyen la presencia o adición de una o más características, números enteros, etapas, operaciones, elementos, componentes y/o grupos de los mismos. Además, en la medida en que los términos "incluyendo", "incluye", "teniendo", "tiene", "con" o sus variantes se utilicen en la descripción detallada y/o en las reivindicaciones, dichos términos pretenden ser inclusivos de forma similar al término "comprende".

Tal como se utiliza en el presente documento, el término "si" puede interpretarse como "cuando" o "tras" o "en respuesta a la determinación" o "en respuesta a la detección", en función del contexto. Del mismo modo, la frase "si se determina" o "si se detecta [una condición o acontecimiento indicado]" puede interpretarse en el sentido de "tras determinar" o "en respuesta a la determinación" o "tras detectar [la condición o acontecimiento indicado]" o "en respuesta a la detección de [la condición o acontecimiento indicado]", en función del contexto.

También se entenderá que, aunque los términos primero, segundo, etc. pueden utilizarse en el presente documento para describir diversos elementos, estos elementos no deben ser limitados por estos términos. Estos términos sólo se utilizan para distinguir un elemento de otro. Por ejemplo, un primer sujeto podría denominarse segundo sujeto y, del mismo modo, un segundo sujeto podría denominarse primer sujeto, sin apartarse del alcance de la presente divulgación. El primer sujeto y el segundo sujeto son ambos sujetos, pero no son el mismo sujeto. Además, los términos "sujeto", "usuario" y "paciente" se utilizan indistintamente en el presente documento.

Tal como se utilizan en el presente documento, los términos "sujeto" o "paciente" se refieren a cualquier ser humano vivo o no vivo (por ejemplo, un humano varón, una humana mujer, un feto, una mujer embarazada, un niño o similar).

En algunas realizaciones, un sujeto es un varón o una mujer de cualquier etapa (por ejemplo, un hombre, una mujer o un niño).

Tal como se utilizan en el presente documento, los términos y expresiones "control", "muestra de control", "referencia", "muestra de referencia", "normal" y "muestra normal" describen una muestra de un sujeto que no tiene una afección concreta o que, por lo demás, está sano. En un ejemplo, un procedimiento como el divulgado en el presente documento puede realizarse en un sujeto que tiene un tumor, en el que la muestra de referencia es una muestra tomada de un tejido sano de dicho sujeto. En algunas realizaciones, se puede obtener una muestra de referencia del sujeto (por ejemplo, para que sirva como control de referencia para el sujeto en un momento determinado). En algunas realizaciones, una muestra de referencia puede obtenerse de una base de datos. La referencia puede ser, por ejemplo, un genoma de referencia que se utiliza para cartografiar las lecturas de secuencia obtenidas de la secuenciación de una muestra del sujeto. Un genoma de referencia puede referirse a un genoma haploide o diploide con el que las lecturas de secuencia de la muestra biológica y de una muestra constitucional pueden alinearse y compararse. Un ejemplo de muestra constitucional puede ser ADN de glóbulos blancos obtenido del sujeto. En un genoma haploide, sólo puede haber un solo nucleótido en cada locus. En un genoma diploide, pueden identificarse loci heterocigóticos; cada locus heterocigótico puede tener dos alelos, en los que cualquiera de ellos puede permitir una coincidencia para el alineamiento con el locus.

Tal como se utiliza en el presente documento, el término "locus" se refiere a una posición (por ejemplo, un sitio) dentro de un genoma, tal como en un cromosoma concreto. En algunas realizaciones, un locus se refiere a una posición de un único nucleótido dentro de un genoma, por ejemplo, en un cromosoma concreto. En algunas realizaciones, un locus se refiere a un pequeño grupo de posiciones de nucleótidos dentro de un genoma, por ejemplo, tal como se define para una mutación (por ejemplo, sustitución, inserción o delección) de nucleótidos consecutivos dentro de un genoma del cáncer. Dado que las células normales de los mamíferos tienen genomas diploides, un genoma normal de mamífero (por ejemplo, un genoma humano) tendrá generalmente dos copias de cada locus en el genoma, o al menos dos copias de cada locus ubicadas en los cromosomas autosómicos, por ejemplo, una copia en el cromosoma autosómico materno y una copia en el cromosoma autosómico paterno.

Tal como se utiliza en el presente documento, el término "alelo" se refiere a una secuencia concreta de uno o más nucleótidos en un locus cromosómico.

Tal como se utiliza en el presente documento, la expresión "alelo de referencia" se refiere a la secuencia de uno o más nucleótidos en un locus cromosómico que es el alelo predominante representado en ese locus cromosómico dentro de la población de la especie (por ejemplo, la secuencia "de tipo salvaje"), o un alelo que está predefinido dentro de un genoma de referencia para la especie.

Tal como se utiliza en el presente documento, la expresión "alelo variante" se refiere a una secuencia de uno o más nucleótidos en un locus cromosómico que no es el alelo predominante representado en ese locus cromosómico dentro de la población de la especie (por ejemplo, no la secuencia "de tipo salvaje"), o no un alelo que está predefinido dentro de un genoma de referencia para la especie.

Tal como se utilizan en el presente documento, las expresiones y términos "variante de un solo nucleótido", "SNV", "polimorfismo de un solo nucleótido" o "SNP" se refieren a una sustitución de un nucleótido por un nucleótido diferente en una posición (por ejemplo, sitio) de una secuencia de nucleótidos, por ejemplo, una lectura de secuencia de un individuo. Una sustitución de una primera nucleobase X por una segunda nucleobase Y puede indicarse como "X>Y". Por ejemplo, un SNP de citosina a timina puede indicarse como "C>T". El término "het-SNP" se refiere a un SNP heterocigoto, en el que el genoma es al menos diploide y al menos una, pero no todas, de las dos o más secuencias homólogas presenta el SNP concreto. Del mismo modo, un "hom-SNP" es un SNP homólogo, en el que cada secuencia homóloga de un genoma poliploide tiene la misma variante en comparación con el genoma de referencia. Tal como se utilizan en el presente documento, la expresión "variante estructural" o el término "VS" se refieren a regiones grandes (por ejemplo, mayores de 1 kb) de un genoma que han sufrido transformaciones físicas, tales como inversiones, inserciones, delecciones o duplicaciones (por ejemplo, véase el análisis de las VS del genoma humano de Spielmann *et al.*, 2018, Nat. Rev. Genetics, 19:453-467).

Tal como se utiliza en el presente documento, el término "indel" se refiere a acontecimientos de inserción y/o delección de tramos de uno o más nucleótidos, ya sea dentro de un único locus génico o a través de múltiples genes.

Tal como se utilizan en el presente documento, las expresiones y el término "variante de número de copias", "CNV" o "variación de número de copias" se refieren a regiones de un genoma que se repiten. Pueden clasificarse como repeticiones cortas o largas, en función del número de nucleótidos que se repiten en las regiones del genoma. Las repeticiones largas suelen referirse a casos en los que genes enteros, o grandes porciones de un gen, se repiten una o más veces.

Tal como se utiliza en el presente documento, el término "mutación" se refiere a un cambio detectable en el material genético de una o más células. En un ejemplo concreto, una o más mutaciones pueden encontrarse en células cancerosas y pueden identificarlas (por ejemplo, mutaciones conductoras y pasajeras). Una mutación puede transmitirse de una célula progenitora a una célula hija. Una persona con conocimientos en la materia apreciará que

una mutación genética (por ejemplo, una mutación conductora) en una célula progenitora puede inducir mutaciones adicionales diferentes (por ejemplo, mutaciones pasajeras) en una célula hija. Una mutación se produce generalmente en un ácido nucleico. En un ejemplo concreto, una mutación puede ser un cambio detectable en uno o más ácidos desoxirribonucleicos o fragmentos de los mismos. Una mutación se refiere generalmente a nucleótidos que se añaden, 5 delecionan, sustituyen, invierten o transponen a una nueva posición en un ácido nucleico. Una mutación puede ser espontánea o inducida experimentalmente. Una mutación en la secuencia de un tejido concreto es un ejemplo de "alelo específico de tejido". Por ejemplo, un tumor puede tener una mutación que dé lugar a un alelo en un locus que no aparece en las células normales. Otro ejemplo de "alelo específico de tejido" es un alelo específico del feto que se produce en el tejido fetal, pero no en el tejido materno.

10 Tal como se utiliza en el presente documento, la expresión "variante genómica" puede referirse a una o más mutaciones, variantes de número de copias, indels, variantes de un solo nucleótido o alelos variantes. Una variante genómica también puede referirse a una combinación de uno o más de los anteriores.

Tal como se utilizan en el presente documento, los términos y la expresión "cáncer", "tejido canceroso" o "tumor" se refieren a una masa anómala de tejido en la que el crecimiento de la masa supera al crecimiento del tejido normal y no está coordinado con aquel. En el caso de los cánceres hematológicos, esto incluye un volumen de sangre u otro líquido corporal que contenga células cancerosas. Un cáncer o tumor puede definirse como "benigno" o "maligno" en 15 función de las siguientes características: grado de diferenciación celular, incluida la morfología y la funcionalidad, tasa de crecimiento, invasión local y metástasis. Un tumor "benigno" puede estar bien diferenciado, tener un crecimiento característicamente más lento que un tumor maligno y permanecer localizado en el lugar de origen. Además, en algunos casos, un tumor benigno no tiene la capacidad de infiltrarse, invadir o metastatizar a lugares distantes. Un tumor "maligno" puede ser poco diferenciado (anaplasia), tener un crecimiento característicamente rápido 20 acompañado de infiltración progresiva, invasión y destrucción del tejido circundante. Además, un tumor maligno puede tener la capacidad de metastatizar a lugares distantes. Por consiguiente, una célula cancerosa es una célula que se encuentra dentro de la masa anómala de tejido cuyo crecimiento no está coordinado con el crecimiento del tejido normal. En consecuencia, una "muestra tumoral" o "biopsia somática" se refiere a una muestra biológica obtenida o 25 derivada de un tumor de un sujeto, tal como se describe en el presente documento.

Tal como se utiliza en el presente documento, la expresión "biopsia somática" se refiere a una biopsia de un sujeto. En algunas realizaciones, la biopsia es de tejido sólido. En algunos casos, se trata de una biopsia líquida.

30 Tal como se utilizan en el presente documento, el término y la expresión "secuenciación", "determinación de la secuencia" y similares tal como se utilizan en el presente documento se refieren en general a todos y cada uno de los procesos bioquímicos que pueden utilizarse para determinar el orden de macromoléculas biológicas, tales como ácidos nucleicos o proteínas. Por ejemplo, los datos de secuenciación pueden incluir la totalidad o una parte de las bases nucleotídicas de una molécula de ácido nucleico, tal como un transcrito de ARNm o un locus genómico.

35 Tal como se utilizan en el presente documento, la expresión y el término "lecturas de secuencia" o "lecturas" se refieren a secuencias de nucleótidos producidas por cualquier proceso de secuenciación descrito en el presente documento o conocido en la técnica. Las lecturas pueden generarse a partir de un extremo de los fragmentos de ácido nucleico ("lecturas de un solo extremo"), y a veces se generan a partir de ambos extremos de los ácidos nucleicos (por ejemplo, lecturas de extremos emparejados, lecturas de dos extremos). La longitud de la lectura de la secuencia suele estar asociada a la tecnología de secuenciación concreta. Los procedimientos de alto rendimiento, por ejemplo, 40 proporcionan lecturas de secuencia cuyo tamaño puede variar entre decenas y cientos de pares de bases (pb). En algunas realizaciones, las lecturas de secuencia tienen una mediana de longitud, una longitud media o promedio de aproximadamente 15 pb a 900 pb de longitud (por ejemplo, aproximadamente 20 pb, aproximadamente 25 pb, aproximadamente 30 pb, aproximadamente 35 pb, aproximadamente 40 pb, aproximadamente 45 pb, aproximadamente 50 pb, aproximadamente 55 pb, aproximadamente 60 pb, aproximadamente 65 pb, 45 aproximadamente 70 pb, aproximadamente 75 pb, aproximadamente 80 pb, aproximadamente 85 pb, aproximadamente 90 pb, aproximadamente 95 pb, aproximadamente 100 pb, aproximadamente 110 pb, aproximadamente 120 pb, aproximadamente 130, aproximadamente 140 pb, aproximadamente 150 pb, aproximadamente 200 pb, aproximadamente 250 pb, aproximadamente 300 pb, aproximadamente 350 pb, aproximadamente 400 pb, aproximadamente 450 pb o aproximadamente 500 pb. En algunas realizaciones, las 50 lecturas de secuencia tienen una mediana de longitud, una longitud media o promedio de aproximadamente 1000 pb, 2000 pb, 5000 pb, 10 000 pb o 50 000 pb o más. La secuenciación por nanoporos, por ejemplo, puede proporcionar lecturas de secuencia cuyo tamaño puede variar de decenas a cientos o miles de pares de bases. La secuenciación paralela de Illumina puede proporcionar lecturas de secuencia que no varían tanto; por ejemplo, la mayoría de las lecturas de secuencia pueden ser inferiores a 200 pb. Una lectura de secuencia (o lectura de secuenciación) puede 55 referirse a la información de secuencia correspondiente a una molécula de ácido nucleico (por ejemplo, una cadena de nucleótidos). Por ejemplo, una lectura de secuencia puede corresponder a una cadena de nucleótidos (por ejemplo, de aproximadamente 20 a aproximadamente 150) procedente de parte de un fragmento de ácido nucleico, puede corresponder a una cadena de nucleótidos en uno o ambos extremos de un fragmento de ácido nucleico, o puede corresponder a nucleótidos de todo el fragmento de ácido nucleico. Una lectura de secuencia puede obtenerse de 60 diversas maneras, por ejemplo, mediante técnicas de secuenciación o utilizando sondas, por ejemplo, en matrices de

hibridación o sondas de captura, o técnicas de amplificación, tal como la reacción en cadena de la polimerasa (PCR) o la amplificación lineal utilizando un solo cebador o la amplificación isotérmica.

Tal como se utilizan en el presente documento, la expresión y el término "segmento de lectura" o "lectura" se refieren a cualquier secuencia de nucleótidos, incluidas lecturas de secuencia obtenidas de un individuo y/o secuencias de nucleótidos derivadas de la lectura de secuencia inicial procedente de una muestra obtenida de un individuo. Por ejemplo, un segmento de lectura puede referirse a una lectura de secuencia alineada, una lectura de secuencia colapsada o una lectura cosida. Además, un segmento de lectura puede referirse a una base nucleotídica individual, tal como una variante de un solo nucleótido.

Tal como se utilizan en el presente documento, las expresiones y el término "profundidad de lectura", "profundidad de secuenciación" o "profundidad" se refieren a un número total de segmentos de lectura de una muestra obtenida de un individuo en una posición, región o locus determinados. El locus puede ser tan pequeño como un nucleótido, tan grande como un brazo cromosómico o tan grande como un genoma entero. La profundidad de secuenciación puede expresarse como "Yx", por ejemplo, 50x, 100x, etc., en la que "Y" se refiere al número de veces que se cubre un locus con una lectura de secuencia. En algunas realizaciones, la profundidad se refiere a la profundidad de secuenciación promedio a través del genoma, a través del exoma, a través del transcriptoma, o a través de un panel de secuenciación selectiva. La profundidad de secuenciación también puede aplicarse a múltiples loci, al genoma completo, en cuyo caso Y puede referirse al número medio de veces que se secuencia un loci o un genoma haploide, un genoma completo, un transcriptoma completo o un exoma completo, respectivamente. Cuando se indica una profundidad media, la profundidad real de los distintos loci incluidos en el conjunto de datos puede abarcar una gama de valores. La secuenciación ultraprofunda puede referirse a una profundidad de secuenciación de al menos 100x en un locus.

Tal como se utiliza en el presente documento, la expresión "amplitud de secuenciación" se refiere a qué fracción de un transcriptoma de referencia concreto (por ejemplo, un exoma de referencia humano), un genoma de referencia concreto (por ejemplo, un genoma de referencia humano), o parte del transcriptoma o genoma se ha analizado. El denominador de la fracción puede ser un genoma con enmascaramiento por repeticiones, por lo que el 100 % puede corresponder a todo el genoma de referencia menos las partes enmascaradas. Un transcriptoma o genoma con enmascaramiento por repeticiones puede referirse a un transcriptoma o genoma en el que las repeticiones de secuencias están enmascaradas (por ejemplo, las lecturas de secuencia se alinean con porciones no enmascaradas del transcriptoma o genoma). Cualquier parte de un transcriptoma o genoma puede estar enmascarado y, de este modo, es posible centrarse en cualquier parte concreta de un exoma o genoma de referencia. La secuenciación amplia puede referirse a la secuenciación y análisis de al menos el 0,1 % del transcriptoma o genoma de referencia.

Tal como se utiliza en el presente documento, la expresión "transcriptoma de referencia" se refiere a cualquier transcriptoma concreto conocido, secuenciado o caracterizado, ya sea parcial o completo, de cualquier tejido de cualquier organismo o patógeno que pueda utilizarse para referenciar secuencias identificadas de un sujeto. Se proporcionan ejemplos de transcriptomas de referencia utilizados para sujetos humanos en la base de datos en línea MiTranscriptome descrita por Iyer *et al.*, 2015, The landscape of long noncoding RNAs in the human transcriptome, Nat. Genet., 47, 199-208, la base de datos CHESS que describen Pertea *et al.*, 2018, CHESS: a new human gene catalog curated from thousands of large-scale RNA sequencing experiments reveals extensive transcriptional noise, Gen. Biol., 19:208, y en la base de datos ENCODE en línea alojada en el proyecto ENCODE.

Tal como se utiliza en el presente documento, la expresión "llamada de expresión" se refiere a una llamada diferencial de expresión de ARN (por ejemplo, una determinación de si una muestra concreta de un sujeto presenta una expresión mayor o menor para un ARN concreto en comparación con un transcriptoma de referencia). En algunas realizaciones, una llamada de expresión se basa, al menos en parte, en recuentos de abundancia de genes.

Tal como se utiliza en el presente documento, la expresión "exoma de referencia" se refiere a cualquier exoma concreto conocido, secuenciado o caracterizado, ya sea parcial o completo, de cualquier tejido de cualquier organismo o patógeno que pueda utilizarse para referenciar secuencias identificadas de un sujeto. Se proporcionan ejemplos de exomas de referencia utilizados para sujetos humanos, así como para muchos otros organismos, en la base de datos GENCODE en línea alojada por el consorcio GENCODE, por ejemplo, la versión 29 (GRCh38.p12) del ensamblaje del exoma humano.

Tal como se utiliza en el presente documento, la expresión "genoma de referencia" se refiere a cualquier genoma concreto conocido, secuenciado o caracterizado, ya sea parcial o completo, de cualquier organismo o patógeno que pueda utilizarse para referenciar secuencias identificadas de un sujeto. En el buscador de genomas en línea alojado por el National Center for Biotechnology Information ("NCBI") o la Universidad de California, Santa Cruz (UCSC), se ofrecen ejemplos de genomas de referencia utilizados para sujetos humanos, así como para muchos otros organismos. Por "genoma" se entiende la información genética completa de un organismo o patógeno, expresada en secuencias de ácidos nucleicos. Tal como se utiliza en el presente documento, una secuencia de referencia o genoma de referencia es a menudo una secuencia genómica ensamblada o parcialmente ensamblada de un individuo o de múltiples individuos. En algunas realizaciones, un genoma de referencia es una secuencia genómica ensamblada o parcialmente ensamblada de uno o más individuos humanos. El genoma de referencia puede considerarse un ejemplo representativo del conjunto de genes o secuencias genéticas de una especie. En algunas realizaciones, un genoma de referencia incluye secuencias asignadas a cromosomas. Algunos ejemplos de genomas humanos de referencia

son, entre otros, NCBI build 34 (equivalente UCSC: hg16), NCBI build 35 (equivalente UCSC: hg17), NCBI build 36.1 (equivalente UCSC: hg18), GRCh37 (equivalente UCSC: hg19) y GRCh38 (equivalente UCSC: hg38).

Tal como se utiliza en el presente documento, el término "ensayo" se refiere a una técnica para determinar una propiedad de una sustancia, por ejemplo, un ácido nucleico, una proteína, una célula, un tejido o un órgano. Un ensayo (por ejemplo, un primer ensayo o un segundo ensayo) puede comprender una técnica para determinar la variación del número de copias de ácidos nucleicos en una muestra, el estado de metilación de ácidos nucleicos en una muestra, la distribución de tamaño de fragmentos de ácidos nucleicos en una muestra, el estado mutacional de ácidos nucleicos en una muestra, o el patrón de fragmentación de ácidos nucleicos en una muestra. Para detectar cualquiera de las propiedades de los ácidos nucleicos mencionadas en el presente documento puede utilizarse cualquier ensayo conocido por una persona con conocimientos ordinarios en la materia. Las propiedades de los ácidos nucleicos pueden incluir una secuencia, identidad genómica, número de copias, estado de metilación en una o más posiciones nucleotídicas, tamaño del ácido nucleico, presencia o ausencia de una mutación en el ácido nucleico en una o más posiciones nucleotídicas, y patrón de fragmentación de un ácido nucleico (por ejemplo, la posición o posiciones nucleotídicas en las que se fragmenta un ácido nucleico). Un ensayo o procedimiento puede tener una sensibilidad y/o especificidad determinadas, y su utilidad relativa como herramienta de diagnóstico puede medirse utilizando la estadística ROC-AUC.

Tal como se utiliza en el presente documento, la expresión "abundancia relativa" puede referirse a una proporción entre una primera cantidad de fragmentos de ácido nucleico que tienen una característica concreta (por ejemplo, se alinean con una región concreta del exoma) y una segunda cantidad de fragmentos de ácido nucleico que tienen una característica concreta (por ejemplo, se alinean con una región concreta del exoma). En un ejemplo, la abundancia relativa puede referirse a una proporción entre el número de transcritos de ARNm que codifican un gen concreto en una muestra (por ejemplo, se alinean con una región concreta del exoma) y el número total de transcritos de ARNm en la muestra.

Tal como se utiliza en el presente documento, dos conjuntos de datos están "equilibrados" con respecto a una característica cuando el porcentaje del número muestras que tienen cada tipo de característica en los dos conjuntos de datos está dentro de un porcentaje establecido de cada uno. A menos que se especifique lo contrario, dos conjuntos de datos están equilibrados con respecto a una característica cuando el porcentaje del número de muestras que tienen cada tipo de característica en los dos conjuntos de datos está dentro del 10 %. Por ejemplo, si en un conjunto de datos de lotes, el 15 % de las muestras son muestras de cáncer de pulmón, el 25 % de las muestras son muestras de cáncer cerebral y el 60 % de las muestras son muestras de cáncer de colon, se considera que un conjunto de datos de referencia está equilibrado con el conjunto de datos de lotes si en el conjunto de datos de referencia, entre el 5 % y el 25 % de las muestras son muestras de cáncer de pulmón, entre el 15 % y el 35 % de las muestras son muestras de cáncer cerebral y entre el 50 % y el 70 % de las muestras son muestras de cáncer de colon. En algunas realizaciones, dos conjuntos de datos están equilibrados con respecto a una característica cuando el porcentaje del número de muestras que tienen cada tipo de característica en los dos conjuntos de datos está dentro del 1 %, dentro del 2 %, dentro del 3 %, dentro del 4 %, dentro del 5 %, dentro del 6 %, dentro del 7 %, dentro del 8 %, dentro del 9 %, dentro del 10 %, dentro del 11 %, dentro del 12 %, dentro del 13 %, dentro del 14 %, dentro del 15 %, dentro del 16 %, dentro del 17 %, dentro del 18 %, dentro del 19 %, dentro del 20 %, dentro del 21 %, dentro del 22 %, dentro del 23 %, dentro del 24 % o dentro del 25 % de cada uno. Generalmente, el equilibrio de una primera característica y el equilibrio de una segunda característica se consideran independientes entre sí. Sin embargo, en algunas realizaciones, el equilibrio de una primera característica y el equilibrio de una segunda característica se consideran conjuntamente. Es decir, en algunas realizaciones, es un elemento compuesto de al menos dos características que está equilibrado. Por ejemplo, el porcentaje de muestras de cáncer cerebral tomadas de un tejido cutáneo y el porcentaje de muestras de cáncer cerebral tomadas de un tejido pulmonar en un conjunto de datos de lotes están equilibrados con el porcentaje de muestras de cáncer cerebral tomadas de un tejido cutáneo y el porcentaje de muestras de cáncer cerebral tomadas de un tejido pulmonar en un conjunto de datos de referencia, a diferencia de equilibrar simplemente el porcentaje de muestras de cáncer cerebral, muestras tomadas de tejido cutáneo y muestras tomadas de tejido pulmonar. En algunas realizaciones, una característica poco frecuente en un conjunto de datos de lotes no estará equilibrada en el conjunto de datos de referencia, por ejemplo, debido a la falta de disponibilidad de un número suficiente de muestras de referencia que compartan la característica poco frecuente.

A continuación, se describen varios aspectos con referencia a aplicaciones de ejemplo a título ilustrativo. Debe entenderse que numerosos detalles específicos, relaciones y procedimientos se exponen para proporcionar una comprensión completa de las características descritas en el presente documento. Sin embargo, una persona con conocimientos ordinarios en la materia reconocerá con facilidad que las características descritas en el presente documento pueden llevarse a la práctica sin uno o más de los detalles específicos o con otros procedimientos. Las características descritas en el presente documento no están limitadas por el orden ilustrado de los actos o acontecimientos, ya que algunos actos pueden aparecer en diferentes órdenes y/o simultáneamente con otros actos o acontecimientos. Además, no todos los actos o acontecimientos ilustrados son necesarios para aplicar una metodología de acuerdo con las características descritas en el presente documento.

A continuación, se hará referencia en detalle a realizaciones, ejemplos de las cuales se ilustran en los dibujos adjuntos. En la siguiente descripción detallada, se exponen numerosos detalles específicos con el fin de proporcionar una

comprensión completa de la presente divulgación. Sin embargo, será evidente para una persona con conocimientos ordinarios en la materia que la presente divulgación puede llevarse a la práctica sin estos detalles específicos. En otros casos, no se han descrito en detalle métodos, procedimientos, componentes, circuitos y redes bien conocidos para no complicar innecesariamente aspectos de las realizaciones.

5 Ejemplos de realizaciones de sistemas.

Ahora que se ha proporcionado una visión general de algunos aspectos de la presente divulgación y algunas definiciones utilizadas en la presente divulgación, se describen detalles de un ejemplo de sistema junto con la figura 1. La figura 1 es un diagrama de bloques que ilustra un sistema 100 de acuerdo con algunas aplicaciones. El sistema 100 en algunas aplicaciones incluye una o más unidades de procesamiento CPU 102 (también denominadas procesadores), una o más interfaces de red 104, una interfaz de usuario 106 que incluye (opcionalmente) una pantalla 108 y un sistema de entrada 110, una memoria no persistente 111, una memoria persistente 112, y uno o más buses de comunicaciones 114 para interconectar estos componentes. Dichos uno o más buses de comunicaciones 114 incluyen opcionalmente circuitos (a veces llamados conjunto de circuitos integrados) que interconectan y controlan las comunicaciones entre los componentes del sistema. La memoria no persistente 111 suele incluir una memoria de acceso aleatorio de alta velocidad, tal como DRAM, SRAM, DDR RAM, ROM, EEPROM, memoria flash, mientras que la memoria persistente 112 suele incluir CD-ROM, discos versátiles digitales (DVD) u otro almacenamiento óptico, casetes magnéticos, cinta magnética, almacenamiento en disco magnético, dispositivos de almacenamiento magnético, dispositivos de almacenamiento en disco óptico, dispositivos de memoria flash u otros dispositivos de almacenamiento de estado sólido no volátil. La memoria persistente 112 incluye opcionalmente uno o más dispositivos de almacenamiento ubicados remotamente de dichas una o más CPU 102. La memoria persistente 112, y el dispositivo o dispositivos de memoria no volátil dentro de la memoria no persistente 112, comprenden un medio de almacenamiento legible por ordenador no transitorio. En algunas aplicaciones, la memoria no persistente 111 o, como alternativa, el medio de almacenamiento legible por ordenador no transitorio, almacena los siguientes programas, módulos y estructuras de datos, o un subconjunto de los mismos, a veces junto con la memoria persistente 112:

- un sistema operativo opcional 116, que incluye procedimientos para manejar diversos servicios básicos del sistema y para realizar tareas dependientes del hardware;
- un módulo opcional de comunicación de red (o instrucciones) 118 para conectar el sistema 100 con otros dispositivos y/o una red de comunicaciones 104;
- un módulo de control de calidad de lotes 120 para realizar el control de calidad de lotes de muestras de secuenciación de ARN;
- uno o más conjuntos de datos de lotes 122, en los que cada conjunto de datos de lotes incluye, para cada muestra 124 de una pluralidad (por ejemplo, un lote) de muestras, al menos una pluralidad correspondiente de lecturas de secuencia 126 y los metadatos de muestra 128 correspondientes, y también se incluye en cada conjunto de datos de lotes un lote de referencia 130 correspondiente emparejado por cohorte;
- un conjunto de datos de muestras de referencia 140 que almacena una o más muestras de referencia 142, en el que cada muestra de referencia incluye al menos una pluralidad correspondiente de secuencias de muestras de referencia 144 y los correspondientes metadatos de muestras de referencia 146;
- un conjunto de datos de control de calidad global 150 para almacenar una o más pruebas de control de calidad de lotes 152, en el que dichas uno o más pruebas de control de calidad de lotes 152 se realizan, a través del módulo de control de calidad 120, en la pluralidad de muestras (por ejemplo, 124-1, ..., 124-A) incluidas en un conjunto de datos de lotes (por ejemplo, el conjunto de datos de lotes 122-1);
- un módulo de control de calidad de una sola muestra 121 para realizar el control de calidad de una sola muestra de secuenciación de ARN; y
- un conjunto de datos de control de calidad de una sola muestra 160 para almacenar una o más pruebas de control de calidad de una sola muestra 162, en el que dichas una o más pruebas de control de calidad de una sola muestra 162 se realizan, a través del módulo de control de calidad 120, en muestras individuales 124 incluidas en un conjunto de datos de lotes 122.

En diversas realizaciones, uno o más de los elementos identificados anteriormente se almacenan en uno o más de los dispositivos de memoria mencionados anteriormente, y corresponden a un conjunto de instrucciones para realizar una función descrita anteriormente. Los módulos, datos o programas identificados anteriormente (por ejemplo, conjuntos de instrucciones) no necesitan aplicarse como programas informáticos, procedimientos, conjuntos de datos o módulos separados, y, por lo tanto, varios subconjuntos de estos módulos y datos pueden combinarse o reorganizarse de otro modo en varias aplicaciones. En algunas aplicaciones, la memoria no persistente 111 almacena opcionalmente un subconjunto de los módulos y estructuras de datos identificados anteriormente. Además, en algunas realizaciones, la

memoria almacena módulos y estructuras de datos adicionales no descritos anteriormente. En algunas realizaciones, uno o más de los elementos identificados anteriormente se almacenan en un sistema informático, distinto al del sistema de visualización 100, que es direccionable por el sistema de visualización 100 para que el sistema de visualización 100 pueda recuperar la totalidad o una parte de dichos datos cuando sea necesario.

- 5 Aunque la figura 1 representa un "sistema 100", la figura pretende ser más una descripción funcional de las diversas características que pueden estar presentes en los sistemas informáticos que un esquema estructural de las aplicaciones descritas en el presente documento. En la práctica, y como reconocen los expertos en la materia, los elementos mostrados por separado podrían combinarse y algunos podrían separarse. Además, aunque la figura 1 muestra determinados datos y módulos en la memoria no persistente 111, algunos o todos estos datos y módulos
- 10 pueden almacenarse en la memoria persistente 112 o en más de una memoria. Por ejemplo, en algunas realizaciones, al menos un conjunto de datos de lotes 122 se almacena en un dispositivo de almacenamiento remoto, que puede formar parte de una infraestructura basada en la nube. En algunas realizaciones, al menos un conjunto de datos de lotes 122 se almacena en una infraestructura basada en la nube. En algunas realizaciones, el conjunto de datos de lotes 122, el módulo de control de calidad de lotes 120, el módulo de control de calidad de una sola muestra 121, el
- 15 conjunto de datos de referencia 140, el conjunto de datos de control de calidad global 150 y/o el conjunto de datos de control de calidad de una sola muestra 160 también pueden almacenarse en el dispositivo o dispositivos de almacenamiento remoto. En algunas realizaciones, se emplean otras configuraciones de almacenamiento de datos y módulos.

Análisis de lotes y muestras

- 20 Habiéndose divulgado los detalles de un sistema 100 de acuerdo con la presente divulgación, a continuación se divulgan detalles relativos a procesos y características del sistema, de acuerdo con diversas realizaciones de la presente divulgación. Específicamente, a continuación se describen ejemplos de procesos con referencia a las figuras 2A, 2B, 15, 16 y 17. En algunas realizaciones, tales procesos y características del sistema son llevados a cabo por los módulos 118, 120 y/o 121, como se ilustra en la figura 1.
- 25 *Bloque 202.* En referencia al bloque 202 de la figura 2A, el procedimiento realiza un control de calidad. En algunas realizaciones, el control de calidad se lleva a cabo en un conjunto de datos de un solo lote (por ejemplo, que comprende una sola celda de flujo de muestras de ARN). En algunas realizaciones, el control de calidad se lleva a cabo simultáneamente en dos o más conjuntos de datos de lotes (por ejemplo, dos o más celdas de flujo de muestras de ARN, en las que cada celda de flujo se analiza en el mismo día o en un día diferente). En algunas realizaciones, el
- 30 control de calidad se lleva a cabo simultáneamente en una pluralidad de conjuntos de datos de lotes (por ejemplo, en una pluralidad de celdas de flujo).

- Haciendo referencia al bloque 204, en algunas realizaciones, los procedimientos descritos en el presente documento se llevan a cabo en un sistema informático que comprende un servidor en la nube. Es decir, en algunas realizaciones, los procedimientos descritos en el presente documento pueden llevarse a cabo total o parcialmente en sistemas
- 35 remotos. Por ejemplo, como se ha descrito anteriormente, en algunas realizaciones, uno o más de los conjuntos de datos se almacenan localmente, y al menos uno de los módulos de control de calidad de lotes 120 y/o el módulo de control de calidad de una sola muestra 121 se almacenan en un servidor en la nube (por ejemplo, en la nube). En algunas realizaciones, el conjunto de datos de referencia 140, el conjunto de datos de control de calidad global 150 y/o el conjunto de datos de control de calidad de una sola muestra 160 también se almacenan en un servidor en la
- 40 nube. En algunas realizaciones, los datos necesarios (por ejemplo, uno o más conjuntos de datos por lotes 122) pueden transmitirse entre los servidores locales y de la nube.

Obtener la información inicial del lote de datos 122

- Bloque 206.* En referencia al bloque 206 de la figura 2A, se obtiene un conjunto de datos de lotes en formato electrónico (por ejemplo, la información de cada muestra del conjunto de datos de lotes se almacena en un archivo .csv). El
- 45 conjunto de datos de lotes comprende, para cada muestra respectiva de una pluralidad (por ejemplo, lote) de muestras, una pluralidad correspondiente de lecturas de secuencia derivadas de la muestra respectiva mediante secuenciación de panel selectiva o del transcriptoma completo. En algunas realizaciones, cada pluralidad correspondiente de lecturas de secuencia se obtiene a partir de una pluralidad de moléculas de ARN o derivados de dicha pluralidad de moléculas de ARN (por ejemplo, derivados tales como ADNc). En algunas realizaciones, cada pluralidad correspondiente de
- 50 lecturas de secuencia se obtiene mediante secuenciación completa del transcriptoma. En algunas realizaciones, una o más pluralidades correspondientes de lecturas de secuencia se derivan de ARN que se ha aislado de un tumor sólido o hematológico (por ejemplo, una biopsia sólida). En algunas realizaciones, una o más pluralidades correspondientes de lecturas de secuencia se derivan de muestras de la línea germinal obtenidas de los respectivos sujetos.

- En algunas realizaciones, una o más pluralidades correspondientes de lecturas de secuencia se generan mediante
- 55 secuenciación de nueva generación. En algunas realizaciones, una o más pluralidades correspondientes de lecturas de secuencia se generan a partir de secuenciación de nueva generación de extremos emparejados de lectura corta. En algunas realizaciones, una o más pluralidades correspondientes de lecturas de secuencia se generan a partir de secuenciación de nueva generación de lectura corta con uno o más controles sembrados. En algunas realizaciones, dichos uno o más controles sembrados calibran la variación en las lecturas de secuencia a través de una población

de células (por ejemplo, el volumen de lecturas de ARN obtenidas de cada célula puede variar significativamente y el sembrado ayuda a normalizar las lecturas a través de un conjunto de células). En algunas realizaciones, una o más pluralidades correspondientes de lecturas de secuencia se obtienen mediante secuenciación de panel selectiva utilizando una pluralidad de sondas.

Los procedimientos para la secuenciación de ARNm son bien conocidos en la técnica. En algunas realizaciones, el ARNm se transcribe inversamente a ADNc antes de la secuenciación. Por ejemplo, los procedimientos de RNA-Seq para su uso de acuerdo con el bloque 210 se divulgan en Nagalakshmi *et al.*, 2008, Science, 320, 1344-1349; y Finotell y Camillo, 2014, Briefings in Functional Genomics, 14(2), 130-142. En algunas realizaciones, la secuenciación del ARNm se lleva a cabo mediante secuenciación del exoma completo ("whole exome sequencing", WES). En algunas realizaciones, la WES se lleva a cabo aislando el ARN de una muestra de tejido, generando un banco de ADNc, seleccionando opcionalmente las secuencias deseadas y/o eliminando las moléculas de ARN no deseadas, y secuenciando a continuación el banco de ADNc, por ejemplo, utilizando técnicas de secuenciación de nueva generación. Para un análisis del uso de técnicas de secuenciación del exoma completo en el diagnóstico del cáncer, véase, Serrati *et al.*, 2016, Onco Targets Ther., 9, 7355-7365 y Cieslik, M. *et al.*, 2015, Genome Res., 25, 1372-1381.

En algunas realizaciones, la secuenciación del ARNm se lleva a cabo mediante secuenciación por nanoporos. Se puede encontrar un análisis del uso de técnicas de secuenciación por nanoporos en el genoma humano en Jain *et al.*, 2018, Nature, 36(4), 338-345. Esta lista no es exhaustiva de los procedimientos de secuenciación de ARN que pueden utilizarse de acuerdo con los procedimientos descritos en el presente documento. En algunas realizaciones, la secuenciación del ARN se lleva a cabo de acuerdo con uno o más procedimientos de secuenciación conocidos en la técnica. Véase, por ejemplo, un análisis de los procedimientos de secuenciación del ARN en Kukurba *et al.*, 2015, Cold Spring Harb. Protoc., 11: 951-969.

Los procedimientos de secuenciación de nueva generación para su uso de acuerdo con los procedimientos descritos en el presente documento se divulgan en Shendure, 2008, Nat. Biotechnology, 26:1135-1145; y Fullwood *et al.*, 2009, Genome Res., 19:521-532. Los procedimientos de secuenciación de nueva generación bien conocidos en la técnica incluyen la tecnología de síntesis (Illumina), la pirosecuenciación (454 Life Sciences), la tecnología de semiconductores iónicos (secuenciación Ion Torrent), la secuenciación en tiempo real de una sola molécula (Pacific Biosciences), la secuenciación por ligadura (secuenciación SOLiD), la secuenciación por nanoporos (Oxford Nanopore Technologies) o la secuenciación de extremos emparejados. En algunas realizaciones, la secuenciación paralela masiva se lleva a cabo utilizando secuenciación por síntesis con terminadores de colorante reversibles.

La técnica RNA-Seq es una metodología para la obtención de perfiles de ARN basada en la secuenciación de nueva generación que permite la medición y comparación de patrones de expresión génica a través de una pluralidad de sujetos. En algunas realizaciones, se generan millones de cadenas cortas, denominadas "lecturas de secuencia", a partir de la secuenciación de posiciones aleatorias de ADNc preparado a partir de los ARN de entrada que se obtienen del tejido tumoral de un sujeto. En algunas realizaciones, los datos de expresión génica de RNA-Seq se generaron a partir de muestras tumorales fijadas en formol e incluidas en parafina utilizando un protocolo de RNA-Seq basado en captura del exoma. A continuación, estas lecturas pueden cartografiarse de modo informático sobre un genoma de referencia para revelar un "mapa transcripcional", en el que el número de lecturas de secuencia alineadas con cada gen da una medida de su nivel de expresión (por ejemplo, abundancia). En algunas realizaciones, los niveles de expresión de RNA-Seq (por ejemplo, recuentos de lecturas en bruto) se normalizan (por ejemplo, para corregir el contenido de GC, la profundidad de secuenciación y/o la longitud del gen). En algunas realizaciones, los procedimientos de cartografiados de lecturas de secuencia de ARN en bruto al transcriptoma, la cuantificación de los recuentos de genes y la normalización se llevan a cabo como se describe en la solicitud de Patente de EE. UU. n.º 62/735 349, titulada "Methods of Normalizing and Correcting RNA Expression Data", presentada el 24 de septiembre de 2018.

En algunas realizaciones alternativas, en lugar de utilizar RNA-Seq, se utilizan micromatrices para examinar el perfil de ARN. Tales micromatrices se divulgan en Wang *et al.*, 2009, Nat. Rev. Genet., 10, 57-63; Roy *et al.*, 2011, Brief Funct. Genomic., 10:135-150; Shendure, 2008, Nat. Methods, 5, 585-587; Cloonan *et al.*, 2008, "Stem cell transcriptome profiling via massive-scale mRNA sequencing", Nat. Methods, 5, 613-619; Mortazavi *et al.*, 2008, "Mapping and quantifying mammalian transcriptomes by RNA-Seq", Nat. Methods, 5, 621-628; y Bullard *et al.*, 2010, "Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments", BMC Bioinformatics, 11, pág. 94.

La primera etapa informática de la canalización del análisis de datos de RNA-Seq es el cartografiado de lecturas: las lecturas se alinean con un genoma o transcriptoma de referencia mediante la identificación de regiones génicas que coinciden con las secuencias de lectura. Para esta tarea se puede utilizar cualquier herramienta de alineación. Véase, por ejemplo, Hatem *et al.*, 2013, BMC Bioinformatics, 14, 184; y Engstrom *et al.*, 2013, Nat. Methods, 10, 1185-1191. En algunas realizaciones, el proceso de cartografiado comienza construyendo un índice del genoma de referencia o de las lecturas, que luego se utiliza para recuperar el conjunto de posiciones en la secuencia de referencia donde es más probable que las lecturas se alineen. Una vez identificado este subconjunto de posibles ubicaciones cartográficas, la alineación se realiza en estas regiones candidatas con algoritmos más lentos y sensibles. Véase, por ejemplo, Flicek y Birney, 2009, Nat. Methods, 6(supl. 11), S6-S12. En algunas realizaciones, la herramienta

de cartografiado es una metodología que hace uso de la pseudoalineación (por ejemplo, alineación de secuencias de lectura con transcritos, pero no con ubicaciones genómicas). Véase, por ejemplo, Bray *et al.*, 2016, Near-optimal probabilistic RNA-Seq quantification, *Nat. Biotech.*, 34, 525-527.

Después del cartografiado, las lecturas alineadas con cada unidad de codificación, tal como un exón, un transcrito o un gen, se utilizan para calcular los recuentos de lecturas, con el fin de proporcionar una estimación de su nivel de abundancia (por ejemplo, expresión). En algunas realizaciones, dicho recuento considera el número total de lecturas que se solapan con los exones de un gen. Sin embargo, debido a que, en algunos casos, algunas de las lecturas de la secuencia se sitúan fuera de los límites de los exones conocidos, otras realizaciones alternativas consideran toda la longitud de un gen, contando también las lecturas de los intrones. Además, en algunas realizaciones, se utilizan lecturas de corte y empalme para modelizar la abundancia de diferentes isoformas de corte y empalme de un gen. Véase, por ejemplo, Trapnell *et al.*, 2010, *Nat. Biotechnol.*, 28, 511-515; y Gatto *et al.*, 2014, *Nucleic Acids Res.*, 42, pág. e71.

Como se ha explicado anteriormente, la cuantificación de la abundancia de transcritos a partir de datos de RNA-Seq se suele aplicar en la canalización del análisis a través de dos etapas informáticas: alineación de las lecturas con un genoma o transcriptoma de referencia, y estimación posterior de las abundancias de transcritos e isoformas basada en las lecturas alineadas. Lamentablemente, las lecturas generadas por las tecnologías RNA-Seq más utilizadas suelen ser mucho más cortas que los transcritos de los que se extraen como muestras. Por consiguiente, en presencia de transcritos con secuencias similares, no siempre es posible asignar de forma unívoca las lecturas de secuencia cortas a un gen específico. Estas lecturas de secuencia se denominan "multilecturas" porque son homólogas con más de una región del genoma de referencia u homólogas con múltiples transcritos del transcriptoma de referencia. En algunas realizaciones, tales multilecturas se descartan, es decir, no contribuyen a los recuentos de abundancia de genes. En algunas realizaciones, se utilizan programas como MMSEQ o RSEM para resolver la ambigüedad. Véanse ejemplos de metodologías utilizadas para resolver las multilecturas en Turro *et al.*, 2011, *Genome Biol.*, 12, pág. R13; y Nicolae *et al.*, *Algorithms Mol. Biol.*, 6, 9.

Otro aspecto de RNA-Seq es la normalización de los recuentos de lecturas de secuencia. En algunas realizaciones, esto incluye la normalización para tener en cuenta las diferentes profundidades de secuenciación. Véase, por ejemplo, Lin *et al.*, 2011, *Bioinformatics*, 27, 2031-2037; Robinson Oshlack, 2010, *Genome Biol.*, 11, R25; y Li *et al.*, 2012, *Biostatistics*, 13, 523-538.

En algunas realizaciones, los recuentos de lecturas de secuencia se normalizan para tener en cuenta el sesgo de longitud de gen. Véase, Finotell y Camillo, 2014, *Briefings in Functional Genomics*, 14(2), 130-142.

En realizaciones en las que una o más pluralidades correspondientes de lecturas de secuencia se generan a partir de una secuenciación de panel selectiva utilizando una pluralidad de sondas, cada sonda respectiva en la pluralidad de sondas representa de manera distintiva una porción diferente de un genoma de referencia. En tales realizaciones, cada lectura de secuencia en la correspondiente pluralidad de lecturas de secuencia corresponde al menos a una sonda de la pluralidad de sondas.

Cada sonda respectiva en la pluralidad de sondas se dirige de forma exclusiva a una porción diferente (por ejemplo, una porción respectiva) de un transcriptoma de referencia (por ejemplo, del transcriptoma de referencia humano). Cada lectura de secuencia en la segunda pluralidad de lecturas de secuencia y cada lectura de secuencia en la tercera pluralidad de lecturas de secuencia corresponde al menos a una sonda de la pluralidad de sondas. En algunas realizaciones, se utiliza la secuenciación del genoma completo, por ejemplo, en lugar de la secuenciación de panel selectiva.

En algunas realizaciones, la segunda pluralidad de lecturas de secuencia tiene una profundidad promedio de al menos 50X a través de la pluralidad de sondas. En algunas realizaciones, la segunda pluralidad de lecturas de secuencia tiene una profundidad promedio de al menos 400X a través de la pluralidad de sondas. En otras realizaciones, la segunda pluralidad de lecturas de secuencia tiene una profundidad promedio de al menos 10X, 15X, 20X, 25X, 30X, 40X, 50X, 75X, 100X, 150X, 200X, 250X, 300X, 400X, 500X o mayor.

En algunas realizaciones, la pluralidad de sondas incluye sondas para al menos trescientos genes diferentes. En algunas realizaciones, la pluralidad de sondas incluye sondas para al menos quinientos genes diferentes. En otras realizaciones, la pluralidad de sondas incluye al menos 50, 100, 150, 200, 250, 300, 400, 500, 600, 700, 800, 900, 1000, 1500, 2000, 3000, 4000, 5000 o más genes diferentes.

Bloque 208. En referencia al bloque 208 de la figura 2A, se determina un lote de referencia emparejado por cohorte para el conjunto de datos de lotes. El lote de referencia emparejado por cohorte está equilibrado en cuanto al emplazamiento del tejido, la pureza del tumor, el tipo de cáncer, la identidad del secuenciador o los datos secuenciados. En algunas realizaciones, un tamaño del lote de referencia emparejado por cohorte (por ejemplo, el número de muestras en el mismo) es igual al tamaño del conjunto de datos de lotes. En algunas realizaciones, el lote de referencia emparejado por cohorte tiene un tamaño diferente al del conjunto de datos de lotes.

Determinación de un conjunto de datos de referencia emparejado por cohorte 130 a partir del conjunto de datos de referencia 140

Haciendo referencia al bloque 210 de la figura 2A, en algunas realizaciones, la determinación de un conjunto de datos de referencia emparejado por cohorte para el conjunto de datos de lotes comprende extraer, para cada muestra en el lote de muestras: i) una pluralidad respectiva de características de secuencia a partir de la pluralidad respectiva de lecturas de secuencia, obteniendo de este modo una pluralidad de características de secuencia del lote, y ii) una pluralidad respectiva de características de metadatos de muestra, obteniendo de este modo una pluralidad de características de metadatos del lote. En algunas realizaciones, se selecciona, a partir de un conjunto de datos de referencia, basándose al menos en parte en la pluralidad de características de secuencia del lote o en la pluralidad de características de metadatos del lote, el conjunto de datos de referencia emparejado por cohorte que comprende una pluralidad de muestras de referencia.

En algunas realizaciones, el conjunto de datos de referencia emparejados por cohorte comprende una pluralidad de muestras de referencia. En algunas realizaciones, cada muestra de referencia de la pluralidad de muestras de referencia comprende una pluralidad correspondiente de lecturas de secuencia derivadas de la muestra de referencia respectiva mediante secuenciación de ARN selectiva o del transcriptoma completo y los metadatos correspondientes para la muestra de referencia respectiva. En algunas realizaciones, las muestras de referencia para un conjunto de datos de referencia emparejado por cohorte se seleccionan basándose, al menos en parte, en las características de la muestra a partir de metadatos de la muestra, tales como los perfiles de transcripción del ARN (por ejemplo, determinados a partir de una pluralidad de lecturas de secuencia), los datos clínicos (por ejemplo, diagnóstico del paciente, resultados del tratamiento, etc.), el sexo, el tipo de biopsia (por ejemplo, biopsia de sangre frente a biopsia sólida) y/o los datos moleculares (por ejemplo, mutaciones genómicas).

En algunas realizaciones, el diagnóstico del paciente comprende el tipo de cáncer y/o el estadio del cáncer. En algunas realizaciones, un tipo de cáncer respectivo para cada muestra en el conjunto de datos de lotes se selecciona del conjunto que consiste en un tumor de un estadio predeterminado de un cáncer cerebral, un estadio predeterminado de un glioblastoma, un estadio predeterminado de un cáncer de próstata, un estadio predeterminado de un cáncer de páncreas, un estadio predeterminado de un cáncer de riñón, un estadio predeterminado de un cáncer colorrectal, un estadio predeterminado de un cáncer de ovario, un estadio predeterminado de un cáncer de endometrio, o un estadio predeterminado de un cáncer de mama.

En algunas realizaciones, el tipo de biopsia comprende una biopsia somática. En algunas realizaciones, una biopsia somática comprende secciones de tejido macrodisecadas fijadas en formol e incluidas en parafina ("formalin fixed paraffin embedded", FFPE), biopsia quirúrgica, biopsia cutánea, biopsia en sacabocados, biopsia prostática, biopsia ósea, biopsia de médula ósea, biopsia con aguja, biopsia guiada por TC, biopsia guiada por ecografía, aspiración con aguja fina, biopsia por aspiración, muestras de tejido fresco o de sangre. En algunas realizaciones, la biopsia somática es de un tumor de mama, un glioblastoma, un tumor de próstata, un tumor pancreático, un tumor renal, un tumor colorrectal, un tumor ovárico, un tumor endometrial, un tumor de mama o una combinación de los mismos. Las biopsias suelen realizarse después de que una o más pruebas clínicas menos invasivas sugieran que un paciente tiene o es probable que tenga uno o más tumores. El tipo de biopsia suele depender de la ubicación del tumor. Por ejemplo, las biopsias de tumores renales suelen realizarse por vía endoscópica, mientras que las biopsias de tumores ováricos suelen consistir en raspados de tejido.

En algunas realizaciones, las mutaciones genómicas comprenden variantes de número de copias, mutaciones somáticas, mutaciones de la línea germinal, indicaciones de inestabilidad de microsatélites, carga mutacional tumoral, indicaciones de carga de patógenos o celularidad tumoral.

Algunos ejemplos de variaciones en el número de copias se describen en Shilien y Malkin, 2009, *Genome Med.*, 1, 62. Las indicaciones de inestabilidad de microsatélites pueden determinarse como describen Buhard *et al.*, 2006, *J. Clinical Onco.*, 24(2), 241. Algunos ejemplos de la determinación de la carga mutacional tumoral se describen en Chalmers *et al.*, 2017, *Genome Med.*, 9, 34. Las indicaciones de carga de patógenos y/o la indicación de infiltración inmunitaria pueden determinarse como se describe, por ejemplo, en Barber *et al.*, 2015, *PLoS Pathog.*, 11(1): e1004558; y Pages *et al.*, 2010, *Oncogene*, 29, 1093-1102. En algunos casos, se determina una indicación de celularidad tumoral a partir de una biopsia somática comparando un número de células cancerosas con el número de células normales obtenidas en la biopsia somática. En algunos casos, se determina una indicación de la celularidad tumoral a partir de una o más imágenes de una biopsia somática (por ejemplo, contando e identificando las células cancerosas frente a las no cancerosas).

Los conjuntos de datos de referencia emparejados por cohorte están, en algunas realizaciones, equilibrados para corresponder lo más estrechamente posible a los tipos de muestras presentes en el conjunto de datos de lotes (por ejemplo, seleccionando muestras de referencia que son similares a las muestras respectivas en la pluralidad de muestras en el conjunto de datos de lotes). En algunas realizaciones, la similitud entre las muestras de referencia y las muestras del conjunto de datos de lotes se determina basándose en al menos una de las características de la muestra descritas anteriormente. En algunas realizaciones, se selecciona un lote de referencia emparejado por cohorte de la base de datos de referencia para incluir tantas muestras de referencia como sea posible (por ejemplo, para mantener un lote de referencia que esté equilibrado en cuanto al emplazamiento del tejido, la pureza del tumor, el tipo

de cáncer, la identidad del secuenciador, la fecha de secuenciación y/o las características de la muestra obtenidas de los metadatos de la muestra).

En algunas realizaciones, cada muestra en un primer subconjunto de muestras en el conjunto de datos de lotes tiene un primer tipo de biopsia correspondiente, y cada muestra en un segundo subconjunto de muestras en el conjunto de

5 datos de lotes tiene un segundo tipo de biopsia correspondiente. En algunas realizaciones, el primer tipo de biopsia o el segundo tipo de biopsia comprende una biopsia somática seleccionada del conjunto que comprende secciones de tejido macrodisecadas fijadas en formol e incluidas en parafina (FFPE), biopsia quirúrgica, biopsia cutánea, biopsia en sacabocados, biopsia de próstata, biopsia ósea, biopsia de médula ósea, biopsia con aguja, biopsia guiada por TC, biopsia guiada por ecografía, aspiración con aguja fina, biopsia por aspiración, muestras de tejido fresco o de sangre.

10 En algunas realizaciones, el primer y segundo tipo de biopsia se identifican en metadatos respectivos para cada muestra en el conjunto de datos de lotes. Para proporcionar un conjunto de datos de referencia emparejado por cohorte equilibrado, cada muestra de referencia de un primer subconjunto de muestras de referencia del conjunto de datos de referencia emparejado por cohorte tiene el primer tipo de biopsia correspondiente, y cada muestra de referencia de un

15 segundo subconjunto de muestras de referencia del conjunto de datos de referencia emparejado por cohorte tiene el segundo tipo de biopsia correspondiente. Por ejemplo, si la pluralidad de muestras del conjunto de datos de lotes comprende un 50 % de muestras con cáncer de mama, un 20 % de muestras con cáncer de pulmón y un 30 % de muestras con cáncer cerebral, entonces el conjunto de datos de referencia emparejado por cohorte incorporará el mayor número de muestras de referencia del conjunto de datos de referencia que sean compatibles con estos porcentajes de tipos de cáncer. En algunas realizaciones, se emplean procedimientos similares para equilibrar un

20 conjunto de datos de referencia emparejado por cohorte con un conjunto de datos de lotes con respecto a otras características de la muestra.

Haciendo referencia al bloque 212 de la figura 2A, en algunas realizaciones, una combinación lineal o no lineal de la pluralidad de características de secuencia del lote y la pluralidad de características de metadatos del lote se determina sometiéndolo a una técnica de reducción de dimensiones. En algunas realizaciones, la técnica de reducción de dimensiones comprende la aproximación y proyección uniforme de la variedad ("Uniform Manifold Approximation and Projection", UMAP). En algunas realizaciones, la técnica de reducción de dimensiones comprende el análisis de componentes principales ("Principal Component Analysis", PCA).

25

Realización de pruebas de control de calidad global en el conjunto de datos de lotes 122

30 **Bloque 214.** En referencia al bloque 214 de la figura 2B, se realizan una o más pruebas de control de calidad global (por ejemplo, pruebas 152) en el conjunto de datos de lotes utilizando al menos el conjunto de datos de referencia emparejado por cohorte (por ejemplo, lote). En algunas realizaciones, dichas una o más pruebas de control de calidad global comprenden pruebas para uno o más efectos de lote de un conjunto que comprende procedimientos de análisis de la canalización de datos bioinformáticos y de secuenciación.

35 Haciendo referencia al bloque 216 de la figura 2B, en algunas realizaciones, dichas una o más pruebas de control de calidad global de lotes comprenden pruebas para uno o más efectos de lote de un conjunto que comprende análisis de la canalización de datos bioinformáticos (por ejemplo, fecha de análisis de la muestra, identidad del secuenciador, tipo de canalización, etc.), contaminación de ADN, manipulación de la muestra (por ejemplo, procedimiento de recogida de la muestra, cambios de reactivo, etc.) y procedimientos de secuenciación (por ejemplo, adaptadores de secuencia UMI frente a UDI).

40

En algunas realizaciones, se utilizan diferentes canalizaciones de datos bioinformáticos para analizar muestras (por ejemplo, basados en el tipo de biopsia, muestras de ácido nucleico libre de células), y el uso de diferentes canalizaciones puede contribuir a los efectos de lote. Por ejemplo, en algunas realizaciones, incluso el tipo de tubo de ensayo utilizado para la recogida de la muestra de sangre (por ejemplo, PAX frente a EDTA) se tiene en cuenta por su

45 posible impacto en los efectos de lote. En algunas realizaciones, los cambios en el instrumental (por ejemplo, máquinas de secuenciación o celdas de flujo) también pueden contribuir a los efectos de lote. En algunas realizaciones, los reactivos con posibles efectos de lote incluyen lotes de sondas, controles (por ejemplo, controles Horizon) y tampones.

Haciendo referencia al bloque 218 de la figura 2B, en algunas realizaciones, el lote de referencia emparejado por cohorte se usa para ajustar cada muestra en el conjunto de datos de lotes a una o más covariables de confusión antes de realizar dichas una o más pruebas de control de calidad global de lotes. En algunas realizaciones, este ajuste comprende la normalización de los niveles de expresión para cada gen en un genoma de referencia (por ejemplo, un genoma de referencia que comparte cada muestra en el conjunto de datos de lotes y cada muestra de referencia en el lote de referencia emparejado por cohorte) para cada pluralidad respectiva a las lecturas de secuencia para cada muestra en el conjunto de datos de lotes. En algunas realizaciones, Mostafavi, 2013, incluye un resumen de procedimientos de normalización pertinentes. Véase PLOS ONE, e68141 en la sección titulada "Unified Representation of Existing Normalization Methods".

50

55

En algunas realizaciones, al menos una muestra en el conjunto de datos de lotes es una muestra de control (por ejemplo, una muestra de control Horizon). En algunas realizaciones, dicha al menos una muestra de control del

conjunto de datos de lotes se utiliza para ajustar cada una de las otras muestras del conjunto de datos de lotes. Las muestras de control Horizon son controles disponibles en el mercado que se derivan de líneas celulares que comprenden una variante de fusión conocida. En algunas realizaciones, se espera que la expresión de la variante de fusión sea constante, independientemente de las condiciones experimentales (por ejemplo, identidad del secuenciador, procedimiento de secuenciación, fecha de secuenciación, etc.). Estos controles Horizon son útiles para normalizar muestras entre y a través de conjuntos de datos de lotes y también para proporcionar información sobre tendencias de secuenciación a lo largo del tiempo (por ejemplo, comparando controles Horizon evaluados en diferentes puntos temporales entre sí). En algunas realizaciones, cualquier muestra de control disponible en el mercado puede utilizarse con los procedimientos descritos en el presente documento.

En algunas realizaciones, una prueba de control de calidad global de lotes respectiva comprende i) determinar un número promedio de lecturas de secuencia por muestra a través del conjunto de datos de lotes, ii) obtener un número promedio de referencia de lecturas de secuencia por muestra a partir de un conjunto de datos de referencia (por ejemplo, o a partir del lote de referencia emparejado por cohorte), y iii) comparar el número promedio de lecturas de secuencia a través del conjunto de datos de lotes con el número promedio de referencia de lecturas de secuencia por muestra. En algunas realizaciones, cuando el número promedio de lecturas de secuencia es inferior al número promedio de lecturas de secuencia de referencia por muestra, el conjunto de datos de lotes no supera la prueba de control de calidad global de lotes correspondiente.

En algunas realizaciones, una prueba de control de calidad global de lotes respectiva comprende i) determinar un porcentaje promedio de lecturas de secuencia cartografiadas por muestra a través del conjunto de datos de lotes, ii) obtener un porcentaje promedio de referencia de lecturas de secuencia cartografiadas por muestra a partir de un conjunto de datos de referencia (por ejemplo, o a partir del lote de referencia emparejado por cohorte), y iii) comparar el porcentaje promedio de lecturas de secuencia cartografiadas a través del conjunto de datos de lotes con el porcentaje promedio de referencia de lecturas de secuencia cartografiadas por muestra. En algunas realizaciones, cuando el porcentaje promedio de lecturas de secuencia cartografiadas es inferior al porcentaje promedio de referencia de lecturas de secuencia cartografiadas por muestra, el conjunto de datos de lotes no supera la respectiva prueba de control de calidad global de lotes.

En algunas realizaciones, los metadatos respectivos para cada muestra en un conjunto de datos de lotes incluyen un tipo de cáncer respectivo. En algunas de estas realizaciones, una prueba de control de calidad global de lotes respectiva comprende aplicar, para cada muestra respectiva del conjunto de datos de lotes, la pluralidad correspondiente de lecturas de secuencia y los metadatos correspondientes a un segundo modelo de clasificación entrenado, proporcionando así el segundo modelo de clasificación entrenado un tipo de cáncer predicho respectivo para cada muestra. En algunas realizaciones, la respectiva prueba de control de calidad global de lotes comprende además comparar, para cada muestra, el respectivo tipo de cáncer de los respectivos metadatos con el respectivo tipo de cáncer predicho. En algunas realizaciones, una o más muestras con un respectivo tipo de cáncer predicho que no coinciden con el respectivo tipo de cáncer conocido no superan la prueba de control de calidad global de lotes. En algunas realizaciones, todo el conjunto de datos de lotes no supera la respectiva prueba de control de calidad global de lotes cuando una o más muestras del conjunto de datos de lotes no superan la prueba de control de calidad global de lotes. En algunas realizaciones, el segundo procedimiento de clasificación entrenado comprende cualquiera de los procedimientos de clasificación descritos en la solicitud de patente provisional de EE. UU. n.º 62/855 750 titulada "Systems and Methods for Multi-Label Cancer Classification" y presentada el 31 de mayo de 2019.

En algunas realizaciones en las que los metadatos respectivos para cada muestra en el conjunto de datos de lotes incluyen un tipo de cáncer respectivo, una prueba de control de calidad global de lotes respectiva comprende determinar un porcentaje de pureza tumoral respectivo para cada muestra respectiva en el conjunto de datos de lotes. En algunas realizaciones, la pureza del tumor se determina basándose al menos en parte en fracciones de alelos variantes, y, en algunas realizaciones, las fracciones de alelos variantes se determinan como se describe en Shin *et al.*, 2017, "Prevalence and detection of low-allele-fraction variants in clinical cancer samples", Nat. Comm., 8, 1377. En algunas realizaciones, cuando una muestra respectiva tiene una pureza tumoral correspondiente de menos del 20 %, menos del 30 %, menos del 40 % o menos del 50 %, entonces la muestra respectiva no supera la prueba de control de calidad global de lotes respectiva. En algunas realizaciones, cuando al menos el 30 %, al menos el 40 %, al menos el 50 % o al menos el 60 % de la pluralidad de muestras del conjunto de datos de lotes no han superado la respectiva prueba de control de calidad global de lotes, entonces el conjunto de datos de lotes no supera la prueba de control de calidad global de lotes.

Bloque 220. En referencia al bloque 220 de la figura 2B, las muestras respectivas del conjunto de datos de lotes que no superan alguna de dichas una o más pruebas de control de calidad global 152 se eliminan del conjunto de datos de lotes o se marcan para su inspección manual. En algunas realizaciones, la etapa de eliminación comprende además proporcionar un conjunto de datos de lotes actualizado que carece de cada una de las muestras respectivas que no superaron alguna de dichas una o más pruebas de control de calidad global.

Informes de muestra

Haciendo referencia al bloque 222, en algunas realizaciones, para cada muestra en el lote de muestras (en algunas realizaciones, incluso aquellas muestras que no superan una prueba de control de calidad global de lotes), se

proporciona un informe de muestra respectivo. En realizaciones alternativas, los informes de muestra se proporcionan sólo para aquellas muestras que superaron todas las pruebas de control de calidad global de lotes (por ejemplo, muestras incluidas en el conjunto de datos de lotes actualizado). En algunas realizaciones, cada informe de muestra respectivo comprende al menos una de un conjunto de llamadas de expresión, una o más terapias emparejadas, o uno o más ensayos clínicos emparejados. En algunas realizaciones, las terapias emparejadas adecuadas se determinan basándose en las llamadas de expresión y en la información sobre el tipo de cáncer. En algunas realizaciones, las terapias emparejadas adecuadas se determinan basándose en pruebas de organoides. Algunos ejemplos de pruebas de organoides y la correlación entre los resultados de las pruebas de organoides y la sensibilidad de la terapia se proporcionan en la solicitud de patente provisional de EE. UU. n.º 62/924 621 titulada "Systems and Methods for Predicting Therapeutic Sensitivity" y presentada el 22 de octubre de 2019. En algunas realizaciones, los ensayos clínicos emparejados apropiados se determinan basándose, al menos en parte, en las llamadas de expresión correspondientes para una muestra respectiva.

En algunas realizaciones, un informe de muestra puede incluir además resúmenes que proporcionan a los pacientes y/o proveedores de atención sanitaria una breve visión general de los hallazgos más importantes del informe de muestra completo. En algunas realizaciones, se proporciona un informe de muestra (por ejemplo, de paciente) como se describe en la solicitud de patente provisional de EE. UU. n.º 62/855 750 titulada "Systems and Methods for Multi-Label Cancer Classification" y presentada el 31 de mayo de 2019.

En algunas realizaciones, cada muestra del lote de muestras, una vez realizado el análisis de control de calidad, se asocia además con los datos clínicos correspondientes. En algunas realizaciones, la asociación entre las muestras de secuencias de ARN y los datos clínicos se utiliza para validar o refinar las llamadas de expresión. En algunas realizaciones, los datos clínicos comprenden mutaciones del ADN, la respuesta del paciente a la terapia, los resultados de experimentos con organoides (por ejemplo, los organoides derivados de un paciente pueden ensayarse para determinar si los organoides son sensibles a terapias emparejadas) y/o imágenes histopatológicas. Algunos ejemplos de imágenes histopatológicas son las teñidas con H&E (hematoxilina y eosina) e IHC (inmunohistoquímica).

Realización de pruebas de control de calidad de una sola muestra

Bloque 230. En referencia al bloque 230 de la figura 2B, en algunas realizaciones, para cada muestra respectiva del lote de muestras, a partir de la pluralidad correspondiente de lecturas de secuencia, se realizan una o más pruebas de control de calidad de una sola muestra en la muestra respectiva. Las muestras respectivas del lote de muestras que no superan alguna de las pruebas de control de calidad de una o varias muestras individuales se eliminan del conjunto de datos de lotes o se marcan para su inspección manual. En algunas realizaciones, cualquier prueba de control de calidad de una sola muestra puede aplicarse a un conjunto de datos de lotes completo. En algunas realizaciones, las pruebas de control de calidad de una sola muestra se llevan a cabo antes de las pruebas de control de calidad de lotes. En algunas realizaciones, las pruebas de control de calidad de una sola muestra se llevan a cabo después de las pruebas de control de calidad de lotes.

En algunas realizaciones, una prueba de control de calidad de una sola muestra respectiva compara el número total de lecturas de secuencia por muestra en el lote de muestras con el número promedio de lecturas de secuencia por muestra de referencia en todo el conjunto de datos de referencia 140. En otras palabras, en algunas realizaciones, una prueba de control de calidad de una sola muestra respectiva comprende i) determinar un número promedio de lecturas de secuencia por muestra en el conjunto de datos de lotes, y ii) obtener un número promedio de lecturas de secuencia de referencia por muestra a partir de un conjunto de datos de referencia (por ejemplo, o del lote de referencia emparejado por cohorte, o de un subconjunto del conjunto de datos de referencia). El respectivo conjunto de datos de una sola muestra compara el número promedio de lecturas de secuencia a través del conjunto de datos de lotes con el número promedio de lecturas de secuencia de referencia por muestra. En algunas realizaciones, cuando el número total respectivo de lecturas de secuencia para una muestra respectiva es inferior al número promedio de referencia de lecturas de secuencia por muestra, la muestra respectiva no supera la prueba de control de calidad respectiva.

En algunas realizaciones, se utiliza una prueba de Wilcoxon para evaluar si el conjunto de datos de lotes no supera la respectiva prueba de control de calidad de una sola muestra. En algunas realizaciones, una prueba de Wilcoxon de dos muestras compara grupos emparejados (por ejemplo, compara las muestras más similares (emparejadas) entre dos grupos). Por ejemplo, esto resulta útil para comparar directamente muestras procedentes de un mismo conjunto de sujetos secuenciadas en sistemas HiSeq1 y HiSeq2. En algunas realizaciones, se utiliza una prueba de Wilcoxon para datos independientes (por ejemplo, en los casos en los que hay alguna diferencia en los lotes que se van a comparar). En algunas realizaciones, se utiliza un umbral de valor de p modificado para determinar si existe una diferencia significativa (por ejemplo, un efecto de lote). En algunas realizaciones, se determina un umbral de valor de p modificado basado en al menos una pluralidad de muestras de referencia de alta calidad del conjunto de datos de referencia 140. En algunas realizaciones, se determina que una muestra de referencia es de alta calidad cuando el número de lecturas correspondiente de la muestra de referencia es de al menos 5 millones de lecturas de secuencia, al menos 10 millones de lecturas de secuencia, al menos 20 millones de lecturas de secuencia, al menos 30 millones de lecturas de secuencia, al menos 40 millones de lecturas de secuencia, al menos 50 millones de lecturas de secuencia, al menos 100 millones de lecturas de secuencia o al menos 200 millones de lecturas de secuencia.

- En algunas realizaciones, una prueba de control de calidad de una sola muestra respectiva comprende aplicar, para cada muestra respectiva en el conjunto de datos de lotes, la pluralidad correspondiente de lecturas de secuencia y los metadatos de muestra correspondientes a un primer modelo de clasificación entrenado, proporcionando así el primer modelo de clasificación entrenado un conjunto de asignaciones de género predichas que comprende una asignación de género predicha respectiva para cada muestra. En algunas realizaciones, la respectiva prueba de control de calidad de una sola muestra comprende además comparar el conjunto de asignaciones de género predichas con un conjunto esperado de asignaciones de género (por ejemplo, para detectar intercambios no intencionados de muestras). En algunas realizaciones, una o más muestras con una respectiva asignación de género predicha que no coincide con el conjunto esperado de asignaciones de género (por ejemplo, una proporción demasiado alta de un género) no supera la respectiva prueba de control de calidad de una sola muestra. En algunas realizaciones, todo el conjunto de datos de lotes no supera la respectiva prueba de control de calidad de una sola muestra cuando el conjunto de asignaciones de género predichas no coincide con el conjunto esperado de asignaciones de género (por ejemplo, cuando parece haber demasiados intercambios de muestras).
- En algunas realizaciones, a modo de ejemplo no limitativo, el primer modelo de clasificación comprende un árbol de decisión. Los algoritmos de árbol de decisión adecuados para su uso como el clasificador del bloque 244 se describen, por ejemplo, en Duda, 2001, *Pattern Classification*, John Wiley & Sons, Inc., Nueva York, págs. 395-396. 2. En algunas realizaciones, el árbol de decisión es una regresión de bosque aleatorio. Un algoritmo específico que puede utilizarse como clasificador del bloque 244 es un árbol de clasificación y regresión ("classification and regression tree", CART). Otros ejemplos de algoritmos de árboles de decisión específicos que pueden utilizarse como el clasificador del bloque 244 incluyen, entre otros, ID3, C4.5, MART y Random Forests. CART, ID3 y C4.5 se describen en Duda, 2001, *Pattern Classification*, John Wiley & Sons, Inc., Nueva York, págs. 396-408 y págs. 411-412. CART, MART y C4.5 se describen en Hastie *et al.*, 2001, *The Elements of Statistical Learning*, Springer-Verlag, Nueva York, capítulo 9, que se incorpora por referencia en su totalidad. Los bosques aleatorios se describen en Breiman, 1999, "Random Forests—Random Features", Technical Report, 567, Statistics Department, U.C. Berkeley, septiembre de 1999.
- En algunas realizaciones, una prueba de control de calidad de una sola muestra respectiva en dichas una o más pruebas de control de calidad de una sola muestra comprende i) determinar, para cada muestra respectiva en el lote de muestras, un número respectivo de lecturas de secuencia cartografiadas no duplicadas (por ejemplo, lecturas que no son el resultado de duplicación por PCR) en la pluralidad de lecturas de secuencia, y ii) comparar el número respectivo de lecturas de secuencia cartografiadas no duplicadas con un número esperado de lecturas de secuencia cartografiadas no duplicadas. Cada lectura de secuencia cartografiada no duplicada se cartografía a una porción correspondiente de un genoma de referencia (por ejemplo, tiene sitios de inicio y fin distintivos en un genoma de referencia). En algunas realizaciones, cuando el número respectivo de lecturas de secuencia cartografiadas no duplicadas es inferior a un número predeterminado de lecturas cartografiadas no duplicadas, la muestra respectiva no supera la prueba de control de calidad de una sola muestra respectiva.
- En algunas realizaciones, el número esperado de lecturas de secuencia cartografiadas no duplicadas se basa en si la muestra se obtuvo de una biopsia sólida o líquida del sujeto (por ejemplo, de un tumor sólido o de una muestra de sangre). En algunas realizaciones, se determina un número respectivo de lecturas duplicadas para cada muestra en el conjunto de datos de lotes (por ejemplo, mediante la identificación de lecturas de secuencia que tienen sitios de inicio y fin idénticos). En algunas realizaciones, el procedimiento proporciona además una representación gráfica del número respectivo de lecturas duplicadas como parte de un informe de la muestra o de un informe global del lote.
- En algunas realizaciones, una prueba de control de calidad de una sola muestra respectiva comprende determinar una puntuación de calidad respectiva para cada posición de lectura de par de bases en la pluralidad correspondiente de lecturas de secuencia de cada muestra respectiva en el conjunto de datos de lotes. En algunas realizaciones, cuando una o más puntuaciones de calidad respectivas de una o más posiciones de lectura de pares de bases respectivas son inferiores a una puntuación de calidad umbral, la muestra respectiva no supera la prueba de control de calidad de una sola muestra respectiva. En algunas realizaciones, la puntuación de calidad umbral comprende 20,0 (por ejemplo, según lo calculado por FastQC). En algunas realizaciones, se determina una puntuación de calidad respectiva para cada lectura de secuencia en la pluralidad de lecturas de secuencia para cada muestra respectiva en el conjunto de datos de lotes, y en algunas de tales realizaciones, se descartan una o más lecturas de secuencia con puntuaciones de calidad correspondientes que son inferiores a una puntuación de calidad de lectura umbral.
- En algunas realizaciones, una prueba de control de calidad de una sola muestra respectiva comprende determinar una puntuación de calidad promedio para cada lectura de secuencia en la pluralidad correspondiente de lecturas de secuencia de cada muestra respectiva en el conjunto de datos de lotes. En algunas realizaciones, cuando la media de las puntuaciones de calidad promedio de la pluralidad correspondiente de lecturas de secuencia es inferior a una puntuación de calidad umbral, la muestra respectiva no supera la prueba de control de calidad de una sola muestra correspondiente. En algunas realizaciones, la puntuación de calidad umbral comprende 20,0 (por ejemplo, según lo calculado por FastQC).
- En algunas realizaciones, una prueba de control de calidad de una sola muestra respectiva comprende, para cada muestra respectiva en el lote de muestras, determinar un porcentaje respectivo de lecturas de secuencia correctamente emparejadas (por ejemplo, el porcentaje de lecturas de secuencia que son lecturas de extremos emparejados y están correctamente emparejadas), en la que, cuando el porcentaje de lecturas de secuencia

correctamente emparejadas es inferior a un umbral de lectura emparejada predeterminado, la muestra respectiva no supera la prueba de control de calidad de una sola muestra respectiva. En algunas realizaciones, el umbral de lectura emparejado predeterminado comprende al menos el 90 %, al menos el 95 % o al menos el 99 %.

5 En algunas realizaciones, una prueba de control de calidad de una sola muestra respectiva comprende determinar, para cada muestra respectiva en el lote de muestras, un número respectivo de genes expresados (por ejemplo, el número de genes que tienen lecturas de secuencia de apoyo distintas de cero). En algunas realizaciones, cuando la puntuación de lecturas expresadas correspondiente es inferior a un número predeterminado de lecturas expresadas, la muestra respectiva no supera la prueba de control de calidad de una sola muestra respectiva. En algunas realizaciones, cuando la muestra respectiva procede de una biopsia sólida, el número predeterminado de genes expresados es de al menos 18 000, al menos 19 000 o al menos 20 000 genes. En algunas realizaciones, cuando la muestra respectiva procede de una biopsia líquida (por ejemplo, hematológica), el número predeterminado de genes expresados es de al menos 15 000, al menos 16 500 o al menos 17 000 genes. Algunos tipos de cáncer comprenden diferentes conjuntos de genes expresados (por ejemplo, algunos tipos de cáncer se diferencian claramente desde el punto de vista transcripcional). Véase, por ejemplo, Li *et al.*, 2017, "Transcriptional landscape of human cancers", *Oncotarget*, 8(21), 34534-34551. En algunas realizaciones, el número predeterminado de genes expresados se determina, al menos en parte, basándose en el tipo de cáncer (por ejemplo, a partir de los metadatos correspondientes) de la muestra respectiva.

20 En algunas realizaciones, una prueba de control de calidad de una sola muestra respectiva comprende determinar, para una muestra respectiva en el lote de muestras, un contenido de GC respectivo de la pluralidad correspondiente de lecturas de secuencia. En algunas realizaciones, cuando el contenido de GC respectivo está fuera de un umbral de contenido de GC predeterminado, la muestra respectiva no supera la prueba de control de calidad de una sola muestra respectiva. En algunas realizaciones, el umbral predeterminado de contenido de GC comprende del 35-60 %, del 40-60 %, del 45-60 %, del 50-60 % o del 55-60 %. El contenido de GC varía mucho entre los genes del genoma humano. Véase, por ejemplo, Versteeg *et al.*, 2003, "The Human Transcriptome Map Reveals Extremes in Gene Density, Intron Length, GC content, and Repeat Patterns for Domains of Highly and Weakly Expressed Genes", *Genome Res.*, 13(9), 1998-2004. El contenido de GC puede afectar a la amplificación de una molécula de ácido nucleico durante la PCR. Véase, por ejemplo, Mammedov *et al.*, 2009, "A Fundamental Study of the PCR Amplification of GC-Rich DNA Templates", *Comput. Biol. Chem.*, 32(6), 452-457.

30 En algunas realizaciones, una prueba de control de calidad de una sola muestra respectiva comprende determinar, para cada muestra respectiva en el lote de muestras, un análisis de contenido de secuencia por base respectivo a través de la pluralidad correspondiente de lecturas de secuencia de la muestra respectiva. En algunas realizaciones, una muestra respectiva no supera la prueba de control de calidad de una sola muestra respectiva cuando una distribución del contenido de A, T, C o G se desvía en más de un porcentaje umbral a través de las posiciones de base representadas colectivamente por la pluralidad correspondiente de lecturas de secuencia para la muestra respectiva. En algunas realizaciones, el porcentaje umbral comprende al menos el 5 %, al menos el 6 %, al menos el 7 %, al menos el 8 %, al menos el 9 %, o al menos el 10 % de deriva según se determina a partir de FastQC.

40 En algunas realizaciones, una prueba de control de calidad de una sola muestra respectiva comprende determinar, para cada muestra respectiva en el lote de muestras, un análisis de contenido de GC por base respectivo a través de la pluralidad correspondiente de lecturas de secuencia. En algunas realizaciones, una muestra respectiva no supera la prueba de control de calidad de una sola muestra respectiva cuando una distribución del contenido de GC se desvía en más de un porcentaje umbral a través de las posiciones de base representadas colectivamente por la pluralidad correspondiente de lecturas de secuencia. En algunas realizaciones, el porcentaje umbral comprende más del 5 %, más del 6 %, más del 7 %, más del 8 %, más del 9 % o más del 10 % de deriva.

45 En algunas realizaciones, una prueba de control de calidad de una sola muestra respectiva comprende determinar, para cada muestra respectiva en el lote de muestras, una distribución correspondiente de contenido de GC por lectura de secuencia a través de la pluralidad correspondiente de lecturas de secuencia. En algunas realizaciones, una muestra respectiva no supera la prueba de control de calidad de una sola muestra respectiva cuando una prueba de bondad de ajuste determina que la distribución respectiva del contenido de GC por lectura de secuencia se desvía de una distribución normal con un nivel de significación umbral (por ejemplo, mediante un análisis con una prueba de bondad de ajuste de chi-cuadrado al nivel de significación de 0,05).

50 En algunas realizaciones, una prueba de control de calidad de una sola muestra respectiva comprende determinar, para cada muestra respectiva en el lote de muestras, un análisis de contenido ausente por base correspondiente a través de la pluralidad correspondiente de lecturas de secuencia que determina un porcentaje de llamadas ausentes para cada posición de base representada por la pluralidad correspondiente de lecturas de secuencia. En algunas realizaciones, una muestra respectiva no supera la prueba de control de calidad de una sola muestra respectiva cuando el porcentaje correspondiente de llamadas ausentes para una posición de base representada por la pluralidad correspondiente de lecturas de secuencia supera un porcentaje umbral. En algunas realizaciones, el porcentaje umbral comprende más del 10 %, más del 15 %, más del 20 % o más del 25 %.

60 En algunas realizaciones, una prueba de control de calidad de una sola muestra respectiva comprende un análisis de la distribución de longitud de lectura de secuencia que determina, para cada muestra respectiva en el conjunto de

datos de lotes, un intervalo respectivo de longitudes de lectura de secuencia a través de la pluralidad correspondiente de lecturas de secuencia. En algunas realizaciones, una muestra respectiva no supera la prueba de control de calidad de una sola muestra respectiva cuando el intervalo respectivo de longitudes de lectura de secuencia se desvía de una expectativa de lectura de secuencia (por ejemplo, si se observan lecturas de secuencia de longitud fija, marcar/eliminar si se observa una distribución de longitudes de lectura de secuencia, si se observan intervalos de lecturas de secuencia, marcar/eliminar si la distribución no satisface un valor de expectativa para la distribución).

En algunas realizaciones, una prueba de control de calidad de una sola muestra respectiva comprende un análisis de secuencias sobrerrepresentadas que determina, para cada muestra respectiva en el conjunto de datos de lotes, si cualquier lectura de secuencia en la pluralidad correspondiente de lecturas de secuencia está sobrerrepresentada. En algunas realizaciones, una muestra respectiva no supera la prueba de control de calidad de una sola muestra respectiva cuando el análisis de secuencias sobrerrepresentadas identifica una o varias secuencias de lectura de secuencia que están representadas en más de un porcentaje umbral de la pluralidad correspondiente de lecturas de secuencia. En algunas de estas realizaciones, el porcentaje umbral comprende al menos el 0,05 %, al menos el 0,10 %, al menos el 0,15 % o al menos el 0,2 %.

Segmentación de módulos en contenedores

Bloque 240. En referencia al bloque 240 de la figura 2B, en algunas realizaciones, dichas una o más pruebas de control de calidad de lotes (por ejemplo, detección de valores atípicos del lote completo) comprenden un primer módulo (por ejemplo, módulo 120), y dichas una o más pruebas de control de calidad de una sola muestra (por ejemplo, detección de valores atípicos de una sola muestra) comprenden un segundo módulo (por ejemplo, módulo 121). En algunas realizaciones, cada uno del primer módulo y el segundo módulo comprenden un Docker respectivo (por ejemplo, un contenedor informático que permite la realización de las pruebas de control de calidad de lotes y las pruebas de control de calidad de una sola muestra independientemente del sistema operativo 116). En algunas realizaciones, el primer módulo y el segundo módulo se ejecutan en un mismo sistema informático. En algunas realizaciones, el primer módulo y el segundo módulo se ejecutan en diferentes sistemas informáticos.

Se ofrecen algunos ejemplos de Dockers (también descritos como "contenedores" o "contenedores Docker") en Boettiger, 2015, "An introduction to Docker for reproducible research, with examples from the R environment", arXiv:1410.0846v1; y Felter *et al.*, 2014, "An updated performance comparison of virtual machines and linux containers", en IEEE International Sympo. Los contenedores Docker suelen ser útiles para facilitar el flujo de trabajo y pueden permitir el uso coordinado de varias aplicaciones. Véase, por ejemplo, Di Tommaso *et al.*, 2015, "The impact of Docker containers on the performance of genomic pipelines", Peer J, 3: e1273. En algunas realizaciones, el uso de dos o más Dockers (por ejemplo, módulos o contenedores) proporciona flexibilidad a la ejecución de los procedimientos descritos en el presente documento (por ejemplo, facilidad de aplicación sin tener en cuenta el tipo de sistema operativo disponible).

En algunas realizaciones, el primer módulo 120 (por ejemplo, el módulo de control de calidad de lotes) ensaya la calidad global del transcriptoma de un lote de muestras de ARN (por ejemplo, una pluralidad de muestras procedente de una celda de flujo completa o un conjunto de celdas de flujo). En algunas realizaciones, el primer módulo 120 evalúa la calidad global del transcriptoma de un lote de muestras de ARN frente a un conjunto equilibrado de muestras procedentes de una referencia (por ejemplo, un lote de referencia emparejado por cohorte).

En algunas realizaciones, las entradas al primer módulo 120 comprenden, para cada muestra en el lote de muestras, i) una pluralidad correspondiente de lecturas de secuencia y ii) metadatos correspondientes (por ejemplo, incluidos al menos uno o más valores bioinformáticos). En algunas realizaciones, cada pluralidad de lecturas de secuencia se normaliza (por ejemplo, como se ha descrito anteriormente con respecto al bloque 218). En algunas realizaciones, el primer módulo 120 incluye o tiene acceso a datos de referencia (por ejemplo, el conjunto de datos de referencia 140). En algunas realizaciones, el conjunto de datos de referencia 140 comprende una pluralidad de muestras de referencia 142 que incluyen, para cada muestra de referencia en la pluralidad de muestras de referencia, una pluralidad correspondiente de lecturas de secuencia 144 y metadatos de referencia 146 correspondientes.

En algunas realizaciones, una pluralidad correspondiente de lecturas de secuencia 144 comprende un archivo .csv o un archivo .parquet. En algunas realizaciones, una pluralidad correspondiente de lecturas de secuencia 144 comprende cualquier formato de archivo conocido en la técnica. En algunas realizaciones, los valores bioinformáticos incluidos en los metadatos de la muestra comprenden valores LIMS (por ejemplo, "Laboratory Information Management Systems", sistemas de gestión de información de laboratorio).

En algunas realizaciones, las pruebas de control de calidad global de lotes evaluadas por el primer módulo comprenden pruebas estadísticas y/o reducción de la dimensionalidad. En algunas realizaciones, estas pruebas estadísticas comprenden cualquier procedimiento para distinguir entre un conjunto de datos de lotes 122 y un lote de referencia correspondiente emparejado por cohorte 130. En algunas realizaciones, las pruebas estadísticas de control de calidad de lotes se llevan a cabo en un subconjunto del conjunto de datos de lotes y el lote de referencia correspondiente emparejado por cohorte (por ejemplo, sólo se comparan muestras de un tipo de cáncer concreto).

- En algunas realizaciones, la evaluación de una o más pruebas de calidad de lotes a lo largo del tiempo comprende un tercer módulo, y en algunas realizaciones, este tercer módulo comprende un tercer Docker respectivo. En algunas realizaciones, el tercer módulo evalúa las tendencias de los lotes de secuencias de ARN a lo largo del tiempo (por ejemplo, en múltiples puntos temporales). Esto es útil también para garantizar la estabilidad de los procedimientos de secuenciación (por ejemplo, evaluando si las muestras de control son similares a lo largo de los puntos temporales).
- En algunas realizaciones, el procedimiento proporciona además (por ejemplo, con posterioridad a la aplicación del módulo 120) un informe de control de calidad global de lotes. En algunas realizaciones, el informe global de lotes comprende al menos i) una lista de una o más muestras del conjunto de datos de lotes que no superan alguna de las pruebas de control de calidad de lotes, y ii) una lista de una o más muestras de referencia del conjunto de datos de referencia 140 que se evaluaron dentro de un periodo de tiempo predefinido (por ejemplo, como se identifica a partir de los metadatos de la muestra de referencia correspondiente).
- En algunas realizaciones, este periodo de tiempo predefinido comprende al menos 1 día, al menos 2 días, al menos 3 días, al menos 4 días, al menos 5 días, al menos 6 días, al menos 7 días, al menos 10 días al menos 14 días, al menos 21 días, al menos 28 días o al menos 30 días. En algunas realizaciones, el informe global de lotes es proporcionado por el tercer módulo. En algunas realizaciones, se proporciona un informe global de lotes respectivo al menos cada 1 día, al menos cada 1 semana, al menos cada 2 semanas, al menos cada 1 mes, al menos cada 3 meses o al menos cada 1 año.
- En algunas realizaciones, las métricas de calidad evaluadas por el tercer módulo comprenden una o más métricas del conjunto de al menos el contenido de GC, los niveles de contaminación (por ejemplo, en especial la contaminación de ADN que surge de la aplicación inadecuada de ADNasa de la muestra biológica), el número de lecturas, el porcentaje de lecturas cartografiadas, la tasa de duplicación de genes, el número de genes expresados en función del número de lecturas, el número de integridad del transcrito, la precisión de la determinación de tumor de origen desconocido, o la precisión de la predicción de género.
- En algunas realizaciones, el tercer módulo proporciona además representaciones gráficas de uno o más resultados de evaluación métrica en comparación con uno o más del conjunto que comprende tiempo, versión de la canalización, tipo de secuenciador, celda de flujo o tipo de cáncer. En algunas realizaciones, estas representaciones gráficas comprenden cualquier representación gráfica descrita en el presente documento o conocida en la técnica.
- En algunas realizaciones, el procedimiento proporciona además una o más representaciones gráficas para características globales del conjunto de datos de lotes. En algunas realizaciones, una representación gráfica respectiva incluye información detallada sobre la característica del conjunto de datos de lotes correspondiente. En algunas realizaciones, el procedimiento proporciona una o más representaciones gráficas de los resultados de una o más pruebas globales de control de calidad global de lotes realizadas de acuerdo con las realizaciones descritas en el presente documento. En algunas realizaciones, el procedimiento proporciona una o más representaciones gráficas de los resultados de una o más pruebas de control de calidad de una sola muestra realizadas de acuerdo con las realizaciones descritas en el presente documento. En algunas realizaciones, una característica del conjunto de datos de lotes comprende una combinación de las respectivas características de metadatos para cada muestra respectiva del lote de muestras del conjunto de datos de lotes. Por ejemplo, en algunas realizaciones, las características de metadatos de cada muestra en un conjunto de datos de lotes se combinan para proporcionar una métrica general (por ejemplo, característica) para el conjunto de datos de lotes.
- Identificación de los efectos de lote técnicos en una canalización de datos de expresión de ARN*
- Otro aspecto de la presente divulgación proporciona un procedimiento de realización del control de calidad en un sistema informático que tiene uno o más procesadores y una memoria que almacena uno o más programas para su ejecución por dichos uno o más procesadores. El procedimiento incluye la obtención, en formato electrónico, de un conjunto de datos de lotes que comprende, para cada muestra de ensayo respectiva de un lote de muestras de ensayo, un perfil de expresión correspondiente que incluye un valor de expresión génica correspondiente para cada gen respectivo de un primer conjunto de genes, y un conjunto de metadatos correspondiente que comprende valores para cada característica respectiva de un primer conjunto de características sobre una muestra.
- El procedimiento incluye determinar, para el conjunto de datos de lotes, un conjunto de datos de referencia emparejados por cohorte que incluye, para cada muestra de referencia respectiva en una pluralidad de muestras de referencia, un perfil de expresión correspondiente que comprende un valor de expresión génica correspondiente para cada gen respectivo en el primer conjunto de genes. Cada muestra de referencia respectiva de la pluralidad de muestras de referencia se asocia con un conjunto de metadatos correspondiente que comprende un valor correspondiente para cada característica respectiva de un segundo conjunto de características sobre la muestra de referencia respectiva. Los valores agregados para cada característica respectiva en un tercer conjunto de una o más características que están presentes tanto en el primer conjunto de características como en el segundo conjunto de características se equilibran entre el conjunto de datos de lotes y el conjunto de datos de referencia emparejado por cohorte.

La reducción de dimensión se realiza en un conjunto de datos combinado que consiste en el perfil de expresión correspondiente para cada muestra de ensayo respectiva en la pluralidad de muestras de ensayo y el perfil de expresión correspondiente para cada muestra de referencia respectiva en la pluralidad de muestras de referencia. Así, se obtiene para cada muestra de ensayo respectiva y cada muestra de referencia respectiva, un conjunto correspondiente de coordenadas incrustadas en un espacio dimensional inferior a las dimensiones del perfil de expresión correspondiente.

El procedimiento comprende además determinar una medida estadística de similitud entre los conjuntos de coordenadas obtenidos para las muestras de ensayo y los conjuntos de coordenadas obtenidos para las muestras de referencia. La medida estadística de similitud se compara con un valor umbral, y el conjunto de datos de lotes se valida para emitir una notificación cuando la medida estadística de similitud satisface el valor umbral, o no se valida cuando la medida estadística de similitud no satisface el valor umbral.

Por ejemplo, la figura 15 ilustra un procedimiento de realización de control de calidad (por ejemplo, en un lote de muestras en comparación con un conjunto de datos de referencia), de acuerdo con algunas realizaciones de la presente divulgación. Se obtiene un conjunto de datos de un lote de muestras de ensayo (por ejemplo, "Lote i"), que incluye una pluralidad de muestras de ARN de una pluralidad de N pacientes 1502 (por ejemplo, 1502-1, 1502-2,...,1502-N). En algunas realizaciones, en referencia al bloque 1504, el conjunto de datos de lotes se obtiene mediante un análisis de secuenciación de las muestras de ensayo (por ejemplo, análisis de exoma por RNA-Seq). El conjunto de datos de lotes incluye un perfil de expresión para cada muestra de ensayo, en el que cada perfil de expresión incluye datos de expresión para cada paciente 1506 respectivo en la pluralidad de N pacientes (por ejemplo, 1506-1, 1506-2,...,1506-N). En algunas realizaciones, los datos de expresión incluyen un valor de expresión génica correspondiente para cada gen respectivo en una pluralidad de genes en la muestra de ensayo (por ejemplo, secuenciado por RNA-Seq). En algunas realizaciones, los datos de expresión incluyen además metadatos que indican valores para una pluralidad de características asociadas con la muestra de ensayo, tales como perfiles de transcripción de ARN (por ejemplo, determinados a partir de una pluralidad de lecturas de secuencia), datos clínicos (por ejemplo, diagnóstico del paciente, resultados del tratamiento, etc.), sexo, tipo de biopsia (por ejemplo, biopsia de sangre frente a biopsia sólida), datos moleculares (por ejemplo, mutaciones genómicas) y/u otras características (por ejemplo, emplazamiento del tejido, pureza del tumor, tipo de cáncer, procedimiento de recogida, identidad del secuenciador y/o fecha de secuenciación).

En algunas realizaciones, en referencia al bloque 1508, los datos de expresión en cada perfil de expresión respectivo se normalizan. En algunas realizaciones, la normalización de los datos de expresión en cada perfil de expresión genera una pluralidad de conjuntos de datos normalizados 1510 (por ejemplo, 1510-1, 1510-2,...,1510-N).

En algunas realizaciones, el procedimiento incluye además obtener un conjunto de datos de referencia emparejados por cohorte 1512 para una pluralidad de muestras de referencia. En algunas realizaciones, el conjunto de datos de referencia emparejados por cohorte se identifica emparejando la proporción de una o más características 1509 de las muestras en el conjunto de datos de lotes con muestras de referencia, por ejemplo, en una base de datos de referencia 1511, que tienen la misma proporción de esas una o más características (por ejemplo, emplazamiento del tejido, pureza del tumor, tipo de cáncer, procedimiento de recogida, identidad del secuenciador, fecha de secuenciación, datos clínicos (por ejemplo, diagnóstico del paciente, resultados del tratamiento, etc.), sexo, tipo de biopsia (por ejemplo, biopsia de sangre frente a biopsia sólida), datos moleculares (por ejemplo, mutaciones genómicas) y/u otras características). El conjunto de datos de referencia emparejados por cohorte incluye un perfil de expresión para cada muestra de referencia en la pluralidad de muestras de referencia. Cada perfil de expresión respectivo incluye un valor de expresión génica correspondiente para cada gen respectivo de la pluralidad de genes (por ejemplo, cuando la pluralidad de genes incluidos en cada perfil de expresión de la muestra de referencia es la misma que la pluralidad de genes incluidos en cada perfil de expresión de la muestra de ensayo).

En algunas realizaciones, la pluralidad de muestras de referencia comprende el mismo número de muestras de referencia que el número de muestras de ensayo en el lote de muestras de ensayo (por ejemplo, N muestras de referencia). En algunas realizaciones, la pluralidad de muestras de referencia se selecciona para el lote de muestras de ensayo basándose en una o más similitudes entre las características asociadas con las muestras de referencia y las muestras de ensayo (por ejemplo, emplazamiento del tejido, pureza del tumor, tipo de cáncer, procedimiento de recolección, identidad del secuenciador y/o fecha de secuenciación). En algunas realizaciones, los metadatos (por ejemplo, los valores de las características) entre el lote de muestras de ensayo y la pluralidad de muestras de referencia se equilibran de manera que la distribución (por ejemplo, las proporciones) de las características de las muestras de ensayo entre la pluralidad de muestras de ensayo del lote de muestras de ensayo sea similar a la distribución (por ejemplo, las proporciones) de las características de las muestras de referencia entre la pluralidad de muestras de referencia.

En algunas realizaciones, el conjunto de datos de referencia emparejado por cohorte se normaliza.

De acuerdo con el procedimiento, en referencia al bloque 1514, la reducción de dimensiones (por ejemplo, PCA, análisis de componentes latentes, regresión de mínimos cuadrados parciales, etc.) se realiza en un conjunto de datos combinado que consiste en el perfil de expresión correspondiente para cada muestra de ensayo respectiva en la pluralidad de muestras de ensayo y el perfil de expresión correspondiente para cada muestra de referencia respectiva

en la pluralidad de muestras de referencia. Para cada muestra de ensayo P respectiva y cada muestra de referencia M respectiva, un conjunto correspondiente de coordenadas 1516-1518 se incrusta en un espacio dimensional inferior (por ejemplo, espacio-m) a las dimensiones del perfil de expresión correspondiente (por ejemplo, 1516-1, 1516-2,...,1516-N, y 1518-1, 1518-2,...,1518-N).

- 5 Haciendo referencia al bloque 1520, el procedimiento incluye además evaluar los valores de componentes utilizando el conjunto de datos combinado tras la reducción de dimensionalidad. El procedimiento incluye la determinación de una medida estadística de similitud entre los conjuntos de coordenadas obtenidos para las muestras de ensayo y los conjuntos de coordenadas obtenidos para las muestras de referencia, y la comparación de la medida estadística de similitud con un valor umbral. En referencia al bloque 1522, el conjunto de datos de lotes se valida para emitir una
- 10 notificación cuando la medida estadística de similitud satisface el valor umbral, y no se valida cuando la medida estadística de similitud no satisface el valor umbral. En algunas realizaciones, cuando la medida estadística de similitud no satisface el valor umbral, el conjunto de datos de lotes se marca para su rechazo y/o evaluación posterior de los efectos de lote. En algunas realizaciones, el análisis posterior identifica muestras individuales o grupos de muestras individuales en el conjunto de datos de lotes que provocan la disimilitud con las muestras de referencia. En algunas
- 15 realizaciones, una o más de estas muestras que provocan la diferencia estadística se eliminan del conjunto de datos de lotes, la prueba de control de calidad global se vuelve a ejecutar con el conjunto de datos de lotes modificado (por ejemplo, en el que se han eliminado las muestras individuales que contribuyen a la disimilitud estadística), y el conjunto de datos modificado se valida si supera la prueba de control de calidad de lotes. En algunas realizaciones, cuando se identifica que una o más muestras contribuyen a un efecto de lote identificado, se determinan y aplican factores de corrección para normalizar dichas una o más muestras al conjunto de datos de referencia. De este modo, estas
- 20 muestras pueden validarse y utilizarse para análisis posteriores. En algunas realizaciones, cuando se identifica que una o más muestras contribuyen a un efecto de lote identificado, dichas una o más muestras se rechazan (por ejemplo, tras una inspección manual o automáticamente). En algunas realizaciones, las muestras rechazadas se vuelven a ensayar en la canalización de datos de expresión de ARN.
- 25 Cabe señalar que los detalles de otros procesos descritos en el presente documento con respecto a otros procedimientos descritos en el presente documento (por ejemplo, los procedimientos mostrados en las figuras 2, 16 y 17) también son aplicables de manera análoga al procedimiento descrito anteriormente con respecto a la figura 15. Por ejemplo, los detalles relativos a la recopilación de datos, el procesamiento de datos, el emparejamiento por cohorte, el análisis de reducción de dimensiones, etc., descritos anteriormente con referencia al procedimiento resumido en la figura 15 tienen opcionalmente una o más de las características de la recopilación de datos, el procesamiento de datos, el emparejamiento por cohorte, el análisis de reducción de dimensiones, etc., descritos en el
- 30 presente documento con referencia a otros procedimientos descritos en el presente documento (por ejemplo, los resumidos en las figuras 2, 16 y 17). En aras de la brevedad, estos detalles no se repiten en este punto.

Validación de un cambio en la canalización de datos bioinformáticos

- 35 Otro aspecto de la presente divulgación proporciona un procedimiento de validación de un cambio en la canalización de datos de expresión de ARN en un sistema informático que tiene uno o más procesadores y una memoria que almacena uno o más programas para su ejecución por dichos uno o más procesadores. El procedimiento incluye la obtención, en formato electrónico, de un conjunto de datos de lotes que comprende, para cada muestra de ensayo respectiva de un lote de muestras de ensayo, un perfil de expresión correspondiente preparado utilizando una primera
- 40 canalización de datos de expresión de ARN. El perfil de expresión correspondiente incluye un valor de expresión génica correspondiente para cada gen respectivo en un primer conjunto de genes. El conjunto de datos de lotes comprende además un conjunto de metadatos correspondiente que comprenden valores para cada característica respectiva en un primer conjunto de características sobre la muestra de ensayo respectiva.
- 45 El procedimiento incluye determinar para el conjunto de datos de lotes, un conjunto de datos de referencia emparejado por cohorte que incluye, para cada muestra de referencia respectiva en una pluralidad de muestras de referencia, un perfil de expresión correspondiente preparado usando una segunda canalización de datos de expresión de ARN (por ejemplo, la canalización existente antes del cambio), que comprende un valor de expresión génica correspondiente para cada gen respectivo en el primer conjunto de genes. Cada muestra de referencia respectiva de la pluralidad de muestras de referencia se asocia con un conjunto de metadatos correspondiente que comprende un valor
- 50 correspondiente para cada característica respectiva de un segundo conjunto de características sobre la muestra de referencia respectiva. Los valores agregados para cada característica respectiva en un tercer conjunto de una o más características que están presentes tanto en el primer conjunto de características como en el segundo conjunto de características se equilibran entre el conjunto de datos de lotes y el conjunto de datos de referencia emparejado por cohorte.
- 55 La reducción de dimensión se realiza en un conjunto de datos combinado que consiste en el perfil de expresión correspondiente para cada muestra de ensayo respectiva en la pluralidad de muestras de ensayo y el perfil de expresión correspondiente para cada muestra de referencia respectiva en la pluralidad de muestras de referencia. Así, se obtiene para cada muestra de ensayo respectiva y cada muestra de referencia respectiva, un conjunto correspondiente de coordenadas incrustadas en un espacio dimensional inferior a las dimensiones del perfil de
- 60 expresión correspondiente.

El procedimiento comprende además determinar una medida estadística de similitud entre los conjuntos de coordenadas obtenidos para las muestras de ensayo y los conjuntos de coordenadas obtenidos para las muestras de referencia. La medida estadística de similitud se compara con un valor umbral, y el cambio en la canalización de datos de expresión de ARN se valida cuando la medida estadística de similitud satisface el valor umbral, o el cambio en la canalización de datos de expresión de ARN no se valida cuando la medida estadística de similitud no satisface el valor umbral.

Por ejemplo, la figura 16 ilustra un procedimiento de validación de un cambio en la canalización de datos de expresión de ARN, de acuerdo con algunas realizaciones de la presente divulgación (por ejemplo, determinar la manejabilidad de un nuevo proceso y/o determinar un factor de corrección para un nuevo proceso). Se obtiene un conjunto de datos de lotes a partir de un lote de muestras de ensayo, en el que el conjunto de datos de lotes incluye un perfil de expresión correspondiente para cada muestra de ensayo respectiva en el lote de muestras de ensayo. En algunas realizaciones, cada perfil de expresión respectivo incluye un valor de expresión génica correspondiente para cada gen respectivo en una pluralidad de genes en la muestra de ensayo (por ejemplo, determinado por una canalización de datos de expresión de ARN). En algunas realizaciones, cada perfil de expresión respectivo incluye además metadatos que indican valores para una pluralidad de características asociadas con la muestra de ensayo, tales como perfiles de transcripción de ARN (por ejemplo, determinados a partir de una pluralidad de lecturas de secuencia), datos clínicos (por ejemplo, diagnóstico del paciente, resultados del tratamiento, etc.), sexo, tipo de biopsia (por ejemplo, biopsia de sangre frente a biopsia sólida), datos moleculares (por ejemplo, mutaciones genómicas) y/u otras características (por ejemplo, emplazamiento del tejido, pureza del tumor, tipo de cáncer, procedimiento de recogida, identidad del secuenciador y/o fecha de secuenciación).

En algunas realizaciones, cada perfil de expresión se prepara usando un análisis de secuenciación (por ejemplo, una canalización de datos de expresión de ARN) de cada muestra de ensayo respectiva en el lote de muestras de ensayo. En algunas realizaciones, cada muestra de ensayo respectiva del lote de muestras de ensayo se somete a un primer análisis de secuenciación (por ejemplo, una primera canalización de datos de expresión de ARN, "vX"). Haciendo referencia al bloque 1602, en algunas realizaciones, un cambio de proceso en el proceso de expresión de ARN comprende cambiar el análisis de secuenciación de una primera canalización (por ejemplo, vX) a un segundo proceso (por ejemplo, vY). Refiriéndose al bloque 1604, en algunas realizaciones, cada muestra de ensayo respectiva del lote de muestras de ensayo se somete además a un segundo análisis de secuenciación (por ejemplo, una segunda canalización de datos de expresión de ARN, "vY").

Así, en algunas realizaciones, el procedimiento comprende obtener un primer conjunto de datos de lotes 1608 (por ejemplo, un conjunto de datos de lotes vX) que incluye, para cada muestra de ensayo respectiva en el lote de muestras de ensayo, un primer perfil de expresión correspondiente obtenido usando un primer proceso (por ejemplo, vX) y un segundo conjunto de datos de lotes (por ejemplo, un conjunto de datos de lotes vY) que incluye, para cada muestra de ensayo respectiva en el lote de muestras de ensayo, un segundo perfil de expresión correspondiente obtenido utilizando un segundo proceso (por ejemplo, vY).

En algunas realizaciones, los datos de expresión en cada primer perfil de expresión respectivo en el primer conjunto de datos de lotes y los datos de expresión en cada segundo perfil de expresión respectivo en el segundo conjunto de datos de lotes se normalizan.

En algunas realizaciones, el procedimiento incluye además obtener un conjunto de datos de referencia emparejados por cohorte 1512 para una pluralidad de muestras de referencia. En algunas realizaciones, el conjunto de datos de referencia emparejados por cohorte se identifica emparejando la proporción de una o más características 1603 de las muestras en el conjunto de datos de lotes con muestras de referencia, por ejemplo, en una base de datos de referencia 1605, que tienen la misma proporción de esas una o más características (por ejemplo, emplazamiento del tejido, pureza del tumor, tipo de cáncer, procedimiento de recogida, identidad del secuenciador, fecha de secuenciación, datos clínicos (por ejemplo, diagnóstico del paciente, resultados del tratamiento, etc.), sexo, tipo de biopsia (por ejemplo, biopsia de sangre frente a biopsia sólida), datos moleculares (por ejemplo, mutaciones genómicas) y/u otras características). El conjunto de datos de referencia emparejados por cohorte incluye un perfil de expresión para cada muestra de referencia en la pluralidad de muestras de referencia. Cada perfil de expresión respectivo incluye un valor de expresión génica correspondiente para cada gen respectivo de la pluralidad de genes (por ejemplo, cuando la pluralidad de genes incluidos en cada perfil de expresión de la muestra de referencia es la misma que la pluralidad de genes incluidos en cada perfil de expresión de la muestra de ensayo).

En algunas realizaciones, el conjunto de datos de referencia emparejado por cohorte es un conjunto de datos emparejado por muestra 1608. Es decir, en algunas realizaciones, las mismas muestras se ejecutan a través de ambas versiones de la canalización de datos de expresión de ARN, y se comparan entre sí como el conjunto de datos de lotes (por ejemplo, generado a partir de las muestras ejecutadas a través de la nueva versión de la canalización de datos de expresión de ARN) y el conjunto de datos de referencia (por ejemplo, generado a partir de las muestras ejecutadas a través de la versión anterior de la canalización de datos de expresión de ARN).

En algunas realizaciones, la pluralidad de muestras de referencia se selecciona para el lote de muestras de ensayo basándose en una o más similitudes entre las características asociadas con las muestras de referencia y las muestras de ensayo procesadas usando el proceso vY (por ejemplo, emplazamiento del tejido, pureza del tumor, tipo de cáncer,

procedimiento de recogida, identidad del secuenciador y/o fecha de secuenciación). En algunas realizaciones, los metadatos (por ejemplo, los valores de las características) entre el lote de muestras de ensayo y la pluralidad de muestras de referencia se equilibran de manera que la distribución (por ejemplo, las proporciones) de las características de las muestras de ensayo entre la pluralidad de muestras de ensayo procesadas mediante el proceso vY son similares a la distribución (por ejemplo, las proporciones) de las características de las muestras de referencia entre la pluralidad de muestras de referencia.

En algunas realizaciones, el conjunto de datos de referencia emparejado por cohorte se normaliza.

De acuerdo con el procedimiento, en referencia al bloque 1610, la reducción de dimensiones (por ejemplo, PCA, análisis de componentes latentes, regresión de mínimos cuadrados parciales, etc.) se realiza en un conjunto de datos combinado que incluye el segundo perfil de expresión correspondiente para cada muestra de ensayo respectiva, procesada usando el proceso vY, en la pluralidad de muestras de ensayo y el perfil de expresión correspondiente para cada muestra de referencia respectiva, procesada usando el proceso vX, en el conjunto de datos de referencia emparejados por cohorte. En algunas realizaciones, el conjunto de datos combinados incluye además el correspondiente primer perfil de expresión para cada muestra de ensayo respectiva procesada mediante el proceso vX en la pluralidad de muestras de ensayo.

Así, para cada muestra de ensayo respectiva y cada muestra de referencia respectiva procesada usando el proceso vX, y para cada muestra de ensayo respectiva procesada usando el proceso vY, un conjunto correspondiente de coordenadas 1612-1614 se incrusta en un espacio dimensional más bajo (por ejemplo, espacio-m) que las dimensiones del perfil de expresión correspondiente (por ejemplo, 1612-1, 1612-2,...,1612-N, y 1614-1, 1614-2,...,1614-N).

Haciendo referencia al bloque 1616, el procedimiento incluye además evaluar los valores de componentes utilizando el conjunto de datos combinado tras la reducción de dimensionalidad. El procedimiento incluye la determinación de una medida estadística de similitud entre los conjuntos de coordenadas obtenidos para las muestras de ensayo y las muestras de referencia procesadas mediante el proceso vX, y las muestras de ensayo procesadas mediante el proceso vY. En referencia al bloque 1618, la medida estadística de similitud se compara con un valor umbral para determinar si existe una varianza significativa entre el proceso vX y el proceso vY. Refiriéndose al bloque 1620, el cambio en la canalización de datos de expresión de ARN se valida cuando la medida estadística de similitud satisface el valor umbral. Refiriéndose al bloque 1622, el cambio en la canalización de datos de expresión de ARN se rechaza y/o se marca para una evaluación posterior del cambio de proceso y/o para la determinación de un factor de corrección para el proceso vY cuando la medida estadística de similitud no satisface el valor umbral.

Cabe señalar que los detalles de otros procesos descritos en el presente documento con respecto a otros procedimientos descritos en el presente documento (por ejemplo, los procedimientos mostrados en las figuras 2, 15 y 17) también son aplicables de manera análoga al procedimiento descrito anteriormente con respecto a la figura 16. Por ejemplo, los detalles relativos a la recopilación de datos, el procesamiento de datos, el emparejamiento por cohorte, el análisis de reducción de dimensiones, etc., descritos anteriormente con referencia al procedimiento resumido en la figura 16 tienen opcionalmente una o más de las características de la recopilación de datos, el procesamiento de datos, el emparejamiento por cohorte, el análisis de reducción de dimensiones, etc., descritos en el presente documento con referencia a otros procedimientos descritos en el presente documento (por ejemplo, los resumidos en las figuras 2, 15 y 17). En aras de la brevedad, estos detalles no se repiten en este punto.

Ampliación de la base de datos de referencia

Otro aspecto de la presente divulgación proporciona un procedimiento para añadir datos de expresión de ARN a una base de datos de referencia, en un sistema informático que tiene uno o más procesadores y una memoria que almacena uno o más programas para su ejecución por dichos uno o más procesadores. El procedimiento incluye la obtención, en formato electrónico, de un nuevo conjunto de datos de expresión. El nuevo conjunto de datos de expresión incluye, para cada muestra de ensayo respectiva de una pluralidad de muestras de ensayo, un perfil de expresión correspondiente preparado utilizando una primera canalización de datos de expresión de ARN, en la que el perfil de expresión correspondiente comprende un valor de expresión génica correspondiente para cada gen respectivo de un primer conjunto de genes. El nuevo conjunto de datos de expresión incluye además un conjunto de metadatos correspondiente que incluyen valores para cada característica respectiva en un primer conjunto de características sobre la muestra de ensayo respectiva.

El procedimiento incluye determinar para el nuevo conjunto de datos de expresión, un conjunto de datos de referencia emparejados por cohorte que incluye, para cada muestra de referencia respectiva en una pluralidad de muestras de referencia, un perfil de expresión correspondiente que comprende un valor de expresión génica correspondiente para cada gen respectivo en el primer conjunto de genes. Cada muestra de referencia respectiva de la pluralidad de muestras de referencia se asocia con un conjunto de metadatos correspondiente que comprende un valor correspondiente para cada característica respectiva de un segundo conjunto de características sobre la muestra de referencia respectiva. Cada perfil de expresión correspondiente a una muestra de referencia de la pluralidad de muestras de referencia procede de la base de datos de referencia. Los valores agregados para cada característica respectiva en un tercer conjunto de una o más características que están presentes tanto en el primer conjunto de

características como en el segundo conjunto de características se equilibran entre el conjunto de datos de lotes y el conjunto de datos de referencia emparejado por cohorte.

La reducción de dimensión se realiza en un conjunto de datos combinado que consiste en el perfil de expresión correspondiente para cada muestra de ensayo respectiva en la pluralidad de muestras de ensayo y el perfil de expresión correspondiente para cada muestra de referencia respectiva en la pluralidad de muestras de referencia. Así, se obtiene para cada muestra de ensayo respectiva y cada muestra de referencia respectiva, un conjunto correspondiente de coordenadas incrustadas en un espacio dimensional inferior a las dimensiones del perfil de expresión correspondiente.

El procedimiento comprende además determinar una medida estadística de similitud entre los conjuntos de coordenadas obtenidos para las muestras de ensayo y los conjuntos de coordenadas obtenidos para las muestras de referencia, y la medida estadística de similitud se compara con un valor umbral. El procedimiento incluye añadir el nuevo conjunto de datos de expresión a la base de datos de referencia cuando la medida estadística de similitud satisface el valor umbral, o, cuando la medida estadística de similitud no satisface el valor umbral, determinar un conjunto de factores de conversión para normalizar los perfiles de expresión en el nuevo conjunto de datos de expresión frente a los perfiles de expresión en la base de datos de referencia, normalizar los perfiles de expresión en el nuevo conjunto de datos de expresión utilizando el conjunto de factores de conversión, obteniendo así un nuevo conjunto de datos de expresión normalizado, y añadir el nuevo conjunto de datos de expresión normalizado a la base de datos de referencia.

Por ejemplo, la figura 17 ilustra un procedimiento de adición de datos de expresión de ARN a una base de datos de referencia, de acuerdo con algunas realizaciones de la presente divulgación (por ejemplo, validación de datos de expresión recién adquiridos para su uso en la actualización de una base de datos de referencia).

Haciendo referencia al bloque 1702, se obtiene un nuevo conjunto de datos de expresión (por ejemplo, de datos de expresión de ARN). En algunas realizaciones, el nuevo conjunto de datos de expresión incluye, para cada muestra de ensayo respectiva en una pluralidad de muestras de ensayo, un perfil de expresión preparado utilizando una primera canalización de datos de expresión de ARN e incluyendo valores de expresión génica para una pluralidad de genes. El nuevo conjunto de datos de expresión incluye además un conjunto de metadatos correspondiente que incluyen valores de características asociadas a la muestra de ensayo.

Haciendo referencia al bloque 1704, en algunas realizaciones, el nuevo conjunto de datos de expresión se normaliza.

En algunas realizaciones, el procedimiento incluye además obtener un conjunto de datos de referencia emparejado por cohorte 1706 para el nuevo conjunto de datos de expresión. En algunas realizaciones, el conjunto de datos de referencia emparejados por cohorte se identifica emparejando la proporción de una o más características 1703 de las muestras en el conjunto de datos de lotes con muestras de referencia, por ejemplo, en una base de datos de referencia 1705, que tienen la misma proporción de esas una o más características (por ejemplo, emplazamiento del tejido, pureza del tumor, tipo de cáncer, procedimiento de recogida, identidad del secuenciador, fecha de secuenciación, datos clínicos (por ejemplo, diagnóstico del paciente, resultados del tratamiento, etc.), sexo, tipo de biopsia (por ejemplo, biopsia de sangre frente a biopsia sólida), datos moleculares (por ejemplo, mutaciones genómicas) y/u otras características). El conjunto de datos de referencia emparejados por cohorte incluye un perfil de expresión para cada muestra de referencia en la pluralidad de muestras de referencia. Cada perfil de expresión respectivo incluye un valor de expresión génica correspondiente para cada gen respectivo de la pluralidad de genes (por ejemplo, cuando la pluralidad de genes incluidos en cada perfil de expresión de la muestra de referencia es la misma que la pluralidad de genes incluidos en cada perfil de expresión de la muestra de ensayo).

En algunas realizaciones, el conjunto de datos de referencia emparejados por cohorte incluye un perfil de expresión para cada muestra de referencia en una pluralidad de muestras de referencia. En algunas realizaciones, cada perfil de expresión respectivo procede de la base de datos de referencia (por ejemplo, una base de datos existente) e incluye un valor de expresión génica correspondiente para cada gen respectivo de la pluralidad de genes (por ejemplo, como se incluye en el nuevo conjunto de datos de expresión).

Cada muestra de referencia respectiva en la pluralidad de muestras de referencia se asocia con un conjunto de metadatos correspondiente que incluyen valores para características asociadas con la muestra de referencia. En algunas realizaciones, el conjunto de datos de referencia emparejado por cohorte se empareja (por ejemplo, se equilibra) de acuerdo con cualquiera de los procedimientos descritos en las figuras 15 y/o 16.

En algunas realizaciones, el conjunto de datos de referencia emparejado por cohorte se normaliza.

De acuerdo con el procedimiento, en referencia al bloque 1708, la reducción de dimensiones (por ejemplo, PCA, análisis de componentes latentes, regresión de mínimos cuadrados parciales, etc.) se realiza en un conjunto de datos combinado que consiste en el perfil de expresión correspondiente para cada muestra de ensayo respectiva en la pluralidad de muestras de ensayo y el perfil de expresión correspondiente para cada muestra de referencia respectiva en la pluralidad de muestras de referencia. Para cada muestra de ensayo respectiva N y cada muestra de referencia respectiva C, un conjunto correspondiente de coordenadas 1710-1712 se incrusta en un espacio dimensional inferior

(por ejemplo, espacio-m) a las dimensiones del perfil de expresión correspondiente (por ejemplo, 1710-1, 1710-2,...,1710-N, y 1712-1, 1712-2,...,1712-N).

Haciendo referencia al bloque 1714, el procedimiento incluye además evaluar los valores de componentes utilizando el conjunto de datos combinado tras la reducción de dimensionalidad. El procedimiento incluye la determinación de una medida estadística de similitud entre los conjuntos de coordenadas obtenidos para las muestras de ensayo y los conjuntos de coordenadas obtenidos para las muestras de referencia. En referencia al bloque 1716, la medida estadística de similitud se compara con un valor umbral para determinar si existe una varianza significativa entre el nuevo conjunto de datos de expresión y los datos de la base de datos existente.

Haciendo referencia al bloque 1718, el nuevo conjunto de datos de expresión se añade a la base de datos de referencia cuando la medida estadística de similitud satisface el valor umbral. En referencia al bloque 1720, cuando la medida estadística de similitud no satisface el valor umbral, se determina un conjunto de factores de conversión para normalizar los perfiles de expresión en el nuevo conjunto de datos de expresión frente a los perfiles de expresión en la base de datos de referencia, y los perfiles de expresión en el nuevo conjunto de datos de expresión se normalizan utilizando el conjunto de factores de conversión, obteniendo así un nuevo conjunto de datos de expresión normalizado, y el nuevo conjunto de datos de expresión normalizado se añade a la base de datos de referencia.

Cabe señalar que los detalles de otros procesos descritos en el presente documento con respecto a otros procedimientos descritos en el presente documento (por ejemplo, los procedimientos mostrados en las figuras 2, 15 y 16) también son aplicables de manera análoga al procedimiento descrito anteriormente con respecto a la figura 17. Por ejemplo, los detalles relativos a la recopilación de datos, el procesamiento de datos, el emparejamiento por cohorte, el análisis de reducción de dimensiones, etc., descritos anteriormente con referencia al procedimiento resumido en la figura 17 tienen opcionalmente una o más de las características de la recopilación de datos, el procesamiento de datos, el emparejamiento por cohorte, el análisis de reducción de dimensiones, etc., descritos en el presente documento con referencia a otros procedimientos descritos en el presente documento (por ejemplo, los resumidos en las figuras 2, 15 y 16). En aras de la brevedad, estos detalles no se repiten en este punto.

Ejemplos de realización

En algunas realizaciones de los sistemas y procedimientos descritos en el presente documento (por ejemplo, los procedimientos resumidos en las figuras 2, 15, 16 y 17, como se ha descrito anteriormente), la obtención del conjunto de datos de lotes comprende, para cada muestra respectiva en el lote de muestras, obtener, en forma electrónica, una pluralidad correspondiente de lecturas de secuencia derivadas de la muestra respectiva mediante secuenciación de ARN selectiva o del transcriptoma completo, y determinar, a partir de la pluralidad correspondiente de lecturas de secuencia, el valor de expresión génica correspondiente para cada gen respectivo en el primer conjunto de genes. En algunas realizaciones, los procedimientos descritos en el presente documento también incluyen una etapa de generación de los datos de secuenciación. Sin embargo, en otros aspectos, los procedimientos descritos en el presente documento comienzan después de que ya se haya realizado la secuenciación. Por ejemplo, en algunas realizaciones, los procedimientos descritos en el presente documento comienzan obteniendo lecturas de secuencia para cada muestra de un lote de muestras, en formato electrónico, determinando un perfil de expresión para cada muestra del lote de muestras basándose en las lecturas de secuencia y, a continuación, realizando uno o más procedimientos de control de calidad, como se describe en el presente documento. Del mismo modo, en algunas realizaciones, los procedimientos descritos en el presente documento comienzan por obtener perfiles de expresión, en formato electrónico, para cada muestra de un lote de muestras, y luego realizando uno o más procedimientos de control de calidad, como se describe en el presente documento.

En algunas realizaciones, para cada muestra respectiva en el lote de muestras, la pluralidad correspondiente de lecturas de secuencia es de al menos 10 000 lecturas de secuencia. En algunas realizaciones, la pluralidad correspondiente de lecturas de secuencia es de al menos 100 000 lecturas de secuencia. En algunas realizaciones, la pluralidad correspondiente de lecturas de secuencia es de al menos 1 000 000 de lecturas de secuencia. En algunas realizaciones, la pluralidad correspondiente de lecturas de secuencia es de al menos 10 000 000 de lecturas de secuencia. En algunas realizaciones, la pluralidad correspondiente de lecturas de secuencia es de 10 000 a 100 000 000 de lecturas de secuencia. En algunas realizaciones, la pluralidad correspondiente de lecturas de secuencia es de 100 000 a 50 000 000 de lecturas de secuencia. En algunas realizaciones, la pluralidad correspondiente de lecturas de secuencia es de 1 000 000 a 50 000 000 de lecturas de secuencia.

En algunas realizaciones, el lote de muestras de ensayo incluye al menos 10 muestras de ensayo. En algunas realizaciones, el lote de muestras de ensayo incluye al menos 25 muestras de ensayo. En algunas realizaciones, el lote de muestras de ensayo incluye al menos 100 muestras de ensayo. En algunas realizaciones, el lote de muestras de ensayo incluye al menos 1000 muestras de ensayo. En algunas realizaciones, el lote de muestras de ensayo incluye al menos 5, al menos 6, al menos 7, al menos 8, al menos 9, al menos 10, al menos 25, al menos 50, al menos 100, al menos 250, al menos 500, al menos 750, al menos 1000, al menos 2500, al menos 5000, al menos 10 000, al menos 100 000, al menos 1 000 000 o más muestras. En algunas realizaciones, el lote de muestras de ensayo incluye de 5 a 100 muestras de ensayo. En algunas realizaciones, el lote de muestras de ensayo incluye de 50 a 500 muestras de ensayo. En algunas realizaciones, el lote de muestras de ensayo incluye de 100 a 1000 muestras de ensayo. En algunas realizaciones, el lote de muestras de ensayo incluye de 1000 a 100 000 muestras de ensayo.

- 5 En algunas realizaciones, el primer conjunto de genes incluye al menos 10 genes. En algunas realizaciones, el primer conjunto de genes incluye al menos 100 genes. En algunas realizaciones, el primer conjunto de genes incluye al menos 1000 genes. En algunas realizaciones, el primer conjunto de genes incluye al menos 10.000 genes. En algunas realizaciones, el lote de muestras de ensayo incluye al menos 5, al menos 6, al menos 7, al menos 8, al menos 9, al menos 10, al menos 25, al menos 50, al menos 100, al menos 250, al menos 500, al menos 750, al menos 1000, al menos 2500, al menos 5000, al menos 10 000, al menos 20 000, al menos 30 000 o más genes.
- 10 En algunas realizaciones, el conjunto de características sobre las que se equilibran los conjuntos de datos de lote y de referencia incluye al menos una característica seleccionada de un emplazamiento de un tejido (el emplazamiento del que se adquirió la muestra biológica), una pureza de tumor, un tipo de cáncer, una identidad de secuenciador y una fecha de secuenciación. En algunas realizaciones, el conjunto de características sobre las que se equilibran los conjuntos de datos del lote y de referencia incluye al menos dos características seleccionadas entre un emplazamiento del tejido, una pureza del tumor, un tipo de cáncer, una identidad del secuenciador y una fecha de secuenciación. En algunas realizaciones, el conjunto de características sobre las que se equilibran los conjuntos de datos del lote y de referencia incluye al menos tres características seleccionadas entre un emplazamiento del tejido, una pureza del tumor, un tipo de cáncer, una identidad del secuenciador y una fecha de secuenciación. En algunas realizaciones, el conjunto de características sobre las que se equilibran los conjuntos de datos del lote y de referencia incluye al menos un emplazamiento del tejido y un tipo de cáncer. En algunas realizaciones, el conjunto de características sobre las que se equilibran los conjuntos de datos del lote y de referencia es un emplazamiento del tejido y un tipo de cáncer.
- 15 En algunas realizaciones, el conjunto de características sobre las que se equilibran los conjuntos de datos de lote y de referencia incluye al menos una característica seleccionada entre un procedimiento de extracción de ácido nucleico, un procedimiento de preparación de bancos de ADNc, un procedimiento de secuenciación de ARN, un tipo de reactivo utilizado y un tipo de equipo utilizado.
- 20 En algunas realizaciones, la pluralidad de muestras de referencia (el conjunto de datos de referencia) incluye al menos 50 muestras de referencia. En algunas realizaciones, la pluralidad de muestras de referencia (el conjunto de datos de referencia) incluye al menos 100 muestras de referencia. En algunas realizaciones, la pluralidad de muestras de referencia (el conjunto de datos de referencia) incluye al menos 500 muestras de referencia. En algunas realizaciones, la pluralidad de muestras de referencia (el conjunto de datos de referencia) incluye al menos 1000 muestras de referencia. En algunas realizaciones, la pluralidad de muestras de referencia (el conjunto de datos de referencia) incluye al menos 5000 muestras de referencia. En algunas realizaciones, la pluralidad de muestras de referencia (el conjunto de datos de referencia) incluye al menos 10.000 muestras de referencia. En algunas realizaciones, la pluralidad de muestras de referencia (el conjunto de datos de referencia) incluye al menos 100.000 muestras de referencia. En algunas realizaciones, la pluralidad de muestras de referencia incluye al menos 5, al menos 6, al menos 7, al menos 8, al menos 9, al menos 10, al menos 25, al menos 50, al menos 100, al menos 250, al menos 500, al menos 750, al menos 1000, al menos 2500, al menos 5000, al menos 10 000, al menos 100 000, al menos 1 000 000 o más muestras. En algunas realizaciones, la pluralidad de muestras de referencia incluye de 5 a 100 muestras de referencia. En algunas realizaciones, la pluralidad de muestras de referencia incluye de 50 a 500 muestras de referencia. En algunas realizaciones, la pluralidad de muestras de referencia incluye de 100 a 1000 muestras de referencia. En algunas realizaciones, la pluralidad de muestras de referencia incluye de 1000 a 100 000 muestras de referencia.
- 25 En algunas realizaciones, el conjunto de muestras de referencia equilibrado con respecto a un conjunto de datos de lotes incluye al menos tantas muestras como hay en el conjunto de datos de lotes. En algunas realizaciones, el conjunto de muestras de referencia equilibrado con respecto a un conjunto de datos de lotes tiene el mismo número de muestras que el conjunto de datos de lotes. En algunas realizaciones, el conjunto de muestras de referencia equilibrado con respecto a un conjunto de datos de lotes incluye al menos un 25 % más de muestras que las que hay en el conjunto de datos de lotes. En algunas realizaciones, el conjunto de muestras de referencia equilibrado con respecto a un conjunto de datos de lotes incluye al menos un 50 % más de muestras que las que hay en el conjunto de datos de lotes. En algunas realizaciones, el conjunto de muestras de referencia equilibrado con respecto a un conjunto de datos de lotes incluye al menos un 100 % más de muestras que las que hay en el conjunto de datos de lotes. En algunas realizaciones, el conjunto de muestras de referencia equilibrado con respecto a un conjunto de datos de lotes incluye al menos 5 veces más muestras que las que hay en el conjunto de datos de lotes. En algunas realizaciones, el conjunto de muestras de referencia equilibrado con respecto a un conjunto de datos de lotes incluye al menos 10 veces más muestras que las que hay en el conjunto de datos de lotes.
- 30 El procedimiento de una cualquiera de las reivindicaciones 27 a 36, en el que la pluralidad de muestras de referencia comprende al menos 1000 muestras de referencia.
- 35 En algunas realizaciones, el valor agregado para una característica respectiva está equilibrado entre el conjunto de datos de lotes y el conjunto de referencia emparejado por cohorte cuando el porcentaje de muestras de ensayo respectivas, en el lote de muestras de ensayo, que tienen un valor respectivo para la característica respectiva está dentro del 2,5 % del porcentaje de muestras de referencia respectivas, en la pluralidad de muestras de referencia, que tienen el mismo valor respectivo para la característica respectiva. Por ejemplo, en algunas realizaciones en las que el tipo de cáncer es una característica que se equilibra entre un conjunto de datos de lotes y un conjunto de datos emparejado por cohorte, si el conjunto de datos de lotes está compuesto por un 20 % de muestras de cáncer cerebral,
- 40
- 45
- 50
- 55
- 60

un 30 % de muestras de cáncer de pulmón y un 50 % de muestras de cáncer de colon, el conjunto de datos de referencia incluiría entre un 17,5 % y un 22,5 % de muestras de cáncer cerebral, entre un 27,5 % y un 32,5 % de muestras de cáncer de pulmón y entre un 47,5 % y un 52,5 % de muestras de cáncer de colon.

5 En algunas realizaciones, el valor agregado para una característica respectiva está equilibrado entre el conjunto de datos de lotes y el conjunto de referencia emparejado por cohorte cuando el porcentaje de muestras de ensayo respectivas, en el lote de muestras de ensayo, que tienen un valor respectivo para la característica respectiva está dentro del 1 %, dentro del 2 %, dentro del 3 %, dentro del 4 %, dentro del 5 %, dentro del 6 %, dentro del 7 %, dentro del 8 %, dentro del 9 %, dentro del 10 %, dentro del 11 %, dentro del 12 %, dentro del 13 %, dentro del 14 %, dentro del 15 %, dentro del 16 %, dentro del 17 %, dentro del 18 %, dentro del 19 %, dentro del 20 %, dentro del 21 %, dentro del 22 %, dentro del 23 %, dentro del 24 %, o dentro del 25 % del porcentaje de muestras de referencia respectivas, en la pluralidad de muestras de referencia, que tienen el mismo valor respectivo para la característica respectiva.

10 En algunas realizaciones, la reducción de dimensiones incluye incrustar, para cada muestra de ensayo respectiva y cada muestra de referencia respectiva, el perfil de expresión correspondiente en una representación bidimensional. En algunas realizaciones, la reducción de dimensiones incluye la incrustación en 2 coordenadas utilizando la aproximación y proyección uniforme de la variedad (UMAP). En algunas realizaciones, la reducción de dimensiones incluye incrustar, para cada muestra de ensayo respectiva y cada muestra de referencia respectiva, el perfil de expresión correspondiente en 2 coordenadas, 3 coordenadas, 4 coordenadas, 5 coordenadas, 6 coordenadas, 7 coordenadas, 8 coordenadas, 9 coordenadas, 10 coordenadas o más coordenadas. En algunas realizaciones, la reducción de dimensiones incluye la incrustación en un sistema de coordenadas inferiores utilizando el análisis de componentes principales (PCA).

Ejemplos

Ejemplo 1: Detección robusta de los efectos de lote de secuenciación en ARN mediante incrustación de baja dimensión con muestras de referencia emparejadas por subtipo

25 Los efectos de lote técnicos, tales como los cambios en los protocolos, los reactivos o la tecnología de secuenciación, pueden invalidar los estudios del transcriptoma a gran escala. Los laboratorios que analizan transcriptomas en distintos tipos de tumores, a lo largo del tiempo o en múltiples instalaciones deben disponer de un procedimiento sistemático para validar la compatibilidad de los datos entre lotes.

30 Un efecto de lote puede manifestarse como un gran cambio en un pequeño número de genes o un pequeño cambio en muchos genes. Un procedimiento robusto de detección del efecto de lote identificará cualquiera de los dos. Además, los resultados de la RNA-Seq masiva dependen del tipo de cáncer y del emplazamiento del tejido. Esto complica la detección del efecto de lote en estudios sobre múltiples tipos de cáncer. Para superar estos retos se desarrolló un procedimiento para evaluar los efectos de lote técnicos en un conjunto heterogéneo de muestras de transcriptoma.

35 Brevemente, las muestras se eligieron a partir de datos de referencia validados para que coincidieran con el conjunto de transcriptomas basándose en el tipo de cáncer y el emplazamiento del tejido. Los perfiles de expresión génica del conjunto de transcriptomas y los datos de referencia correspondientes se integraron en 2 coordenadas utilizando la aproximación y proyección uniforme de la variedad (UMAP). Las propiedades de agrupación de UMAP son muy adecuadas para la detección de efectos de lote. A continuación, se realizaron pruebas de la U de Mann-Whitney en las coordenadas x e y UMAP. Si cualquiera de las pruebas arroja un valor de p inferior a un umbral, por ejemplo, 0,01, es probable que exista un efecto de lote.

40 Como primer ejemplo, este procedimiento se aplicó para determinar si se crean efectos de lote cuando se utilizan diferentes metodologías de extracción de sangre. Brevemente, se generaron datos de RNA-Seq para muestras de sangre emparejadas por cohorte y emparejadas por tejido recogidas utilizando tubos de recogida PAX o tubos de recogida EDTA. Todas las demás etapas de preparación de muestras, recogida de datos y procesamiento de datos se realizaron de la misma manera para todas las muestras. A continuación, los datos de RNA-Seq se incrustaron en 2 coordenadas utilizando UMAP (figura 3A; 302 = tubos de recogida PAX; 304 = tubos de recogida EDTA). Las pruebas de la U de Mann-Whitney se aplicaron por separado a las coordenadas x e y de la incrustación UMAP. Tal como se muestra en las figuras 3B y 3C, ambas pruebas de la U de Mann-Whitney indicaron una diferencia estadísticamente significativa entre los datos de RNA-Seq generados a partir de sangre recogida en tubos de recogida PAX y sangre recogida en tubos de recogida EDTA ($p = 7,08E-10$), lo que evidencia efectos de lote derivados del uso de diferentes metodologías de recogida de sangre.

55 Como segundo ejemplo, este procedimiento se aplicó para determinar si se crean efectos de lote cuando se utilizan diferentes canalizaciones de datos bioinformáticos para el análisis de datos de RNA-Seq. Brevemente, se generaron datos de RNA-Seq para muestras emparejadas por cohorte y emparejadas por tejido que se procesaron utilizando la canalización STAR o la canalización Kallisto (véase, Dobin A. *et al.*, Bioinformatics, 29(1): 15-21 (2013) (en el que se describe STAR) y Bray N.L. *et al.*, Nature Biotechnology, 34:525-27 (2016) (en el que se describe Kallisto). Todas las etapas de preparación de muestras, recopilación de datos y procesamiento de datos se realizaron de la misma manera para todas las muestras, excepto por la diferencia en la canalización de datos bioinformáticos utilizada para alinear los datos de RNA-Seq y cuantificar la abundancia de transcritos. A continuación, los datos de RNA-Seq se incrustaron

en 2 coordenadas utilizando UMAP (figura 4A; 402 = alineación Star; 404 = alineación Kallisto). Las pruebas de la U de Mann-Whitney se aplicaron por separado a las coordenadas x e y de la incrustación UMAP. Tal como se muestra en las figuras 4B y 4C, ambas pruebas de la U de Mann-Whitney indicaron una diferencia estadísticamente significativa entre los datos RNA-Seq alineados utilizando el algoritmo STAR y los datos RNA-Seq alineados utilizando el algoritmo Kallisto ($p = 2,80E-9$ y $p = 1,69E-19$), lo que evidencia efectos de lote derivados del uso de diferentes algoritmos de alineación de RNA-Seq.

Por último, el procedimiento se utilizó para analizar fuentes habituales propensas a efectos de lote técnicos, a saber, celdas de flujo, actualizaciones de una canalización de datos bioinformáticos y secuenciadores, para cada efecto. Para cada clase técnica (celda de flujo, canalización y secuenciador), el procedimiento se ejecutó en 15 submuestras por característica y se calcularon las tasas de falsos descubrimientos corregidas por Benjamini-Hochberg en todas las submuestras. Las distribuciones presentadas en las figuras 5A-5C representan la mediana de la FDR calculada en las submuestras.

Así, el procedimiento descrito anteriormente es un procedimiento eficaz y fácil de aplicar para ensayar automáticamente los efectos de lote técnicos y de software a través de múltiples tipos de cáncer y emplazamientos del tejido.

Ejemplo 2: Corrección de lotes aplicada tras el rediseño de la sonda de captura

Los procedimientos de captura de RNA-Seq tienen una serie de ventajas sobre los procedimientos de captura de ARNm total, en especial cuando se analiza la expresión génica en muestras FFPE. Por ejemplo, la selección de poliA en procedimientos de captura de ARNm total funciona de forma subóptima en muestras FFPE porque las moléculas de ARN se fragmentan durante el proceso de fijación. Así, muchos fragmentos de ARNm no son capturados porque ya no están asociados a una cola de poli A. Por el contrario, dado que los procedimientos de captura de RNA-Seq aíslan fragmentos de ARNm utilizando sondas diseñadas para la secuencia codificante del ARNm diana, estos procedimientos se ven significativamente menos afectados por la fragmentación. Además, las metodologías de captura de RNA-Seq se ven menos afectadas por la merma de ribosomas y de hemoglobina cuando se utilizan muestras hematológicas.

Sin embargo, se observó que los cambios en el diseño de las sondas de captura del exoma utilizadas para RNA-Seq de captura dan lugar a efectos de lote técnicos menores. Para corregir estos efectos de lote, se seleccionaron 450 muestras representativas de una diversidad de tipos de cáncer que habían sido previamente secuenciadas con un conjunto de sondas de captura del exoma de primera generación para su resecuenciación con un conjunto de sondas de captura del exoma de segunda generación. A continuación, se determinaron los factores de corrección lineal para cada gen comparando los resultados originales de la secuenciación del exoma generados utilizando el conjunto de sondas de captura del exoma de primera generación con los nuevos resultados de la secuenciación del exoma generados utilizando el conjunto de sondas de captura del exoma de segunda generación.

Se comprobó que las correcciones lineales por gen eran suficientes para eliminar todas las diferencias sistemáticas entre los dos conjuntos de datos. Para cada gen i , el valor de expresión corregido a $v1$ E_{ci} se calculó como:

$$E_{ci} = (E_i * m_i) + b_i$$

donde E_i es el nivel de expresión no corregido determinado para el gen i (log2 TPM) utilizando el conjunto de sondas de captura del exoma de segunda generación, m_i es un factor de corrección de la pendiente para el gen i , y b_i es un factor de corrección del intercepto para el gen i . Estos factores de corrección de la pendiente y el intercepto para cada gen se aprenden para que coincidan con las distribuciones $v1$ y $v2$ corregidas en conjuntos de datos coincidentes. Esto se optimizó minimizando una función de pérdida ponderada que puede tener en cuenta el estado de las muestras emparejadas. Al utilizar esta información lateral, los factores de corrección resultantes funcionarán de forma robusta para muestras secuenciadas de cualquier tipo de cáncer, ya que se puede aplicar la misma corrección lineal a todas las muestras procesadas utilizando el conjunto de sondas del exoma de segunda generación.

Para analizar el efecto de los factores de corrección, se realizó un análisis de componentes principales (PCA) sobre los valores de expresión de 100 muestras, que representaban 37 tipos de cáncer diferentes, que se procesaron cada una dos veces; una vez utilizando tanto el conjunto de sondas de captura del exoma de primera generación como el conjunto de sondas de captura del exoma de segunda generación. Estas 100 muestras no formaban parte de la cohorte de entrenamiento utilizada para generar los factores de corrección. El PCA se realizó en primer lugar sobre los valores de expresión no corregidos determinados utilizando el conjunto de sondas de captura del exoma de segunda generación y, a continuación, sobre los valores de expresión corregidos utilizando los factores de corrección descritos anteriormente.

Como se muestra en la figura 6A, cuando se utilizaron valores de expresión no corregidos para el análisis PCA, los efectos de lote técnicos fueron claramente observables por la asociación del tercer componente principal con el tipo de ensayo (+ = conjunto de sondas de captura del exoma de primera generación; O = conjunto de sondas de captura del exoma de segunda generación; las líneas conectan muestras emparejadas). Sin embargo, como se muestra en la

figura 6B, cuando se utilizaron valores de expresión no corregidos para el análisis PCA, ninguno de los componentes principales se asoció con el tipo de ensayo, y todas las muestras agruparon muestra y tipo de cáncer.

Ejemplo 3: Identificación de los efectos de lote técnicos derivados de diferencias en el lugar de recogida de muestras biológicas y del procedimiento de conservación

5 Los cánceres hemáticos pueden recogerse de sangre total, médula ósea y, a veces, conservarse en parafina fijada con formol (FFPE). Sin embargo, las diferencias en estas metodologías de recogida de muestras dan lugar a la introducción de efectos de lote técnicos. Brevemente, los datos de expresión de ARN de muestras de cáncer emparejadas por cohorte recogidas mediante muestreo de sangre, muestreo de médula ósea o conservadas en FFPE se analizaron mediante análisis de reducción de dimensiones, incrustados en 2 coordenadas utilizando UMAP, de acuerdo con algunas realizaciones de la divulgación. Los resultados presentados en la figura 7 ilustran que las muestras del transcriptoma están agrupadas y separadas por FFPE frente al tubo de sangre/médula ósea EDTA en el eje y, y por médula ósea frente a total en el eje. Esta separación se debe tanto a diferencias biológicas como técnicas y requiere un emparejamiento de referencia adecuado.

15 *Ejemplo 4: Identificación de los efectos de lote técnicos derivados de diferencias en la metodología de extracción de ARN*

Los diferentes procedimientos de extracción y los diferentes productos químicos utilizados durante el análisis del transcriptoma pueden introducir un efecto de lote. Por ejemplo, los efectos de lote pueden surgir cuando las muestras de ARN se extraen antes de que la muestra se envíe para la secuenciación (extracción externa), por ejemplo, por un médico, en lugar de justo antes del análisis de RNA-Seq (extracción interna). Brevemente, los datos de expresión de ARN de muestras de cáncer emparejadas por cohorte que se obtuvieron antes o después del aislamiento de ARN se analizaron mediante análisis de reducción de dimensiones, incrustados en 2 coordenadas utilizando UMAP, de acuerdo con algunas realizaciones de la divulgación. Los resultados presentados en la figura 8 ilustran que las muestras extraídas de sangre por una fuente externa se agrupan por separado de las muestras extraídas internamente. Un segundo efecto de lote se observa mediante la separación de muestras FFPE extraídas internamente de sangre y médula ósea, como se describe en el ejemplo 3.

Ejemplo 5: Identificación de los efectos de lote técnicos derivados de diferentes lotes de reactivos

Las metodologías de RNA-Seq de captura incluyen una etapa en la que un banco de fragmentos de ADNc se enriquece utilizando sondas de captura. En un ejemplo de esto, dos muestras de bancos de fragmentos de ADNc preparados para varias muestras de cáncer se hibridaron con dos lotes del mismo conjunto de sondas oligonucleotídicas biotiniladas que son complementarias con regiones genómicas de interés. A veces, estos bancos de sondas de captura pueden ser a su vez un conjunto de diferentes diseños de captura genómica. La fabricación de lotes de sondas y la agrupación de bancos de captura pueden introducir efectos de lote. A continuación, los datos de expresión de ARN generados utilizando ambos lotes de sondas de captura se analizaron mediante un análisis de reducción de dimensiones PCA. Los resultados presentados en la figura 9 ilustran un efecto de lote introducido por los diferentes lotes de sondas, detectado en PC8 (eje x).

Ejemplo 6: Análisis de los efectos de lote técnicos derivados de la diferente plexidad de la captura de hibridación

La plexidad de hibridación se refiere al número de muestras de ADNc agrupadas durante la captura selectiva. Los ensayos pueden abarcar desde una sola muestra en una agrupación hasta más de una docena. En este experimento, se secuenciaron 9 muestras tumorales y 2 controles celulares en 3 condiciones de plexidad (grupos de una sola muestra, 3x y 6x). Tras la secuenciación del ARN, los datos de expresión generados utilizando muestras preparadas en las diferentes condiciones de plexidad se analizaron mediante un análisis de reducción de dimensiones, incrustado en 2 coordenadas utilizando UMAP, de acuerdo con algunas realizaciones de la divulgación. Los resultados presentados en la figura 10 ilustran que las muestras emparejadas se agruparon, independientemente de las condiciones de plexidad utilizadas, lo que indica que la plexidad no introduce ningún efecto de lote en el análisis del transcriptoma.

Ejemplo 7: Análisis de los efectos de lote técnicos derivados de diferentes números de ciclos de amplificación por PCR

En algunas metodologías de RNA-Seq, la PCR posterior a la captura es una etapa de amplificación después de que los fragmentos de amplicón hayan sido capturados por sondas, y antes de la secuenciación. Los fragmentos no unidos se lavan y los fragmentos restantes se amplifican durante un número determinado de ciclos (a mayor número de ciclos, mayor amplificación). Demasiados ciclos pueden dar lugar a una tasa de duplicación desequilibrada en función de las características de las secuencias. En este experimento, se determinó el efecto del número de ciclos de amplificación (entre 7-9) en 6 tumores y 1 muestra de control. Tras la secuenciación del ARN, los datos de expresión generados utilizando muestras preparadas en las diferentes condiciones de amplificación (7-9 ciclos) se analizaron mediante un análisis de reducción de dimensiones, incrustado en 2 coordenadas utilizando UMAP, de acuerdo con algunas realizaciones de la divulgación. Los resultados presentados en la figura 11 ilustran que las muestras emparejadas se agruparon, independientemente del número de ciclos de amplificación utilizados, lo que indica que la variación en el número de ciclos de amplificación PCR entre 7-9 no introduce ningún efecto de lote en el análisis del transcriptoma.

Ejemplo 8: Análisis de los efectos de lote técnicos derivados de diferentes molaridades de carga del secuenciador

La molaridad de carga se refiere a la cantidad de muestra que se carga en el secuenciador. Normalmente, una molaridad demasiado baja puede provocar una mayor tasa de duplicación y datos más ruidosos. En este experimento del efecto de lote, se secuenciaron 11 muestras tumorales y 3 muestras de control en 3 condiciones de molaridad (0,7, 1 y 1,5 uM). Tras la secuenciación del ARN, los datos de expresión generados utilizando muestras preparadas con las diferentes molaridades de carga se analizaron mediante análisis de reducción de dimensiones, incrustados en 2 coordenadas utilizando UMAP, de acuerdo con algunas realizaciones de la divulgación. Los resultados presentados en la figura 12 ilustran que las muestras emparejadas se agruparon, independientemente de la molaridad de carga utilizada, lo que indica que la variación de la molaridad de carga entre 0,7 y 1,5 uM no introduce ningún efecto de lote en el análisis del transcriptoma.

Ejemplo 9: Análisis de los efectos de lote técnicos derivados de los cambios en la química de los reactivos de secuenciación

El cambio en los reactivos de secuenciación de Illumina es un cambio habitual y propio de la marca en los productos químicos utilizados con el fin de ajustarse mejor a las características adicionales a su tecnología. Un ejemplo reciente de cambio de reactivos se adaptó para permitir lecturas adicionales en los índices moleculares universales ("universal molecular indices", UMI). En el experimento de control del efecto de lote, se secuenciaron 28 muestras con dos versiones de reactivo para verificar que no se detectaba ningún efecto de lote entre las versiones anterior y actual de los reactivos. Tras la secuenciación del ARN, los datos de expresión generados a partir de las muestras preparadas con las diferentes versiones de los reactivos se analizaron mediante un análisis de reducción de dimensiones, incrustado en 2 coordenadas utilizando UMAP, de acuerdo con algunas realizaciones de la divulgación. En general, las muestras se agrupan por muestra (figura 13; línea de conexión) y los reactivos tienen una influencia pequeña, pero dentro de un intervalo aceptable, en la varianza del transcriptoma.

Conclusión

Los procedimientos descritos en el presente documento proporcionan procedimientos de control de calidad mejorados para evaluar lotes de muestras de secuenciación de ARN. Con una precisión mejorada y una resolución más alta que los procedimientos anteriores, los algoritmos predictivos proporcionados en el presente documento pueden utilizarse para identificar muestras individuales y lotes enteros que cumplan las normas de control de calidad. Con este mayor control de calidad, la información utilizada para proporcionar diagnósticos a los pacientes y determinar los tratamientos adecuados será más precisa, lo que permitirá mejorar los diagnósticos y recomendar tratamientos más informados a los pacientes.

Referencias citadas y realizaciones alternativas

La presente invención puede aplicarse como un producto de programa de ordenador que comprende un mecanismo de programa de ordenador incrustado en un medio de almacenamiento legible por ordenador no transitorio. Por ejemplo, el producto de programa de ordenador podría contener los módulos de programa mostrados en la figura 1 y/o como se describe en las figuras 2A y 2B. Estos módulos de programa pueden almacenarse en un CD-ROM, un DVD, un producto de almacenamiento en disco magnético, una llave USB o cualquier otro producto de almacenamiento no transitorio de datos o programas legibles por ordenador.

REIVINDICACIONES

1. Un procedimiento de identificación de los efectos de lote en una o más muestras de secuenciación de ARN, comprendiendo dicho procedimiento:

- 5 en un sistema informático que tiene uno o más procesadores, y una memoria que almacena uno o más programas para su ejecución por dichos uno o más procesadores:
- a) obtener, en formato electrónico, un conjunto de datos de lotes que comprende, para cada muestra de ensayo respectiva de un lote de muestras de ensayo,
- 10 un perfil de expresión génica correspondiente que comprende un valor de expresión génica correspondiente para cada gen respectivo de un primer conjunto de genes, y
- un conjunto de metadatos correspondiente que comprende valores para cada característica respectiva en un primer conjunto de características sobre la muestra de ensayo respectiva;
- en el que el lote de muestras de ensayo comprende al menos 10 muestras de ensayo;
- b) determinar, para el conjunto de datos de lotes, un conjunto de datos de referencia emparejado por cohorte seleccionando, de una base de datos de referencia que comprende una pluralidad de muestras de la base de datos de referencia, un subconjunto de las muestras de la base de datos de referencia, identificando así
- 15 una pluralidad de muestras de referencia emparejadas por cohorte, en el que:
- el conjunto de datos de referencia emparejado por cohorte comprende, para cada muestra de referencia emparejada por cohorte respectiva en la pluralidad de muestras de referencia emparejadas por cohorte,
- 20 un perfil de expresión génica correspondiente que comprende un valor de expresión génica correspondiente para cada gen respectivo en el primer conjunto de genes,
- en el que:
- cada muestra de referencia respectiva emparejada por cohorte en la pluralidad de muestras de referencia emparejadas por cohorte se asocia con un conjunto de metadatos correspondiente que comprende un valor correspondiente para cada característica respectiva en un segundo conjunto de características sobre la muestra de referencia respectiva, en el que el segundo conjunto de características comprende un tercer conjunto de una o más características que también están presentes en el primer conjunto de características, y
- 25 para cada característica respectiva en el tercer conjunto de una o más características, el valor agregado para la característica respectiva en el tercer conjunto de una o más características que están presentes tanto en el primer conjunto de características como en el segundo conjunto de características se equilibra entre el conjunto de datos de lotes y el conjunto de datos de referencia emparejado por cohorte, en el que el valor agregado de una característica respectiva se equilibra entre el conjunto de datos de lotes y el conjunto de datos de referencia emparejado por cohorte cuando el porcentaje de muestras de ensayo respectivas, en el lote de muestras de ensayo, que tienen un valor respectivo para la característica respectiva está dentro del 25 % del porcentaje de muestras de referencia respectivas, en la pluralidad de muestras de referencia, que tienen el mismo valor respectivo para la característica respectiva;
- 30 c) realizar una reducción de dimensiones en un conjunto de datos combinado que consiste en el perfil de expresión génica correspondiente para cada muestra de ensayo respectiva en la pluralidad de muestras de ensayo y el perfil de expresión génica correspondiente para cada muestra de referencia respectiva emparejada por cohorte en la pluralidad de muestras de referencia emparejadas por cohorte, en el que la reducción de dimensiones es un análisis de componentes principales, un análisis de componentes latentes, una regresión de mínimos cuadrados parciales, o una aproximación y proyección uniforme de la variedad, obteniendo de este modo:
- 40 para cada muestra de ensayo respectiva y cada muestra de referencia respectiva, un primer conjunto correspondiente de coordenadas incrustadas en un espacio dimensional inferior a las dimensiones del perfil de expresión génica correspondiente, y
- para cada muestra de referencia respectiva emparejada por cohorte, un segundo conjunto correspondiente de coordenadas incrustadas en un espacio dimensional inferior a las dimensiones del perfil de expresión génica correspondiente en el conjunto de datos de referencia emparejado por cohorte;
- 50

- d) determinar una medida estadística de similitud entre el primer conjunto de coordenadas correspondiente y los segundos conjuntos de coordenadas correspondientes obtenidos para las muestras de ensayo y los conjuntos de coordenadas obtenidos para las muestras de referencia; y
- 5 e) comparar la medida estadística de similitud con un valor umbral para identificar efectos de lote en las muestras de secuenciación de ARN por medio de:
- validar el conjunto de datos de lotes para emitir una notificación cuando la medida estadística de similitud satisface el valor umbral, o
- no validar el conjunto de datos de lotes para emitir una notificación cuando la medida estadística de similitud no satisface el valor umbral.
- 10 2. El procedimiento de la reivindicación 1 comprende además, cuando la medida estadística de similitud no satisface el valor umbral, marcar el conjunto de datos de lotes para su posterior revisión.
3. El procedimiento de la reivindicación 1 o 2, en el que la obtención del conjunto de datos de lotes comprende, para cada muestra de ensayo respectiva del lote de muestras de ensayo:
- 15 obtener, en formato electrónico, una pluralidad correspondiente de lecturas de secuencia derivadas de la muestra de ensayo respectiva mediante secuenciación de ARN selectiva o del transcriptoma completo; y
- determinar, a partir de la pluralidad correspondiente de lecturas de secuencia, el valor de expresión génica correspondiente para cada gen respectivo del primer conjunto de genes.
4. El procedimiento de la reivindicación 3, en el que para cada muestra de ensayo respectiva del lote de muestras de ensayo, la pluralidad correspondiente de lecturas de secuencia comprende al menos 10 000 lecturas de secuencia.
- 20 5. El procedimiento de una cualquiera de las reivindicaciones 1 a 4, en el que el primer conjunto de genes comprende al menos 10 genes.
6. El procedimiento de una cualquiera de las reivindicaciones 1 a 5, en el que el primer conjunto de genes comprende al menos 10.000 genes.
- 25 7. El procedimiento de una cualquiera de las reivindicaciones 1a 6, en el que el tercer conjunto de una o más características comprende una característica seleccionada del grupo que consiste en el emplazamiento del tejido, la pureza del tumor, el tipo de cáncer, la identidad del secuenciador y la fecha de secuenciación.
8. El procedimiento de una cualquiera de las reivindicaciones 1 a 7, en el que el tercer conjunto de una o más características comprende una característica seleccionada entre un procedimiento de extracción de ácido nucleico, un procedimiento de preparación de bancos de ADNc, un procedimiento de secuenciación de ARN, un tipo de reactivo utilizado y un tipo de equipo utilizado.
- 30 9. El procedimiento de una cualquiera de las reivindicaciones 1 a 8, en el que la pluralidad de muestras de referencia comprende al menos 100 muestras de referencia o la pluralidad de muestras de referencia comprende al menos 1000 muestras de referencia.
- 35 10. El procedimiento de una cualquiera de las reivindicaciones 1 a 9, en el que el valor agregado para una característica respectiva está equilibrado entre el conjunto de datos de lotes y el conjunto de referencia emparejado por cohorte cuando el porcentaje de muestras de ensayo respectivas, en el lote de muestras de ensayo, que tienen un valor respectivo para la característica respectiva, está dentro del 5 % del porcentaje de muestras de referencia respectivas, en la pluralidad de muestras de referencia, que tienen el mismo valor respectivo para la característica respectiva.
- 40 11. El procedimiento de una cualquiera de las reivindicaciones 1 a 10, en el que la reducción de dimensiones comprende incrustar, para cada muestra de ensayo respectiva y cada muestra de referencia respectiva, el perfil de expresión génica correspondiente en una representación bidimensional.
- 45 12. El procedimiento de una cualquiera de las reivindicaciones 1 a 10, en el que, para cada muestra de ensayo respectiva de un lote de muestras de ensayo, el perfil de expresión génica correspondiente se determina a partir de lecturas de secuencia derivadas de la muestra respectiva mediante secuenciación de ARN selectiva o del transcriptoma completo.
13. El procedimiento de la reivindicación 12, en el que:
- la secuenciación de ARN selectiva utiliza una pluralidad de sondas;
- 50 cada sonda de la pluralidad de sondas se dirige de forma única a una porción respectiva de un transcriptoma de referencia, y

cada lectura de secuencia en la correspondiente pluralidad de lecturas de secuencia corresponde al menos a una sonda de la pluralidad de sondas.

14. El procedimiento de la reivindicación 3 o 4, que comprende además:

5 realizar, para cada muestra de ensayo respectiva del lote de muestras de ensayo, una o más pruebas de control de calidad de una sola muestra en la muestra de ensayo respectiva; y

10 eliminar las muestras de ensayo respectivas del lote de muestras de ensayo que no superen alguna de dichas una o más pruebas de control de calidad de una sola muestra o marcar para inspección manual las muestras de ensayo respectivas que no superen alguna de dichas una o más pruebas de control de calidad de una sola muestra, en el que una prueba de control de calidad de una sola muestra respectiva en dichas una o más pruebas de control de calidad de una sola muestra comprende:

15 1) i) determinar, para una muestra de ensayo respectiva del lote de muestras de ensayo, un número respectivo de lecturas de secuencia cartografiadas no duplicadas en la pluralidad correspondiente de lecturas de secuencia, en el que cada lectura de secuencia cartografiada no duplicada se cartografía a una porción correspondiente de un genoma de referencia; y ii) comparar el número respectivo de lecturas de secuencia cartografiadas no duplicadas con un número esperado de lecturas de secuencia cartografiadas no duplicadas, en el que, cuando el número respectivo de lecturas de secuencia cartografiadas no duplicadas es inferior a un número predeterminado de lecturas cartografiadas no duplicadas, la muestra respectiva no supera la prueba de control de calidad de una sola muestra respectiva,

20 2) determinar, para una muestra de ensayo respectiva del lote de muestras de ensayo, un porcentaje respectivo de lecturas de secuencia correctamente emparejadas en la pluralidad correspondiente de lecturas de secuencia, en el que, cuando el porcentaje de lecturas de secuencia correctamente emparejadas es inferior a un umbral predeterminado de lecturas emparejadas, la muestra de ensayo respectiva no supera la prueba de control de calidad de una sola muestra respectiva,

25 3) determinar, para una muestra de ensayo respectiva del lote de muestras de ensayo, un contenido de GC respectivo de la pluralidad correspondiente de lecturas de secuencia, en el que cuando el contenido de GC respectivo está fuera de un umbral de contenido de GC predeterminado, la muestra de ensayo respectiva no supera la prueba de control de calidad de una sola muestra respectiva, o

30 4) determinar, para una muestra de ensayo respectiva del lote de muestras, un número respectivo de genes expresados utilizando la pluralidad correspondiente de lecturas de secuencia, en el que, cuando la puntuación correspondiente de lecturas expresadas es inferior a un número predeterminado de lecturas expresadas, la muestra de ensayo respectiva no supera la prueba de control de calidad de una sola muestra respectiva.

35 15. El procedimiento de una cualquiera de las reivindicaciones 1 a 14, en el que el procedimiento comprende además proporcionar, para cada muestra de ensayo del lote de muestras de ensayo, un informe de muestra respectivo, en el que cada informe de muestra respectivo comprende al menos uno de un conjunto de llamadas de expresión, una o más terapias emparejadas, o uno o más ensayos clínicos emparejados.

40 16. Un medio de almacenamiento no transitorio legible por ordenador que almacena al menos un programa configurado para ser ejecutado por un ordenador, comprendiendo dicho al menos un programa instrucciones para realizar un procedimiento según una cualquiera de las reivindicaciones 1 a 15.

17. Un sistema informático, que comprende:

al menos un procesador;

una memoria que almacena al menos un programa para ser ejecutado por dicho al menos un procesador;

45 comprendiendo dicho al menos un programa instrucciones para realizar un procedimiento según una cualquiera de las reivindicaciones 1 a 15.

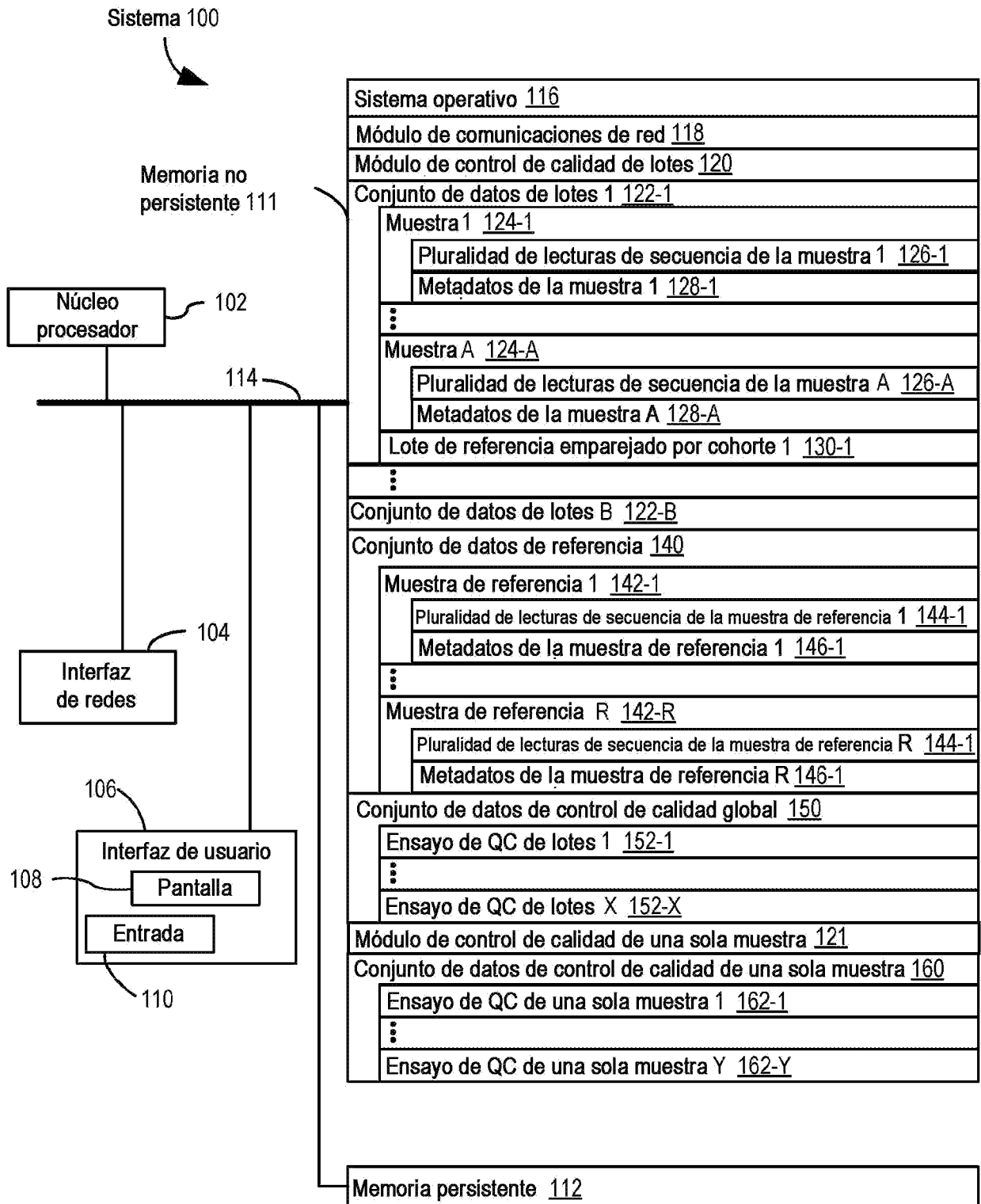


FIG. 1

202

206 Obtener, en formato electrónico, un conjunto de datos de lotes que comprende, para cada muestra respectiva en un lote de muestras, una pluralidad correspondiente de lecturas de secuencia derivadas de la muestra respectiva por secuenciación de ARN selectiva o del transcriptoma completo y los metadatos correspondientes para la muestra respectiva.



208 Determinar, para el conjunto de datos de lotes, un lote de referencia emparejado por cohorte. El lote de referencia emparejado por cohorte se equilibra para el emplazamiento del tejido, la pureza del tumor, el tipo de cáncer, la identidad del secuenciador o la fecha de secuenciación.

210 Determinar un conjunto de datos de referencia emparejado por cohorte para el conjunto de datos de lotes que comprende extraer, para cada muestra en el lote de muestras: i) una pluralidad respectiva de características de secuencia a partir de la pluralidad respectiva de lecturas de secuencia, obteniendo así una pluralidad de características de secuencia del lote, y ii) una pluralidad respectiva de características de metadatos de la muestra, obteniendo así una pluralidad de características de metadatos del lote; y seleccionar, a partir de un conjunto de datos de referencia, basándose, al menos en parte, en la pluralidad de características de secuencia del lote o en la pluralidad de características de metadatos del lote, el conjunto de datos de referencia emparejado por cohorte que comprende una pluralidad de muestras de referencia.

212 Determinar una combinación lineal o no lineal de la pluralidad de características de secuencia del lote y la pluralidad de características de metadatos del lote, sometiendo la pluralidad de características de secuencia del lote y la pluralidad de características de metadatos del lote a una técnica de reducción de dimensiones.



FIG. 2A

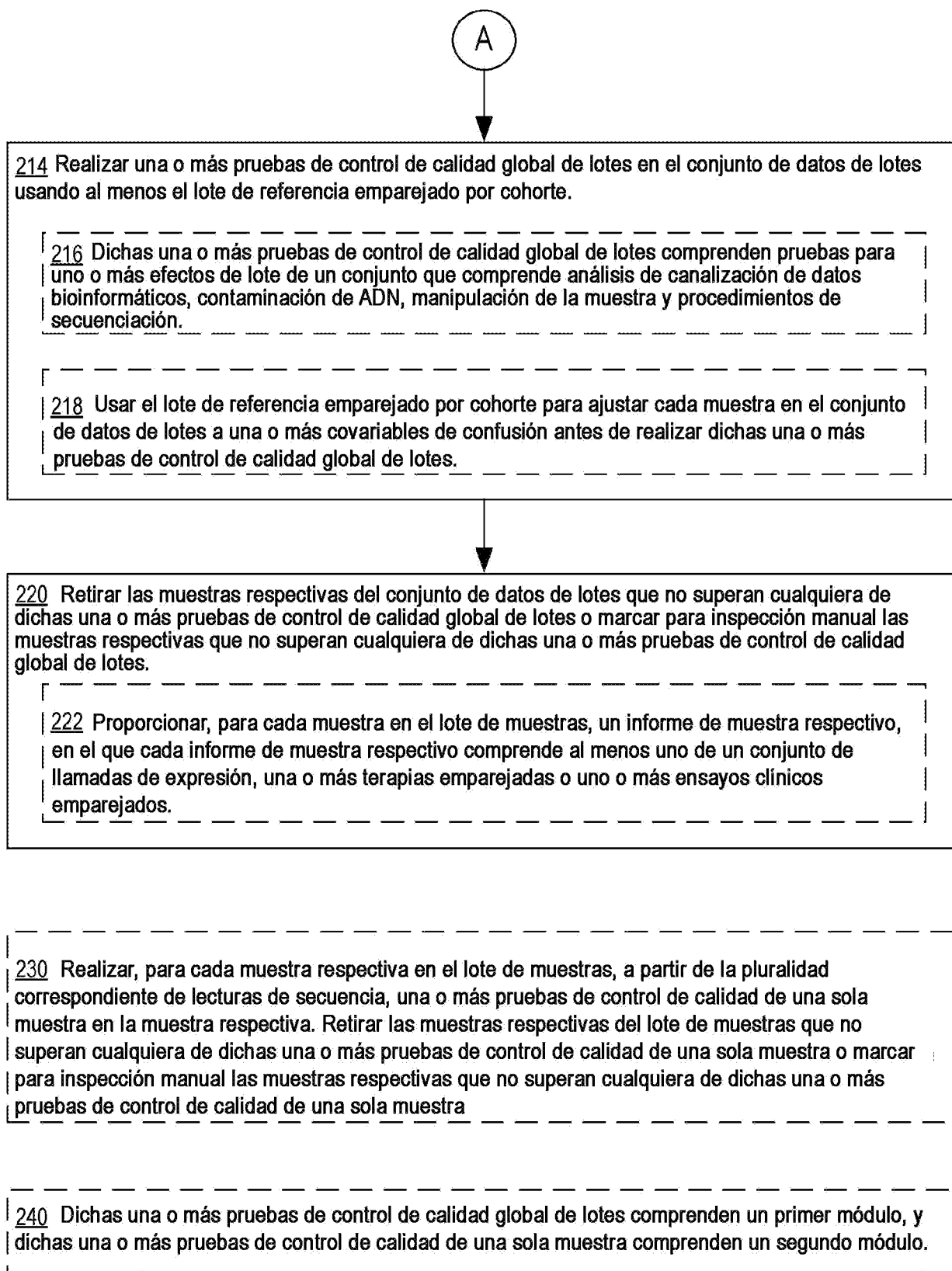


FIG. 2B

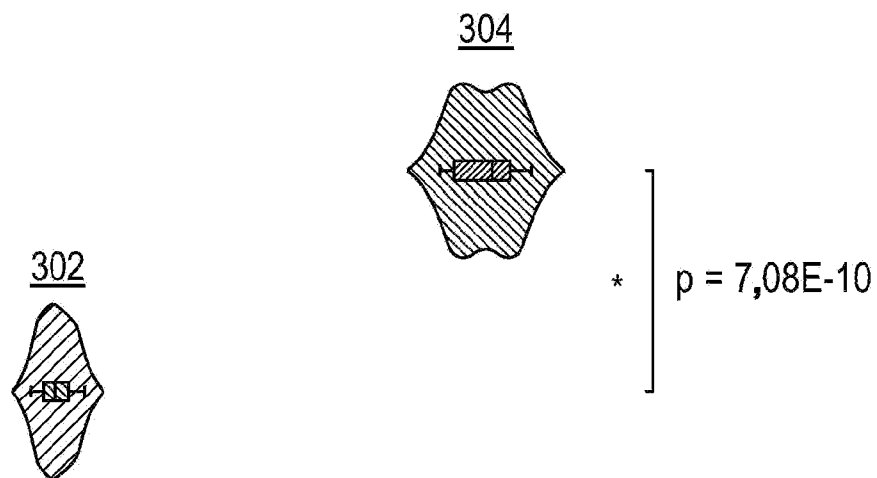
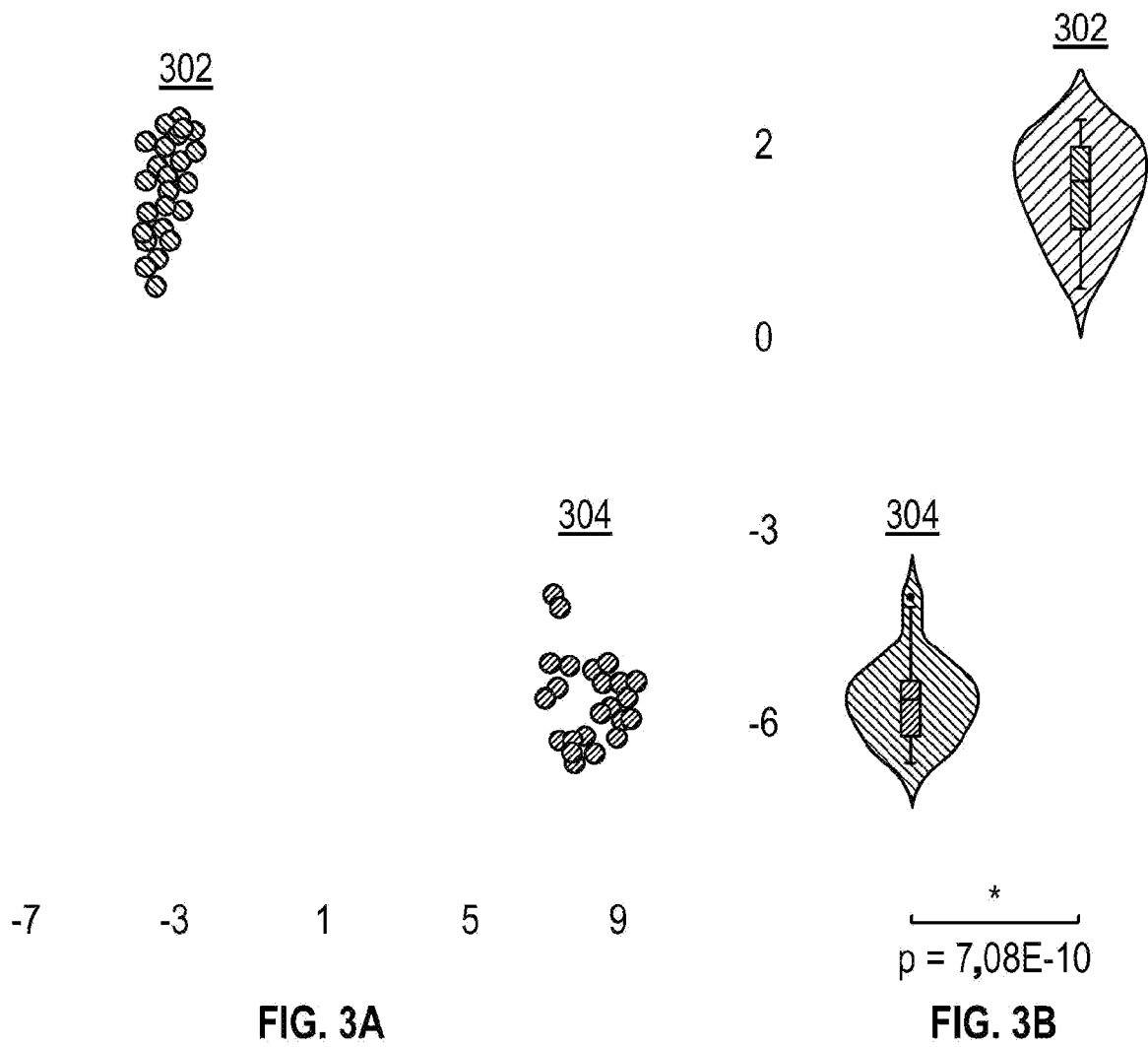


FIG. 3C

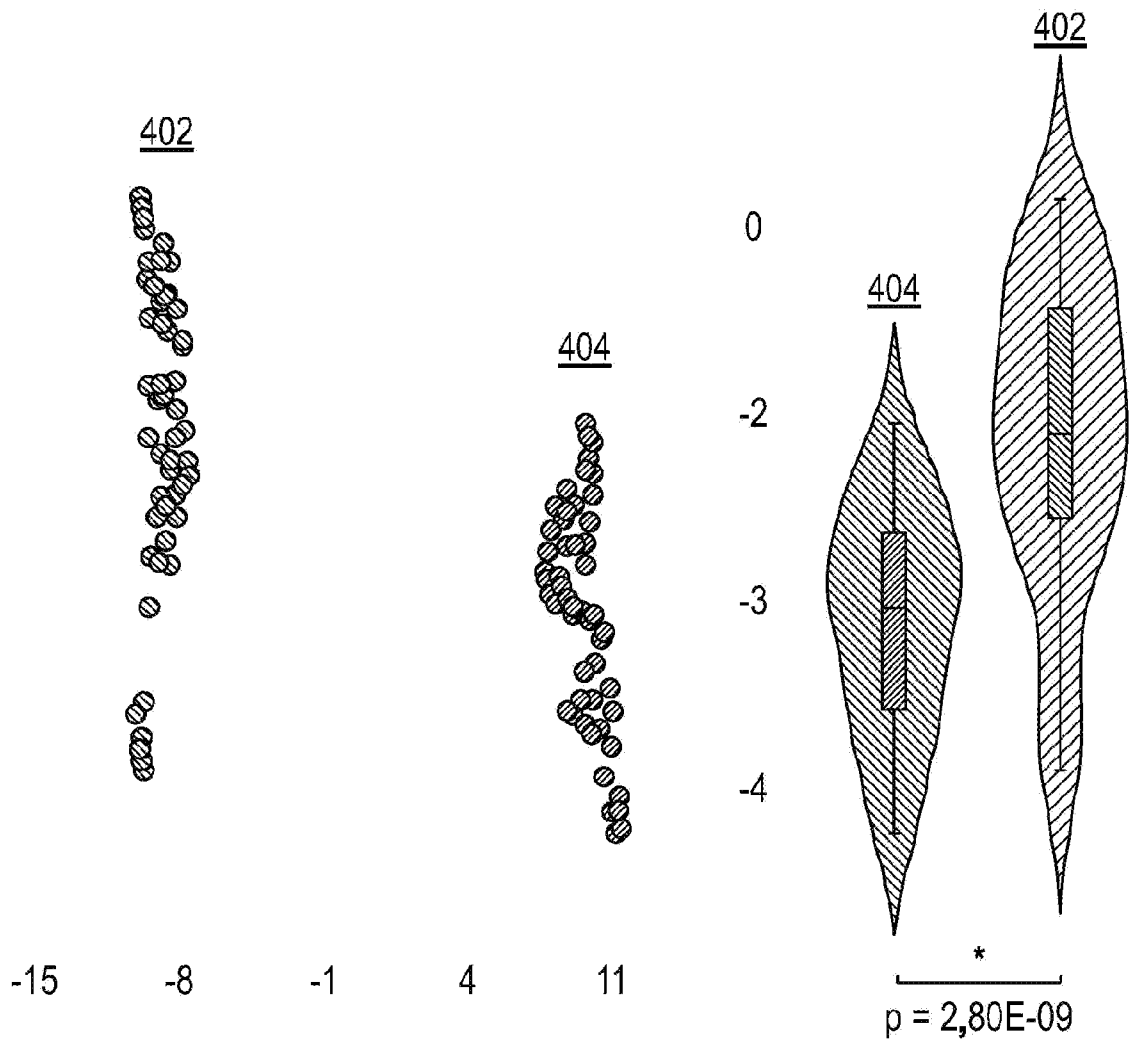


FIG. 4A

FIG. 4B

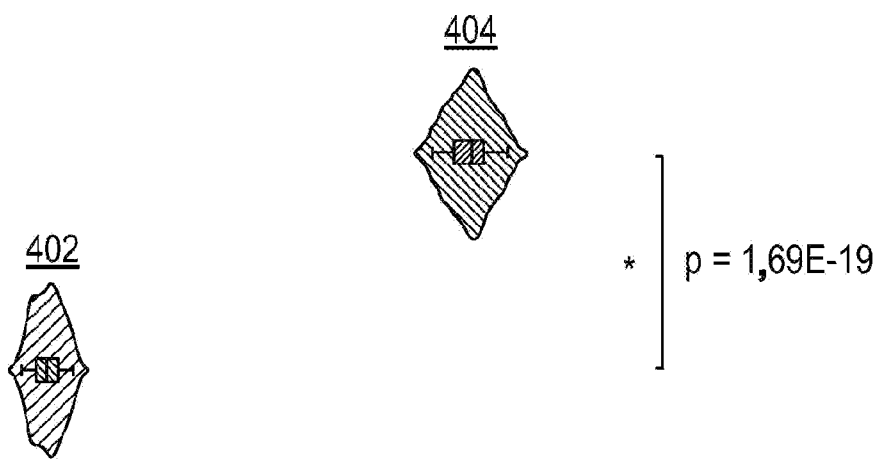


FIG. 4C

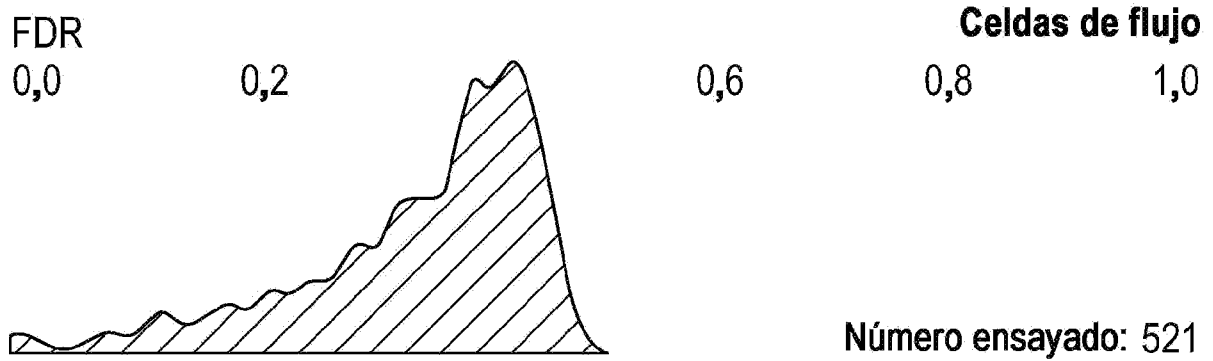


FIG. 5A

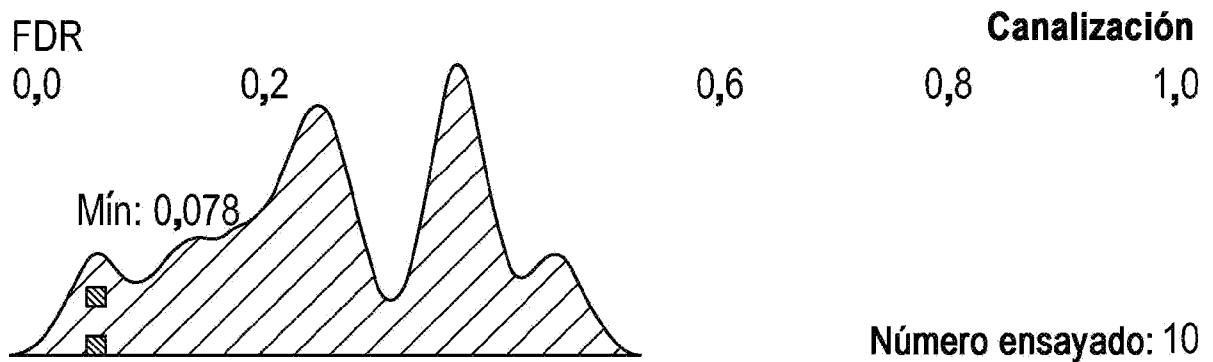


FIG. 5B

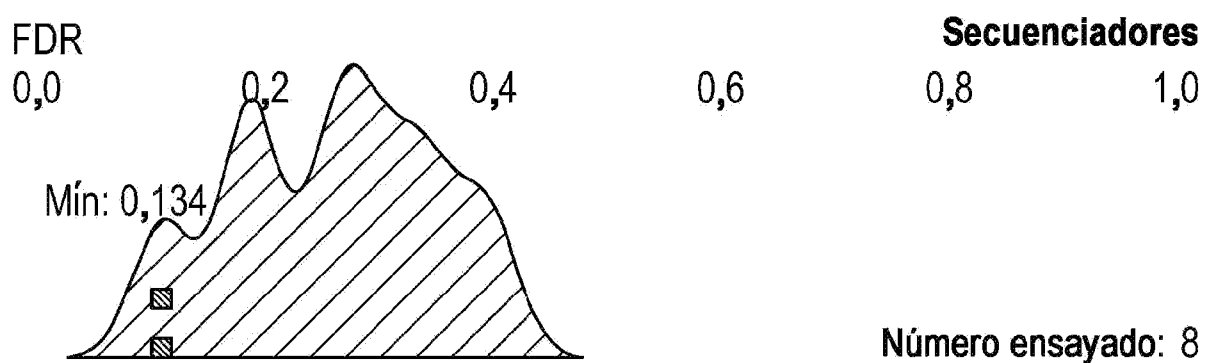


FIG. 5C

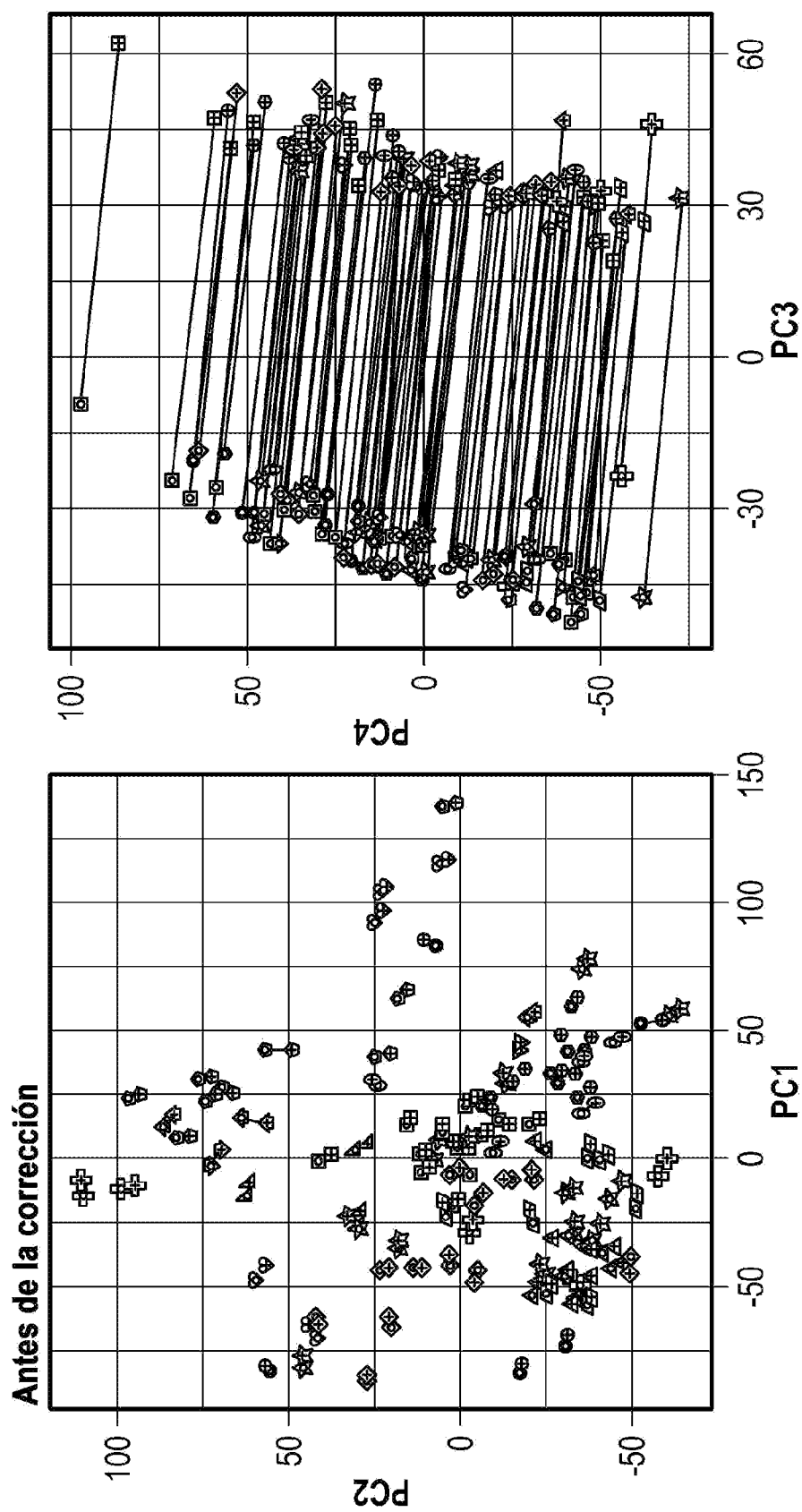


FIG. 6A

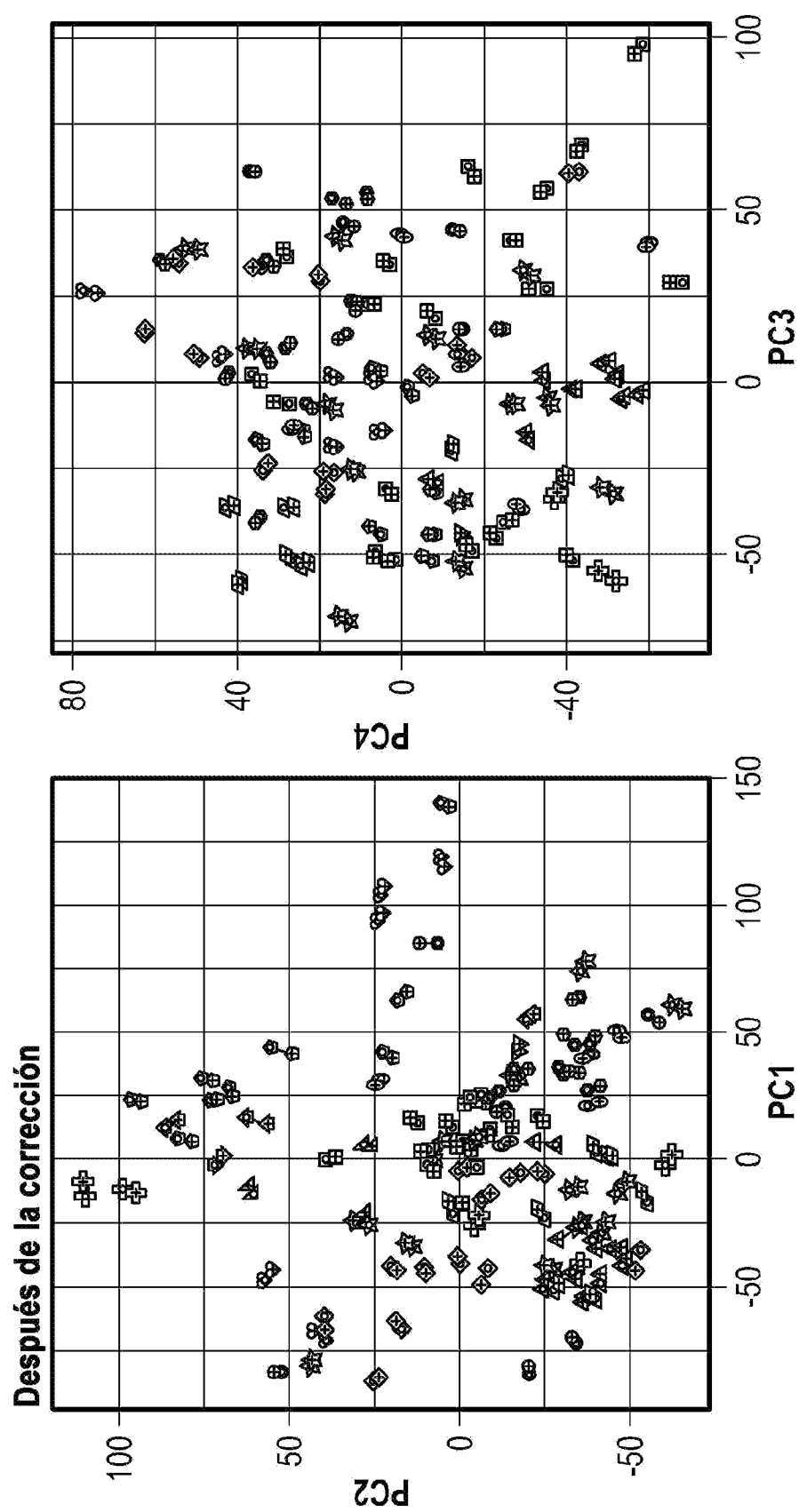


FIG. 6B

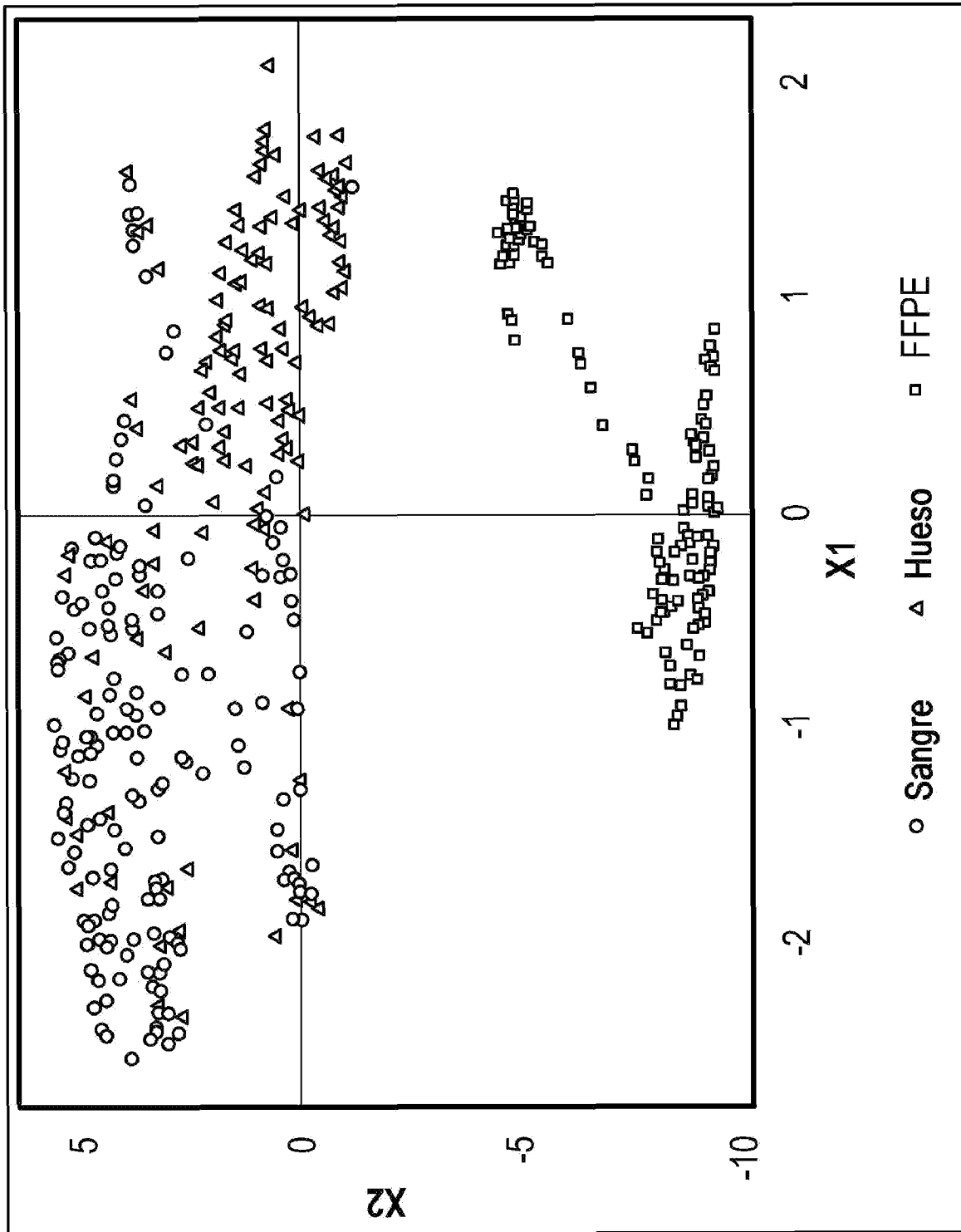


FIG. 7

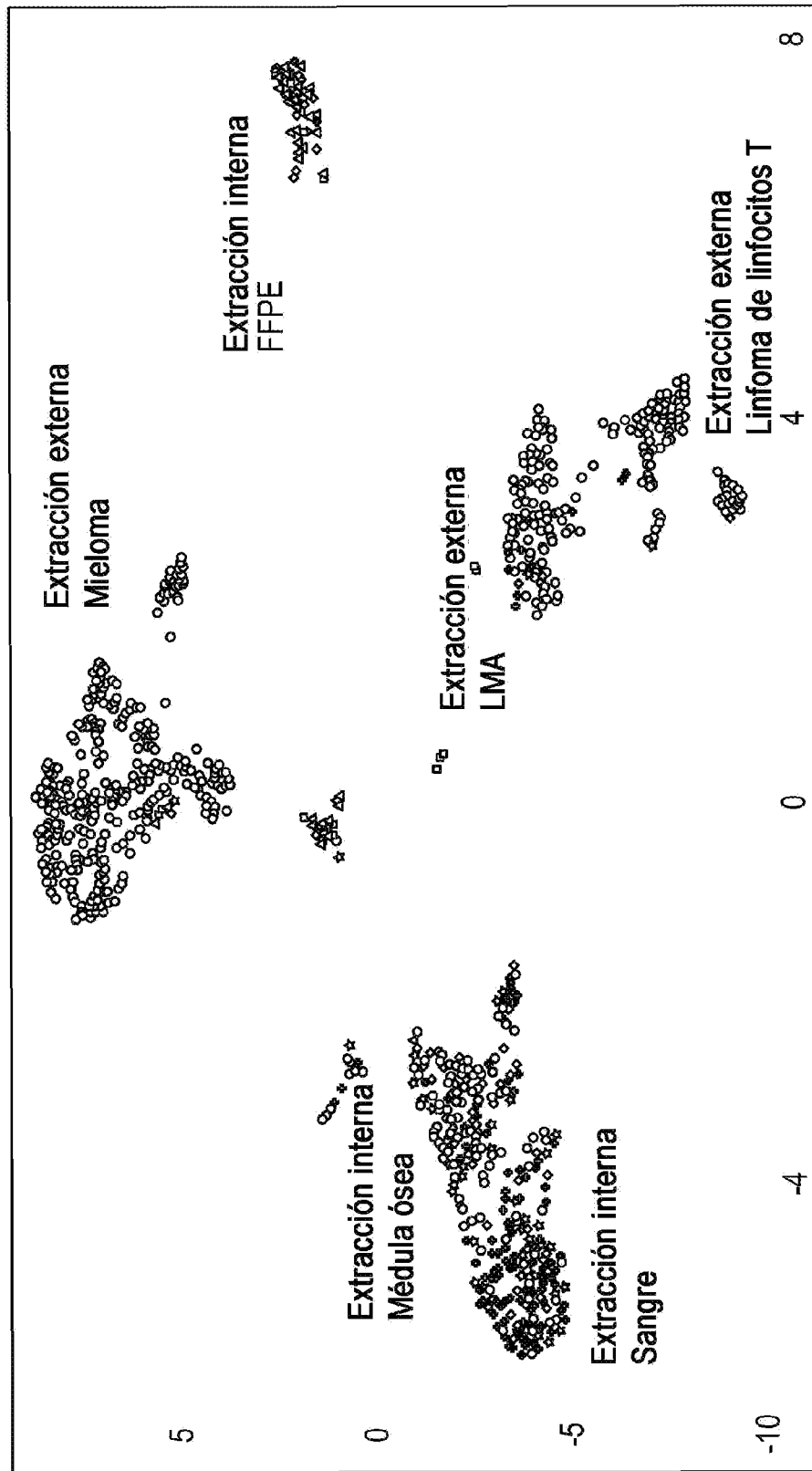


FIG. 8

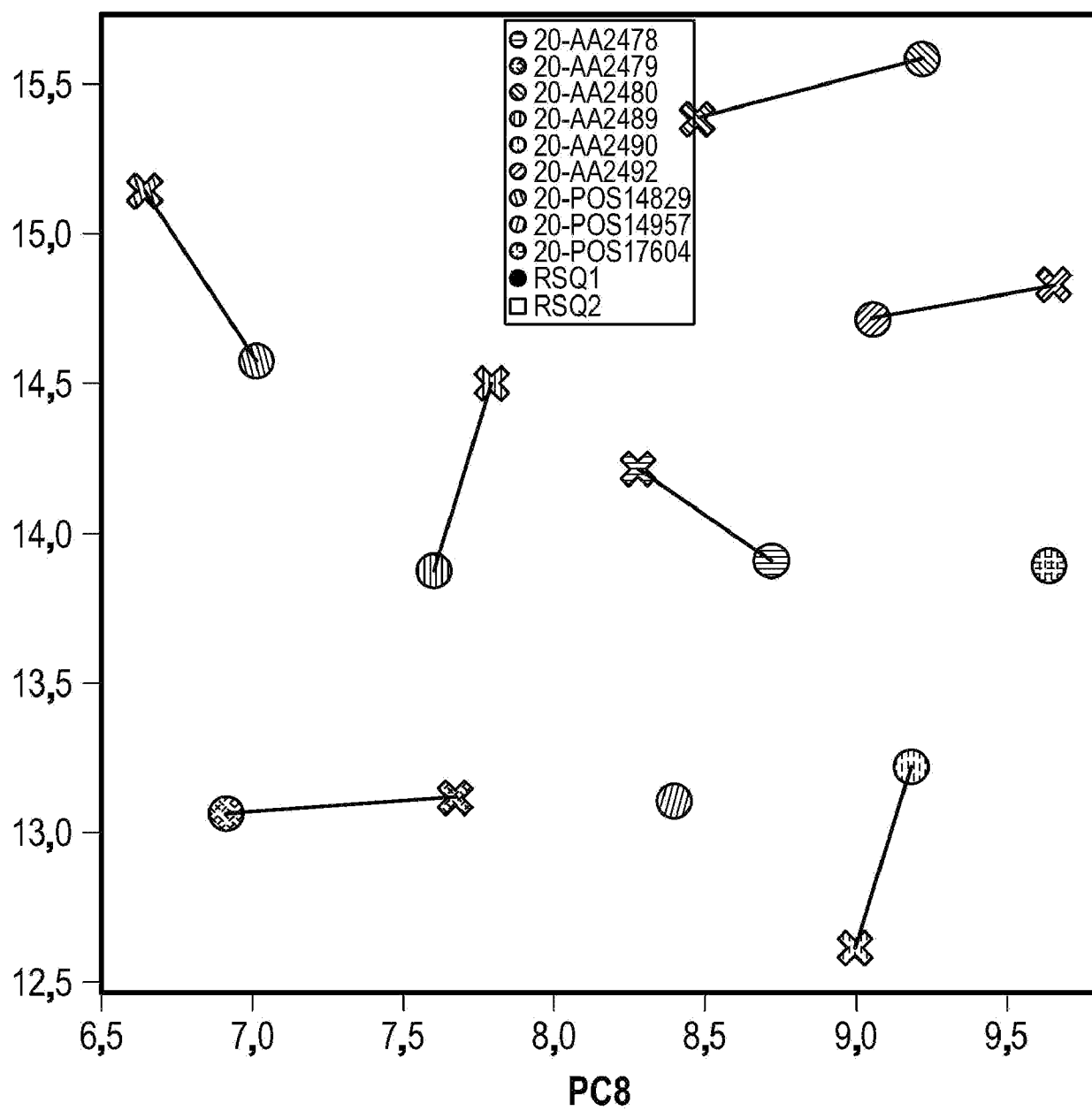


FIG. 9

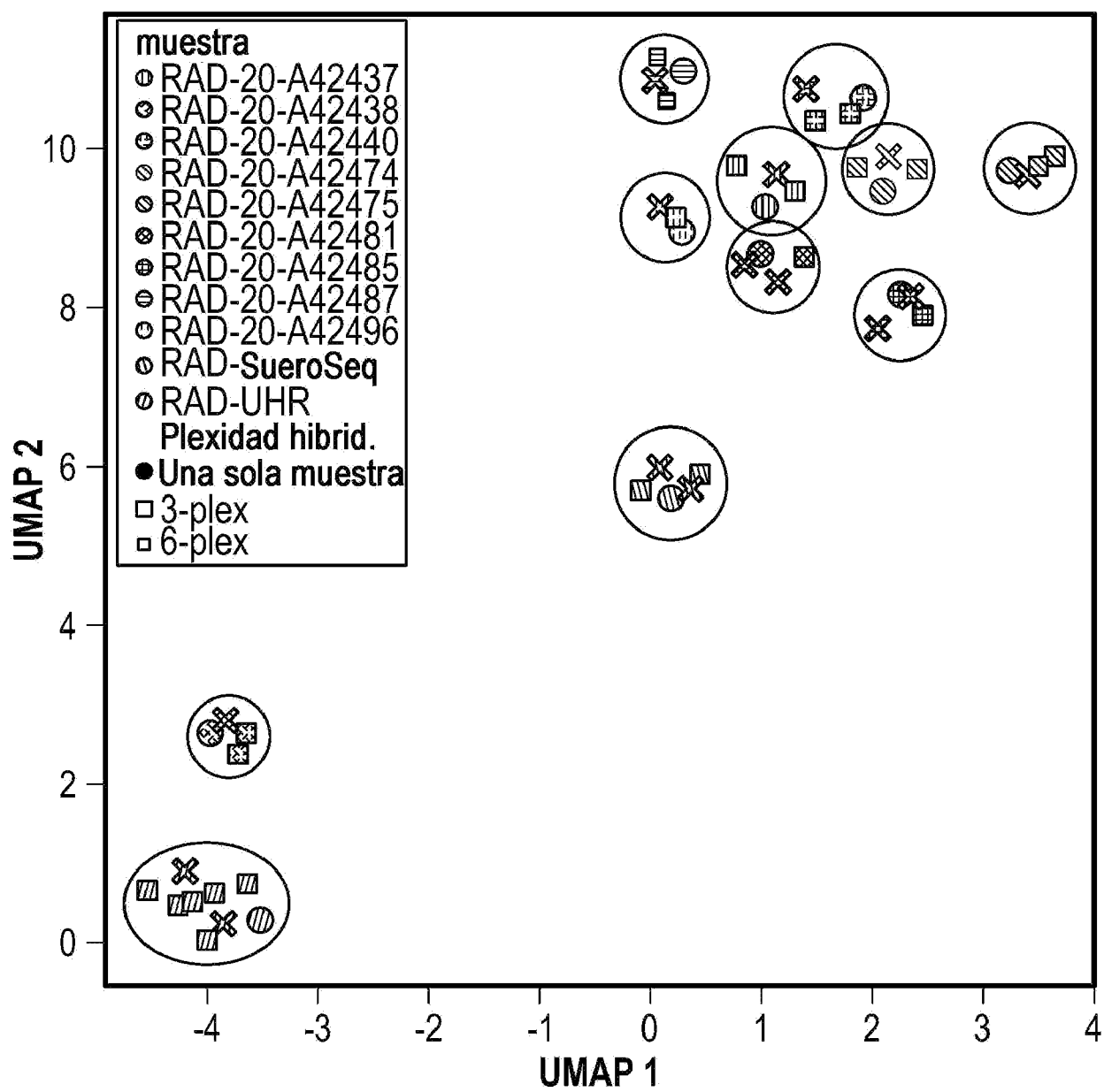


FIG. 10

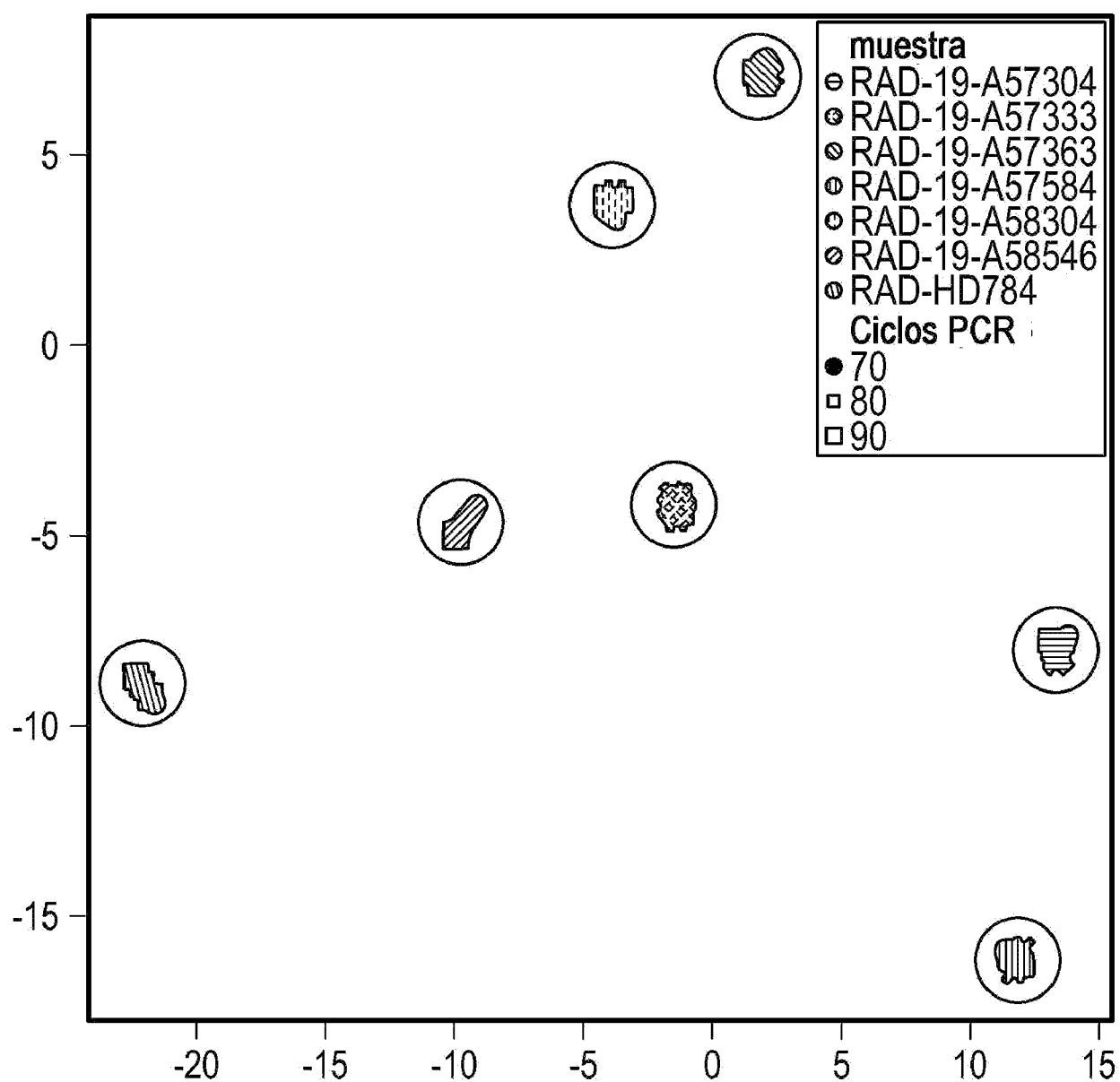


FIG. 11

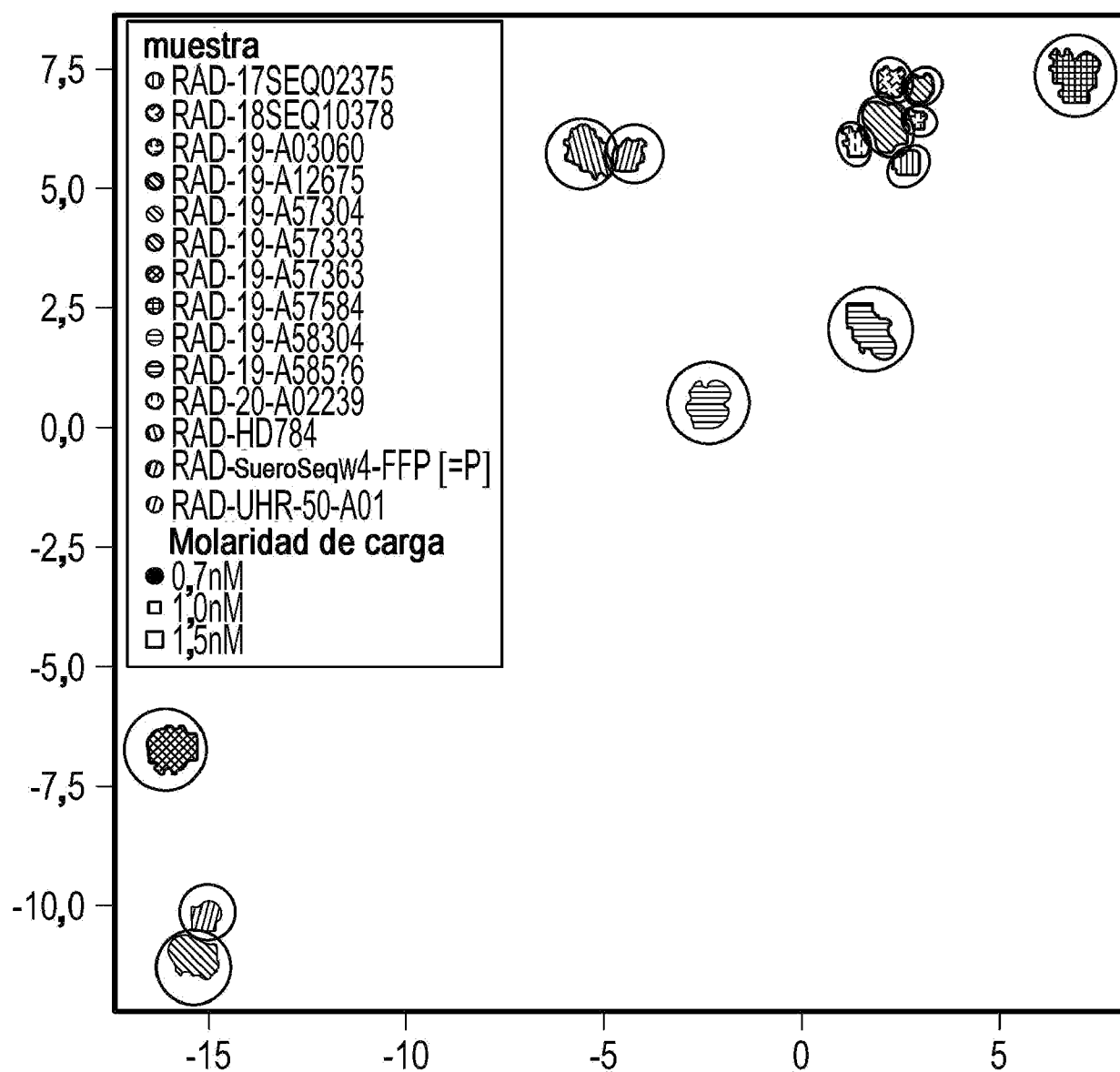


FIG. 12

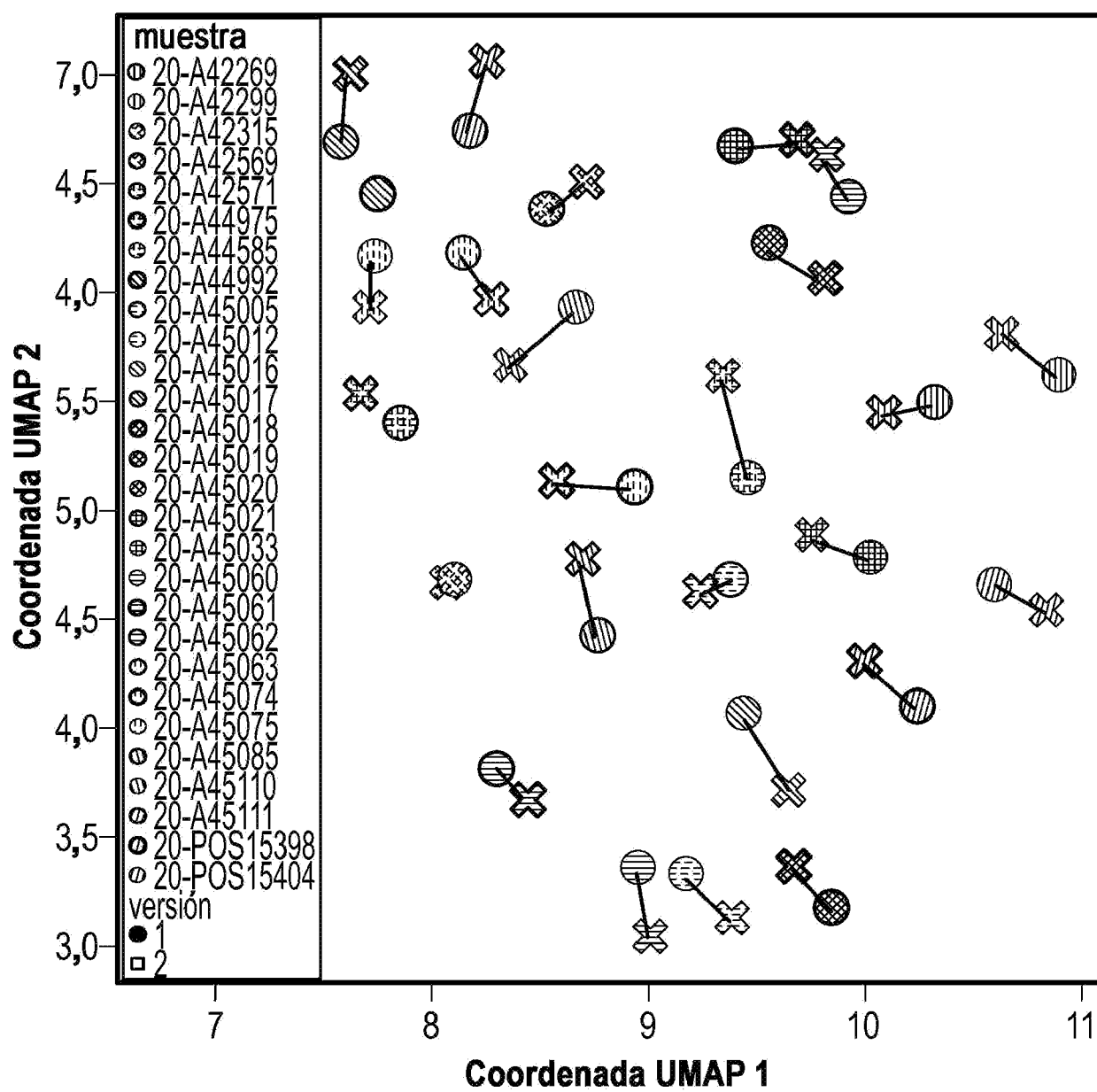


FIG. 13

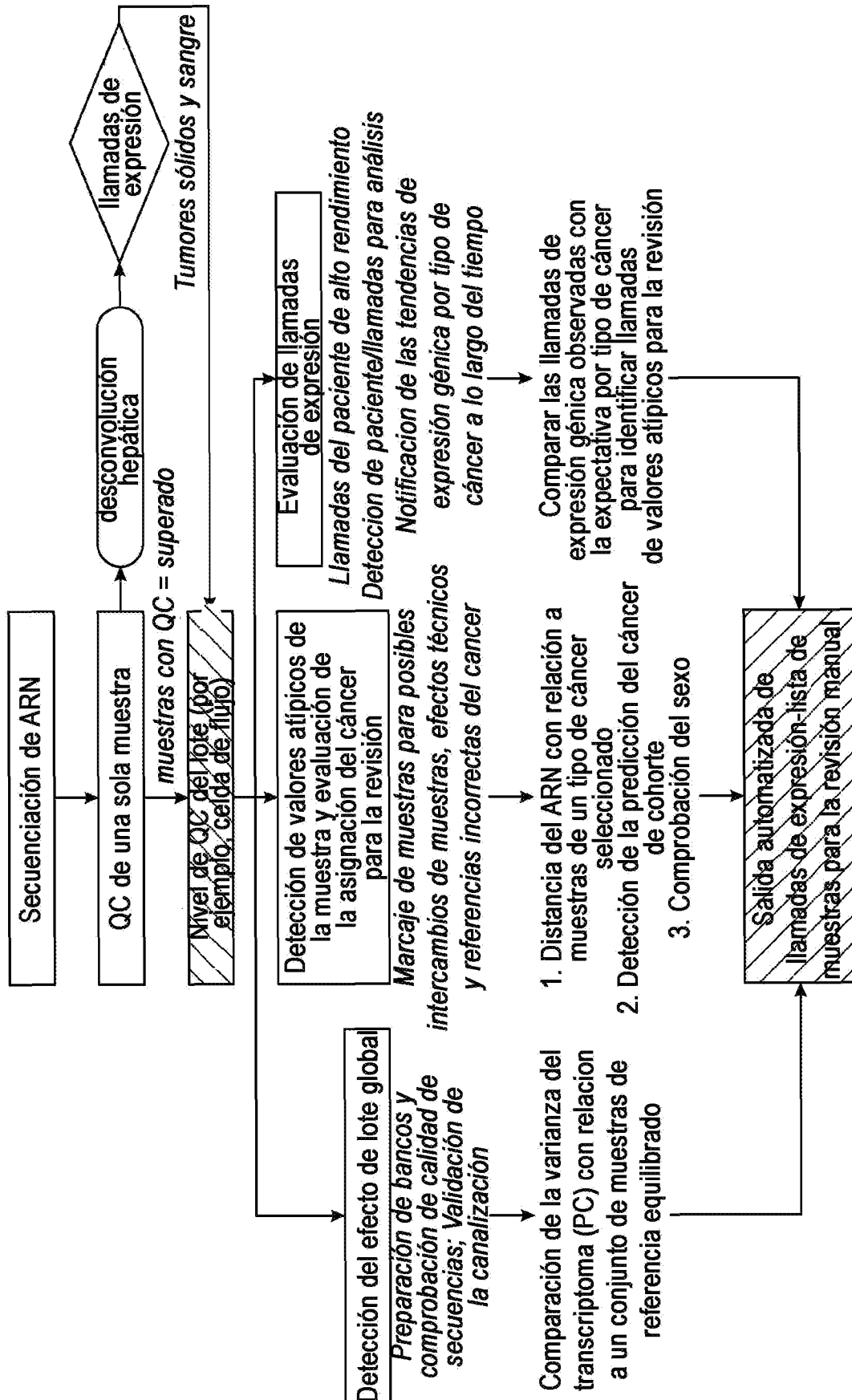


FIG. 14

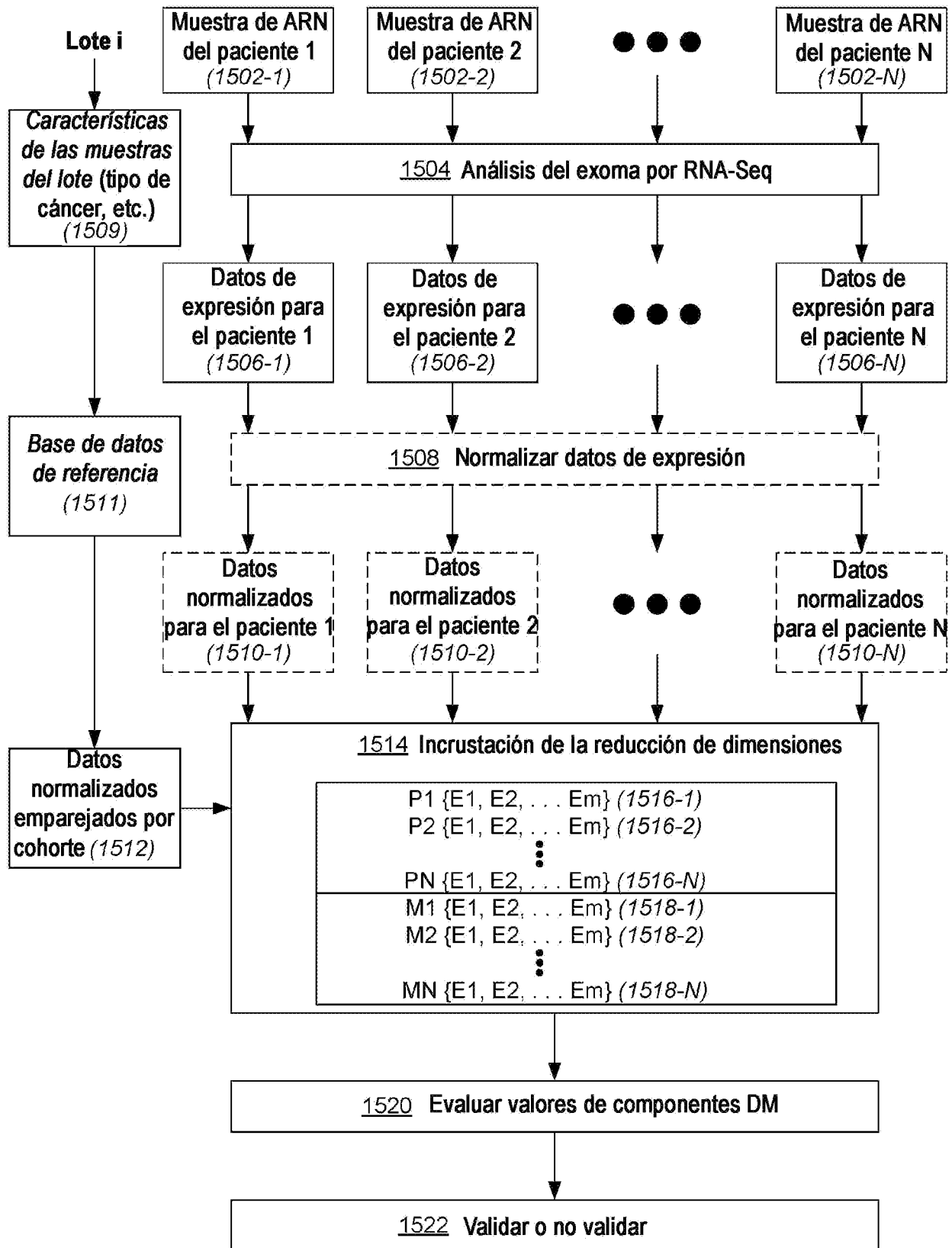


FIG. 15

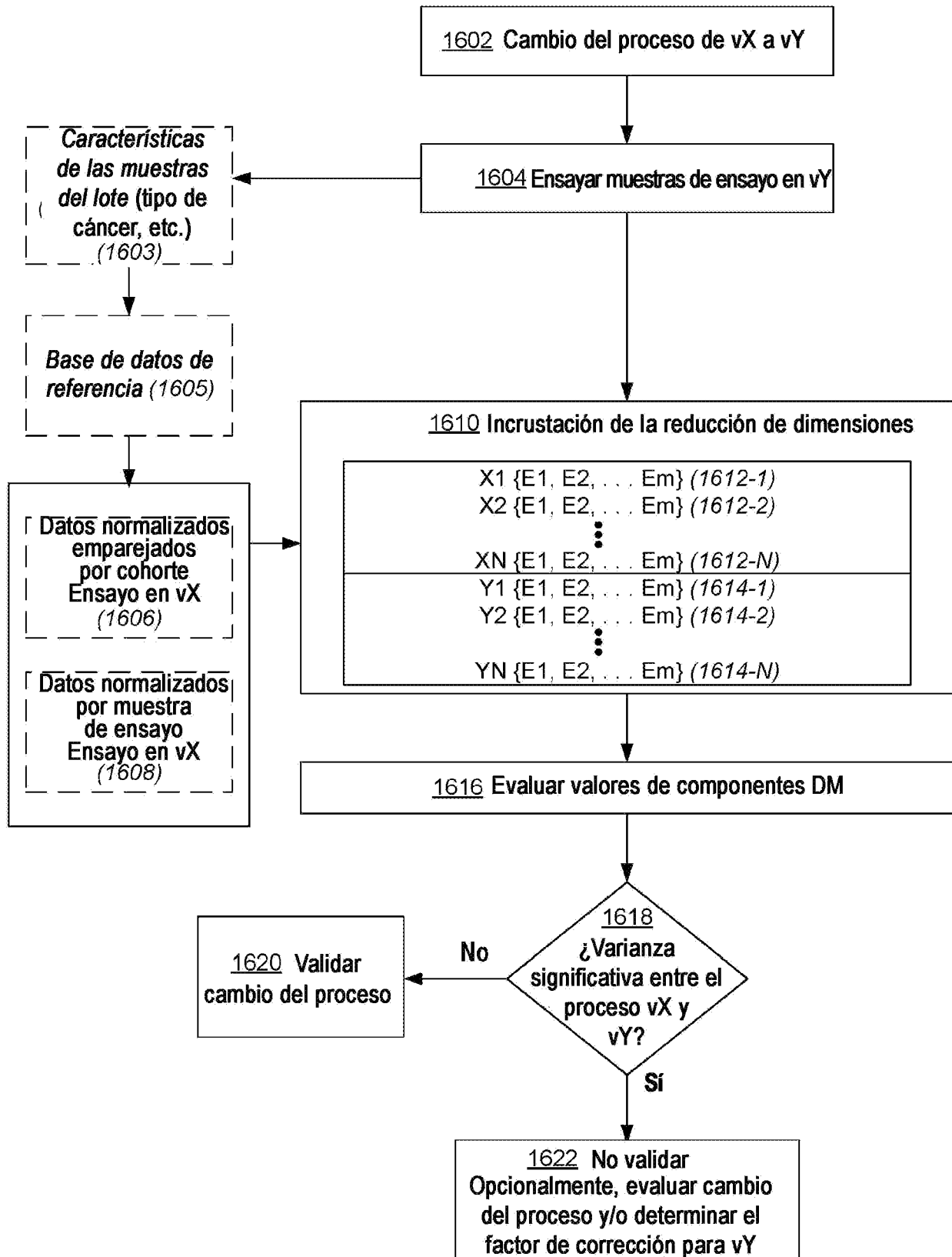


FIG. 16

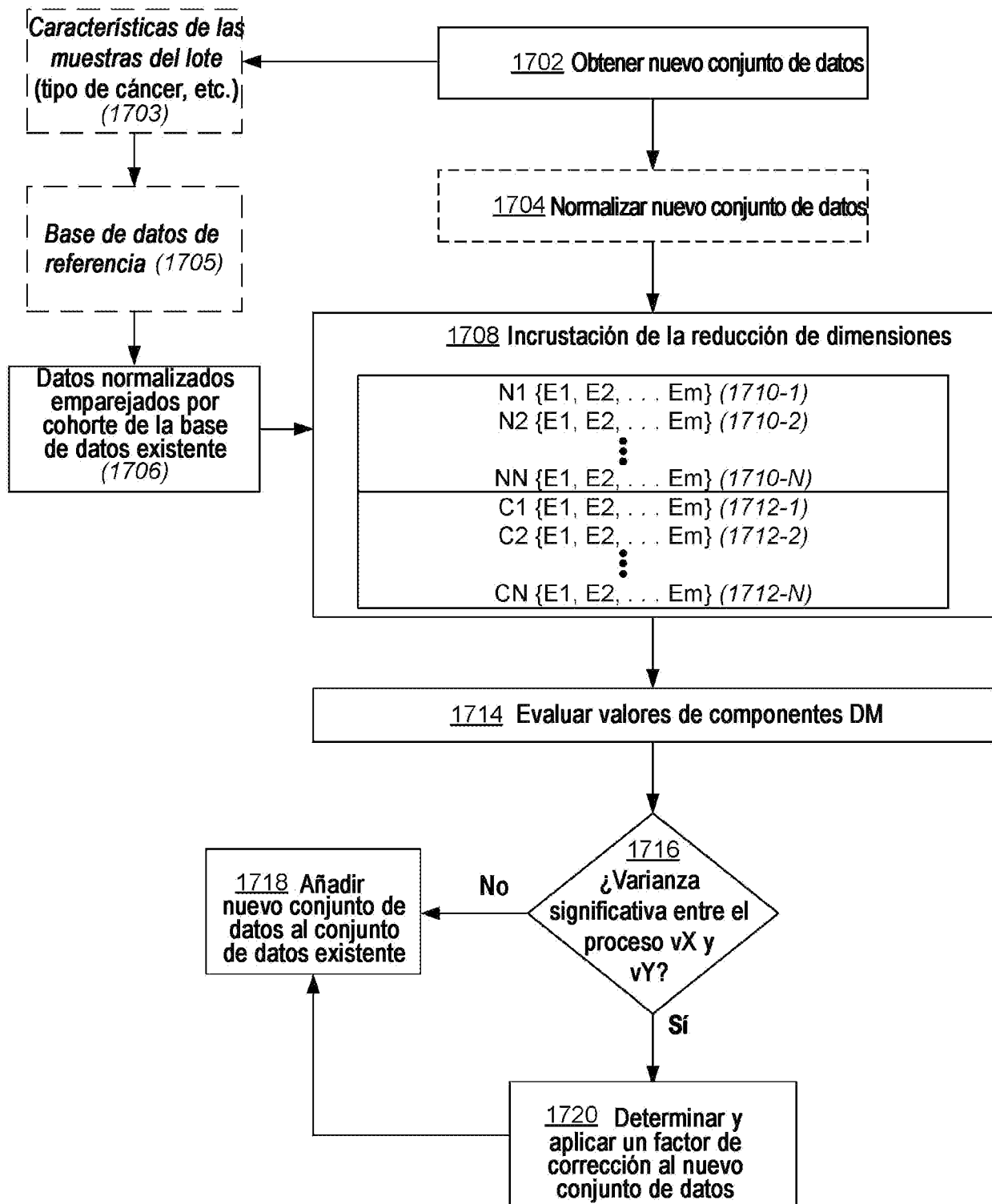


FIG. 17